

CHAPTER 2

DESCRIPTIVE ANALYSIS OF ENGLISH LOANWORDS IN KOREAN

This chapter presents a large scale quantitative analysis of English loanwords in Korean. The analysis is based on a list of 10,000 orthographically and phonologically aligned English words attested as loanwords in Korean, and it details a number of previously unreported effects of orthography on the phonological adaptation of English loanwords in Korean. The loanwords analyzed here are also used as data in a series of experiments on English-Korean transliteration ³ and identifying English loanwords in Korean ⁴.

The remainder of this chapter describes the data set and aspects of English loanword adaptation in Korean. Section 2.1 deals with details of the construction of the data set including criteria for inclusion, data formatting, obtaining English phonological representations, and aligning orthographic and phonological forms. Section 2.2 presents an analysis of how orthography influences the adaptation of English loanwords in Korean, particularly with respect to vowels.

2.1 Construction of the Data Set

This analysis is based on a list of 10,000 English words attested as loanwords in Korean. The majority of the words (9686) come from the National Institute of the Korean Language's (NIKL) list of foreign words (NIKL, 1991) after removing duplicate entries, proper names and non-English words. Entries considered duplicates in the

NIKL list are spelling variants like *traveller/traveler*, *analog/analogue*, *hippy/hippie*, etc. The remainder (314) were manually extracted from a variety of online Korean text sources.

The original NIKL list of foreign words used in Korean contains 20,420 items from a number of languages, including Italian, French, Japanese, Greek, Latin, Hindi, Hebrew, Mongolian, Russian, German, Sanskrit, Arabic, Persian, Spanish, Vietnamese, Malaysian, Balinese, Dutch, and Portuguese. Non-English words are often labeled according to their etymological source, whereas English words (the majority) are not labeled.

In many cases, however, a word which follows a non-English pattern of adaptation is not labeled. For example, certain terms like *acetylase* and *amidase* are labeled in the NIKL list as German, whereas terms like *catalase* and *aconitase* are not labeled. However, the latter items are pronounced in Korean following the sound patterns of the labeled German words – in particular, the final syllable is given as /aatʃe/, as shown in Table 2.1. This pronunciation contrasts with other words ending

Etymological Label	Orthographic Form	Kr. Orthography	Kr. Pronunciation
German	acetylase	아세틸라아제	/aset ^h illaatʃe/
German	amidase	아미다아제	/amitaatʃe/
None	catalase	카탈라아제	/k ^h at ^h alaaʃe/
None	aconitase	아코니타아제	/ak ^h onit ^h aaʃe/

Table 2.1: Example of labeled and unlabeled German loanwords

in the orthographic sequence *-ase*, which are realized in Korean as /eisi/ as would be expected on the basis of the English pronunciation (Table 2.2).

Unlabeled words whose pronunciation matched labeled non-English words were removed, as were words not contained in an online dictionary (American Heritage Dictionary, 2004). The ultimate decision to include a word as English came down

Etymological Label	Orthographic Form	Kr. Orthography	Kr. Pronunciation
None	pericalse	페리클레이스	/p ^h erik ^h ileisi/
None	base	베이스	/peisi/

Table 2.2: Example of unlabeled English loanwords

to a subjective judgment: if the word was recognized as familiar, it was included; otherwise, it was discarded.

Each entry in the list corresponds to an orthographically distinct English word and consists of four tab-separated fields: English spelling, English pronunciation, linearized hangul transliteration, and orthographic hangul transliteration. The first three fields in each entry are aligned at the the character level. An example entry is shown below.

s-pi-der s-pY-dX- s|paid~- 스파이더

Figure 2.1: Example loanword alignment

The list is stored in a single, UTF-8 encoded text file, with one entry per line. UTF-8 is a variable length character encoding for Unicode symbols that uses one byte to encode the 128 US-ASCII characters and uses three bytes for Korean characters. Because it is a plain text file, it is not tied to any proprietary file format and can be opened with any modern text editor.

2.1.1 Romanization

Korean orthography is based on an alphabetic system that is organized into syllabic blocks containing two to four characters each. In standard Korean character encodings such as EUC-KR or UTF-8, each syllabic block is itself coded as a unique character.

This means that there is no longer an explicit internal representation of the individual orthographic characters composing that syllable. For example, in UTF-8 the Korean characters `ㅅ`, `ㅌ`, and `ㄴ` are represented as `'\u1112'`, `'\u1161'`, and `'\u1102'`, respectively. However, the Korean syllable composed of these characters, `한`, is not represented as `'\u1112\u1161\u1102'` but as its own character `'\uD55C'`. Therefore, determining character-level mappings (i.e., phoneme-to-phoneme or letter-to-letter) between Korean and English words is possible only by converting the syllabic blocks of Korean orthography into a linear sequence of characters. One way to do this is to convert hangul representations into an ASCII-based character representation.

For romanization of the data set, priority was given to a one-to-one mapping from hangul letters to ASCII characters because this simplifies many string-based operations like aligning and searching. Multicharacter representations such as Yale romanization (Martin, 1992) or phonemic representations like those in the CMU Pronouncing Dictionary (Weide, 1998) require additional processing or an additional delimiter between symbols. Furthermore, the symbol delimiter must be distinct from the word delimiter.

As much as possible, romanization of the data set is phonemic in the sense that it uses ASCII characters that are already in use as IPA symbols. Consonant transliteration follows Yoon and Brew (2006), which in turn is based on Revised Romanization of Korean. We modified this transliteration scheme so that tense consonants are single character and velar nasal is single character. Table 2.3 (left column) shows the list of consonant equivalences. Vowels were romanized on the basis of the IPA transliterations given in Yang (1996: 251, Table III), using the ASCII equivalents from the Hoosier Mental Lexicon (HML) (Nusbaum, Pisoni, and Davis, 1984). Vowel equivalents are shown in Table 2.3, right column. This dissertation uses Yale

Consonants			Vowels		
Hangul	IPA	Romanized	Hangul	IPA	Romanized
ㄱ	/k/	g	ㅏ	/a/	a
ㅋ	/k*/	G	ㅑ	/æ/	@
ㄴ	/n/	n	ㅓ	/ʌ/	^
ㄷ	/t/	d	ㅕ	/ɛ/	e
ㅌ	/t*/	D	ㅗ	/o/	o
ㄹ	/l/	l	ㅜ	/u/	u
ㅁ	/m/	m	ㅣ	/i/	i
ㅂ	/p/	b	ㅡ	/ɨ/	l
ㅃ	/p*/	B	ㅛ, ㅝ, ㅟ, etc.	/ja, jʌ, jæ/	y+ vowel
ㅅ	/s/	s			
ㅆ	/s*/	S			
ㅇ	/ŋ/	N			
ㅈ	/tʃ/	j			
ㅉ	/tʃ*/	J			
ㅊ	/tʃ ^h /	c			
ㅌ	/t ^h /	t			
ㅋ	/k ^h /	k			
ㅍ	/p ^h /	p			

Table 2.3: Romanization key for transliteration of Korean words into English

romanization to represent Korean orthographic sequences and IPA-based transliteration when pronunciation is of primary importance, following Yang (1996) and Yoon and Brew (2006).

2.1.2 Phonemic Representation

2.1.2.1 Source of Pronunciations

English pronunciations in the data set are represented with the phonemic alphabet used in the HML (Nusbaum et al., 1984). The chief motivation for choosing this phonological representation was ease of processing, which in practical terms means an

ASCII-based, single character per phoneme pronunciation scheme. Pronunciations for English words were derived from two main sources: the HML (Nusbaum et al., 1984) and the Carnegie Mellon Pronouncing Dictionary (CMUDICT) (Weide, 1998). The HML contains approximately 20,000 words, and CMUDICT contains approximately 127,000. Loanwords contained in neither of these two sources were transcribed with reference to pronunciations given in the American Heritage Dictionary (2004).

2.1.2.2 Standardizing Pronunciations

There are several differences between the transcription conventions used in the HML and CMUDICT which had to be standardized for consistent pronunciation. The relevant differences are briefly summarized below, followed by the procedure used for normalizing these differences and standardizing pronunciations.

1. Different alphabets. CMUDICT uses an all-capital phoneme set, with many phonemes represented by two characters (e.g., AA /a/, DH /ð/, etc.). Two-character phones requires using an additional delimiter to separate unique symbols. The HML uses upper and lower case letters, with only one character per phoneme, which does not require an additional delimiter.
2. CMUDICT represents three levels of lexical stress with indices 0, 1, or 2 attached to vowel symbols; the HML does not explicitly represent suprasegmental stress. For example, *chestnut* CEsⁿt (HML) versus CH EH1 S N AH2 T (CMUDICT).
3. The HML distinguishes two reduced vowels (| /i/ vs. x /ə/); CMUDICT treats both as unstressed schwa (AH0 /ə/). For example, *wicked* wIk|d (HML) and W IH1 K AH0 D (CMUDICT) versus *zebra* zibrx (HML) and Z IY1 B R AH0 (CMUDICT).

4. The HML uses distinct symbols for syllabic liquids and nasals; CMUDICT treats these as unstressed schwa followed by a liquid or nasal. For example, *tribal* trYbL (HML) versus T R AY1 B AH0 L (CMUDICT); *ardent* ardNt (HML) versus AA1 R D AH0 N T (CMUDICT).
5. CMUDICT consistently transcribes /oɪ/ sequences as AO R ɔɪ where HML transcribes them as or /oɪ/. For example, *sword* sord (HML) versus S A01 R D (CMUDICT); *sycamore* sIkxmor versus S IH1 K AH0 M A02 R (CMUDICT).

CMUDICT pronunciations were converted to HML pronunciations using the following procedure. In general, information was removed when it could be done so unambiguously rather than attempting to add information from one scheme into the other.

1. CMUDICT unstressed schwa AH0 was converted to HML unstressed schwa x. For example, *action* AE1 K SH AH0 N → AE1 K SH x N; *callous* K AE1 L AH0 S → K AE1 L x S.
2. CMUDICT stressed schwa AH1 or AH2 was converted to HML stressed schwa ^ . For example, *blowgun* B L OW1 G AH2 N → B L OW1 G ^ N; *blood* B L AH1 D → B L ^ D.
3. Remaining stress information was deleted from CMUDICT vowels. For example, *blowgun* B L OW1 G ^ N → B L OW G ^ N; *callous* K AE1 L x S → K AE L x S
4. CMUDICT AO R was converted to HML o r. For example, *sword* S AO R D → S o r D; *sycamore* S IH K x M AO R → S IH K x M o r.
5. Remaining CMUDICT symbols were converted to their HML equivalents using the equivalence chart shown in Table 2.4.

6. HML syllabic liquids and nasals were converted to an unstressed schwa + non-syllabic liquid (nasal) sequence. HML syllabics were expended with schwa following CMUDICT as this made mapping to Korean ㅇ / Λ / easier. For example, *tribal* trYbL \rightarrow trYbxl; *ardent* ardNt \rightarrow ardxNt.
7. HML reduced vowel | /i/ was converted to schwa x. For example, *abandon* xb@nd|n \rightarrow xb@ndxn; *ballot* b@l|t \rightarrow b@lxt.
8. The distinction between HML X / \mathfrak{x} / and R / \mathfrak{z} / was removed. For example, *affirm* xFRm \rightarrow xfXm.

HML	CMUDICT	Example	HML	CMUDICT	Example
a	AA	odd	b	B	be
@	AE	at	C	CH	cheese
^	AH1, AH2	above, hut	d	D	dee
x	AH0	about	D	DH	thee
c	AO	ought	f	F	fee
W	AW	cow	g	G	green
Y	AY	hide	h	HH	he
E	EH	Ed	J	JH	gee
R	ER	hurt	k	K	key
e	EY	ate	l	L	lee
I	IH	it	m	M	me
i	IY	eat	n	N	knee
o	OW	oat	G	NG	ping
O	OY	toy	p	P	pee
U	UH	hood	r	R	read
u	UW	two	s	S	sea
			S	SH	she
			t	T	tea
			T	TH	theta
			v	V	vee
			w	W	we
			y	Y	yield
			z	Z	zee
			Z	ZH	seizure

Table 2.4: Hoosier Mental Lexicon and CMUDICT symbol mapping table.

2.1.3 Alignments

In order to look at the influence of both orthography and pronunciation on English loanwords in Korean, we wanted a three-way, character level alignment between an English orthographic form, its phonemic representation, and corresponding linearized Korean transliteration. English spellings were automatically aligned with their pronunciations using the iterative, expectation-maximization based alignment algorithm detailed in Deligne, Yvon, and Bimbot (1995). The Korean transliteration was aligned with the English pronunciation using a simplified version of the edit-distance procedure detailed in Oh and Choi (2005). The algorithm described in Oh and Choi (2005) assigns a range of substitution costs depending on a set of conditions that describe the relation between a source and target symbol. For example, if the source and target symbol are phonetically similar, a cost of 0 is assigned; an alignment between a vowel and a semi-vowel incurs a cost of 30; an alignment between phonetically dissimilar vowels costs 100, and aligning phonetically dissimilar consonants costs 240. Manually constructed phonetic similarity tables are used to determine the relation between source and target symbols.

We tried a simpler strategy of assigning consonant-consonant or vowel-vowel alignments a low cost consonant-vowel alignments a high cost and found that values of 0 and 10, respectively, performed reasonably well. These costs were determined by trial and error on a small sample. Because there are symbols in one representation that don't have a counterpart in the other (e.g., Korean epenthetic vowels or English orthographic characters that are not pronounced), it is necessary to insert a special null symbol indicating a null alignment. The null symbol is '-'. The resulting alignments are all the same length. The costs assigned determine alignments that tend to obey the following constraints.

1. consonants align with consonants; vowels align with vowels

English Spelling k a n g a r o o

English Pronunciation k @ G g X - u -

Korean k @ N g ^ l u -

2. ‘silent vowels’ align with the null character

English Spelling m a r i n e

English Pronunciation m X - i n -

Korean m ^ l i n -

3. phonemes align at the left edge of orthographic character clusters

English Spelling f i - g h t -

English Pronunciation f Y - - - t -

Korean p a i - - t |

4. Korean epenthetic vowels align with the null character in the English orthography and pronunciation

English Spelling s - m o k e -

English Pronunciation s - m o k - -

Korean s | m o k - |

Because the accuracy of the alignments is crucial to the quality of any analyses of the data set, each alignment was checked by hand and corrected if necessary to ensure that the above constraints are satisfied.

This representation of the correspondences between English and Korean characters makes it easy to possible to derive alignments between any two levels sans the third by deleting correspondences between the null character. For example, alignments between English spelling and pronunciation can be obtained by deleting a ‘-’ that arises from Korean vowel epenthesis:

English Spelling s - m o k e - → s m o k e

English Pronunciation s - m o k - - → s m o k -

Alignments between English pronunciation and Korean can be obtained by deleting a ‘-’ that arises from silent orthographic characters:

English Pronunciation s - m o k - - → s - m o k -

Korean s | m o k - | → s | m o k |

Many-to-many correspondences between two levels may be obtained by consuming the null character in either level and concatenating symbols at both levels. For example, correspondences between English phones and orthographic character sequences can be obtained as:

English Spelling f | i - g h | t - → f igh t

English Pronunciation f | Y - - - | t - → f Y t

Correspondences between English spelling and Korean can be obtained as:

English Spelling f | i - g h | t - → f igh t

Korean p | a i - - | t | → p ai t|

Correspondences between English pronunciation and Korean can be obtained as:

English Pronunciation f | Y - - - | t - → f Y t

Korean p | a i - - | t | → p ai t|

2.2 Analysis of English Loanwords in Korean

In recent years, computational and linguistic approaches to the study of English loanwords in Korean have developed in parallel, with little sharing of insights and techniques. Computational approaches are oriented towards practical problem solving, and are framed in terms of identifying a function that maximizes the number of correctly transformed inputs. Linguistic analyses are oriented towards finding evidence for a particular theoretical point of view and are framed in terms of identifying general linguistic principles that account for a given set of observations. One of the main differences between these two approaches is the relative importance each places on the role of source language orthography in determining the form of a borrowed word. English orthography figures prominently in computational approaches. Early work derived mappings directly between English and Korean spellings (e.g., Kang

and Choi, 2000a), while later work considers the joint contribution of orthographic and phonological information (e.g., Oh and Choi, 2005).

Many linguistic analyses of loanword adaptation, however, consider orthography a confound, as in Kang (2003: 234):

“problem of interference from normative orthographic conventions”

or uninteresting, as in Peperkamp (2005: 10):

“Given the metalinguistic character of orthography, adaptations that are (partly) based on spelling correspondences are of course of little interest to linguistic analyses”

Linguistic accounts of English loanword adaptation in Korean instead focus on whether the mechanisms of loanword adaptation are primarily phonetic or phonological. Other analyses of loanword adaptation in other languages acknowledge that orthography interacts with these mechanisms (e.g., Smith (2008) on English loanword adaptation in Japanese).

This section looks at some influences of orthography on English loanwords in Korean, and shows that English spelling accounts for substantially more of the variation in Korean vowel adaptation than phonetic similarity does. The relevance of this correlation is illustrated for the case of variable vowel epenthesis following word final voiceless stops, and discussed more generally for understanding English loanword adaptation in Korean.

The Korean Ministry of Culture and Tourism (1995) published a set of phonological adaptation rules that describe the changes that English phonemes undergo when they are borrowed into Korean. Example rules are shown below (Korean Ministry of Culture and Tourism, 1995: p. 129: 1(1), 2).

1. after a short vowel, word-final voiceless stops ([p], [t], [k]) are written as codas (p, s, k)
book [bʊk] → *puk*
2. i is inserted after word-final and pre-consonantal voiced stops ([b], [d], [g])
signal [sɪgnəl] → *sikinəl*

These rules were implemented as regular expressions in a Python script and applied to the phonological representations of English words in the data set (this procedure is explained in detail in Chapter 3 Section 3.3.1). The output of the program was compared to the attested Korean forms, and the proportion of times the rule applied as predicted was calculated for each English consonant. These results are shown in Table 2.5.

Stops		Fricatives		Nasals		Glides	
p	0.990	f	0.999	m	1.000	r	0.988
t	0.989	v	0.985	n	0.997	l	0.987
k	0.990	θ	0.978	ŋ	0.983	w	0.967
b	0.996	ð	1.000			j	0.859
d	0.996	s	0.975				
g	0.984	z	0.733				
		ʃ	0.985				
		ʒ	1.000				
		tʃ	0.951				
		dʒ	0.969				
		h	0.983				

Table 2.5: Accuracy by phoneme of phonological adaptation rules. Mean = 0.97

In general the rules do a good job of predicting the borrowed form of English consonants in Korean. On average, consonants were realized as predicted by the phonological conversion rules 97% of the time. The prediction rates for /z/ and /j/ were substantially below the mean at 0.73 and 0.86, respectively. Based on Korean

Ministry of Culture and Tourism (1995: p. 129: 2, 3(1)) the following rules for the adaptation of English /z/ in Korean loanwords were implemented:

1. word-final and pre-consonantal [z] → ㅈ *tʃi*

jazz [jæz] → 재즈 /tʃæʃi/

2. otherwise, [z] → ㅅ /tʃ/

zigzag [zɪgzæg] → 지그재그 /tʃikiʃæki/

/z/ occurred 704 times in English words in the data set; it was realized according to the rule as ㅅ *tʃ* 512 times and realized as ㅈ *s* 188 times. In 117 of these cases, the unpredicted form corresponds to English word-final /z/ representing the plural morpheme (orthographic ‘-s’). Examples include words like *users* /juzəz/ → 유저스 /juʃʌsi/, *broncos* /bræŋkəz/ → 브롱코스 /pɪlɔŋk^hosi/, and *bottoms* /batəmz/ → 보텀스 /pɒt^hʌmsi/. The contingency table in 2.6 shows how often /z/ is realized as predicted with respect to the English grapheme spelling it. The χ^2 significance test indicates that /z/ is significantly more likely to become ㅈ *s* in Korean when the English spelling contains a corresponding ‘s’ than when it does not (Yates’ $\chi^2 = 100.547, df = 1, p < 0.001$).

	s	¬s	English Orthography
/z/ → ㅈ <i>tʃ</i>	300	212	
/z/ → ㅈ <i>s</i>	185	3	

Table 2.6: Contingency table for the transliteration of ‘s’ in English loanwords in Korean

Although this result indicates that English spelling is a more reliable indicator of the adapted form of /z/ than its phonological identity alone, it does not tease apart the question of whether low level phonetics or morphological knowledge of English is responsible for this adaptation pattern. English word-final /z/ often devoices (e.g.

Smith, 1997); if the adaptation of these words is based on [s] rather than /z/, these cases would be regularly handled under the rule for the adaptation of English /s/. Alternatively, these borrowed forms may represent knowledge of the morphological structure of the English words, in which a distinction between \varkappa *ʃ* and \wedge *s* is maintained in the borrowed forms.

The following rule predicts the appearance of English /j/ in English loanwords in Korean (Korean Ministry of Culture and Tourism, 1995):

$$[j] \rightarrow y.$$

/j/ occurred 368 times in English loanwords in the data set; 275 of these cases were adapted as the predicted *j* (e.g., *yuppie* /jʌpi/ → 여피 /jʌp^{hi}/), while 35 were adapted as *i* (e.g., *billion* /bɪljən/ → 빌리언 /pilliʌn/) and 58 were adapted as \emptyset (e.g., *cellular* /sɛljʊlə/ → 셀룰러 /sellullʌ/). These cases are examined separately in the χ^2 tables 2.7 and 2.8. Table 2.7 shows how often English /j/ transliterates as Korean ㅇ| /i/ with respect to whether the English spelling contains a corresponding ‘i’. The

	i	¬i
j→j	7	64
j→ \emptyset	29	4

Table 2.7: Contingency table for the transliteration of /j/ in English loanwords in Korean

results of the χ^2 test indicate that when the English orthography contains the vowel ‘i’, /j/ is more likely to be transliterated as ㅇ| /i/ (Yates’ $\chi^2 = 57.192$, $df = 1$, $p < 0.001$). Table 2.8 shows how often English /j/ is produced in the adapted form with respect to whether the English orthography contains a corresponding character. The results of the χ^2 test indicate that /j/ shows a tendency to drop when the orthography does not support its inclusion (e.g., *cellular*) ($\chi^2 = 4.725$, $df = 1$, $p \leq 0.03$).

	y	∅
j→j	54	204
j→∅	5	53

Table 2.8: Contingency table for the transliteration of ‘i’ in English loanwords in Korean

Whereas the behavior of English consonants in loanwords in Korean is reliably expressed with a handful of phonological rules, the behavior of vowels is considerably less constrained. Table 2.9 shows the number of transliterations found in the data set for each English vowel. The average number of transliterations per vowel is 8.46.

English Vowel	Number of Korean Transliterations
a	7
æ	6
ɔ	6
e	11
ʊ	5
ɪ	9
o	10
i	9
u	6
ʒ	15
ə	12
ɛ	9
ʌ	5

Table 2.9: Average number of transliterations per vowel in English loanwords in Korean

Korean Ministry of Culture and Tourism (1995) does not provide phonological rules describing the adaptation of English vowels to Korean. However, Yang (1996) provides acoustic measurements of the English and Korean vowel systems. Based on this data, it is possible to estimate the acoustic similarity of the English and Korean

vowels, and examine the relation between the cross language vowel similarity and transliteration frequency. The prediction is that acoustically similar Korean vowels will be substituted for their English counterparts more frequently than non-similar vowels. Recognizing that acoustic similarity is not necessarily the best predictor of perceptual similarity (e.g., Yang, 1996), we nonetheless applied two measures of vowel distance and correlated each with transliteration frequency.

The first measurement was the Euclidean distance between vowels using F1 through F3 measurements for English and Korean vowels from Yang (1996):

$$(2.1) \sqrt{\sum_{i=1}^3 (F_{Ei} - F_{Ki})^2}$$

The notion of a perceptual $F2'$ has been recognized as relevant since Carlson, Granström, and Fant (1970) introduced it for accounting for the perceptual integration of the higher formants. We calculated $F2'$ according to the formula in Padgett (2001: 200):

$$(2.2) F2' = F2 + \frac{F3 - F2}{2} \times \frac{F2 - F1}{F3 - F1}$$

and applied the Euclidean distance formula in 2.3 to calculate vowel distance:

$$(2.3) \sqrt{(F_{E1} - F_{K1})^2 + (F_{E2'} - F_{K2'})^2}$$

The correlation between vowel distance and frequency of transliteration in an acoustic-perceptual space is very weak. Table 2.10 shows the associated correlations between each of the distance measures and vowel transliteration frequency.

Measure	Correlation
Euclidean	-0.256
$F2'$	-0.331

Table 2.10: Correlation between acoustic vowel distance and transliteration frequency

However, in many cases the Korean vowel corresponds to a normative “IPA reading” of the English orthographic vowel, regardless of its actual pronunciation. A much stronger correlation is found between the number of ways a vowel is written in English and the number of adaptations of that vowel in Korean ($r = 0.92$). For example, ə is represented orthographically in a variety of ways in English (e.g., action, Atlanta, cricket, coxswain, instrumentalism) and shows a variety of realizations in loanwords in Korean (e.g., 액션 *ayksyen*, 애틀랜타 *aythullayntha*, 크리케트 *khulikeythu*, 콕스웨인 *khoksuweyin*, 인스트루멘탈리즘 *insuthulwumeynthellicum*). This correlation is depicted graphically in Figure 2.2.

Finally, we note an orthography-sensitive distinction that concerns epenthesis following word final voiceless stops. Kang (2003) observes that English tense vowels preceding a voiceless stop often trigger final vowel epenthesis. The standard conversion rules also specify this phenomenon, in terms of vowel length (Korean Ministry of Culture and Tourism, 1995: 1.3). Examples are shown in Table 2.11.

In English, orthographic ‘o’ is typically pronounced one of two ways: /o/ (e.g. hope, smoke) and /a/ (e.g., pot, lock). These words are typically borrowed into Korean in one of two ways, as well. English words containing pre-final /o/ are typically produced in Korean with *o* ‘ㅜ’ plus epenthesis (e.g., rope 로프 *lophu*, smoke *sumokhu*). However, many English words pronounced /a/ are borrowed with /o/ ‘ㅜ’ as well, presumably on the basis of the English orthography (e.g., hardtop 하드토프 *hatuthop*, headlock 헤드록 *heytulok*, etc.). Although the form of the adapted

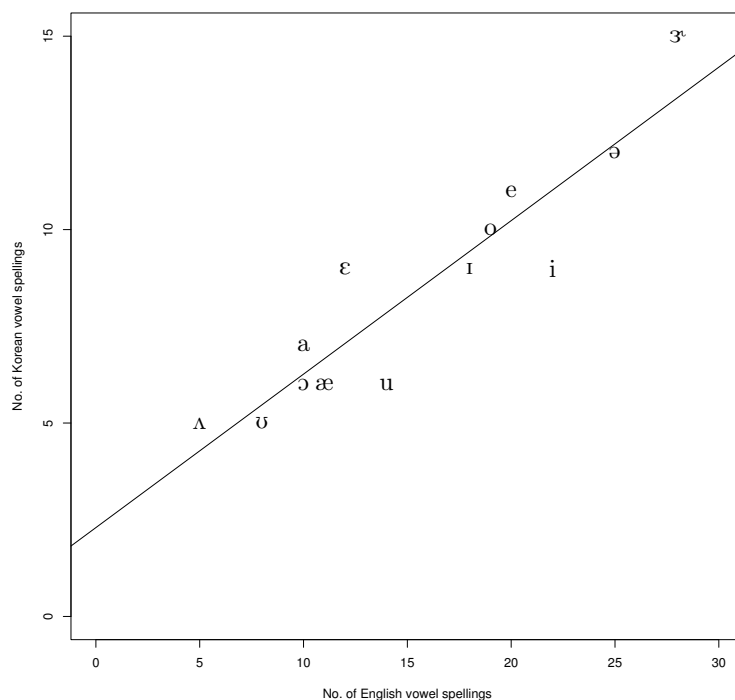


Figure 2.2: Correlation between number of loanword vowel spellings in English and Korean

vowel is the same in both cases, epenthesis is significantly less likely to occur for orthographically derived /o/ than when /o/ corresponds to the English pronunciation as well (Yates' $\chi^2 = 107.57$; $df = 1$; $p < .0001$). Examples are given in Table 2.12, which contains a breakdown of the epenthesis data for /o/ by identity of the following stop. For /k/ and /p/, epenthesis is very unlikely when the English letter 'o' is pronounced /a/; for /t/, orthographically derived /o/ is as likely to epenthesize as pronunciation-based /o/¹. In essence, the Korean phonology preserves a distinction

¹This difference may reflect morphophonemic constraints on final /t/ in Korean nouns (Kang, 2003).

English	Korean
rope	로프 <i>lophu</i>
smoke	스모크 <i>sumokhu</i>
part	파트 <i>phathu</i>
make	메이크 <i>meyikhu</i>

Table 2.11: Examples of final stop epenthesis after long vowels in English loanwords in Korean

between phonologically and orthographically derived /o/ in terms of epenthesis on the final voiceless stop.

Eng. Pron.	Examples	Epenthesis	No Epenthesis
/ap/	desktop/데스크톱 <i>teysukhuthop</i> turboprop/터보프로프 <i>thepophulophu</i> [†]	0	27
/op/	rope/로프 <i>lophu</i> [†] soap/소프 <i>sophu</i> [†]	32	0
/ak/	hemlock/헴록 <i>heymlok</i> smock/스목 <i>sumok</i>	5	36
/ok/	spoke/스포크 <i>suphokhu</i> [†] stroke/스트로크 <i>suthulokhu</i> [†]	15	0
/at/	ascot/애스콧 <i>aysukhos</i> boycott/보이콧 <i>poikhos</i>	11	12
/ot/	tugboat/터그보트 <i>thekupothu</i> [†] vote/보트 <i>pothu</i> [†]	26	0

Table 2.12: Vowel epenthesis after voiceless final stop following Korean /o/. [†] indicates epenthesis

2.3 Conclusion

This chapter described the preparation of a set of English-Korean loanwords that is aligned at the character level to show correspondences between English spelling, pronunciation and the Korean form of borrowed English words. This is the only

resource of its kind that is freely available for unrestricted download: <http://purl.org/net/kbaker/data>. Several analyses of the data were presented which highlight previously unreported observations about the influence of orthography on English loanword adaptation in Korean. Orthography has a particularly noticeable influence on the realization of vowel in English loanwords in Korean. Vowel adaptation is not reliably predicted from the phonological representation of vowels in English source words in the absence of orthographic information, whereas consonant transliteration is reliably captured by a small set of phonological conversion rules.

The analysis presented here also identified cases where English orthography interacts with the Korean phonological process of word final vowel epenthesis following voiceless stops. These findings are important for accounts of English loanword adaptation in Korean because they provide a quantification of the extent to which orthography influences the form of borrowed words, and indicate that accounts of loanword adaptation which focus exclusively on the phonetics or phonology of the adaptation process are overlooking important factors that shape the realization of English loanwords in Korean. The next chapters use the data set described here in a series of experiments on automatic English-Korean transliteration and foreign word identification.

	/a/	/o/	English pronunciation of 'o'
Korean /o/ 'ㅏ', with Epenthesis	16	73	
Korean /o/ 'ㅏ', no Epenthesis	75	0	

Table 2.13: Relation between voiceless final stop epenthesis after /o/ 'ㅏ' and whether the Korean form is based on English orthography 'o' or phonology /a/. $\chi^2 = 107.57$; $df = 1$; $p < .001$