LETTER-TO-SOUND RULES FOR KOREAN

Kyuchul Yoon, Chris Brew and Mary Beckman

Department of Linguistics
The Ohio State University
1712 Neil Avenue, Columbus, Ohio, USA
{kyoon|cbrew|mbeckman}@ling.ohio-state.edu

ABSTRACT

We describe a Korean letter-to-sound (LTS) subsystem built within the Festival TTS framework [1]. The subsystem has two major components. In the first step Korean orthography is deterministically mapped, without consideration of context, to a phonemic representation close to the ISO TR 11941:1996 transcription scheme [2]. The second step is an implementation of the most important morphophonemic rules of Korean. This second component is somewhat more complex, and needs to respond to context, syllable structure, morphology and (ideally, but not implemented here) etymology. We outline the design principles of the TTS system of which this LTS module will be a part and describe the evaluation of the LTS rules against a large corpus.

1. INTRODUCTION

This paper motivates and describes a set of letter-to-sound rules which we have developed for standard Korean. This work is part of a larger effort to develop a suite of demonstration speech synthesizers for a variety of typologically distinct languages. This is in turn related to a yet larger effort to make a case for a particular view of the place of segmental variation in relation to prosody and morphology.

For the Korean system, we focused first on LTS rules, in order to use them in designing a database of reasonably familiar phrases that can be recorded to test hypotheses about the effects of prosodic position on segmental production and to extract positionally-specific diphones. That is, our immediate goal is to select phrases from a (text-only) corpus of Korean Newswire [3], subject to the usual practical constraints: we need the number of phrases to be small enough that they can be recorded in a single session, but we require these phrases to cover as much as possible of the linguistic repertoire of the Korean speaker whose voice we model.

Thanks to SBC/Ameritech for a fellowship awarded to the second and third authors, supporting the first.

2. KOREAN LETTERS AND SOUNDS

Most Korean writing today is completely in hangul (perhaps with interpolations of roman alphabet text). This means two things. First, hangul conversion covers much of grapheme-to-phoneme conversion and the problems of providing the right Sino-Korean reading for a Chinese character can be circumvented for most online texts such as the Newswire corpus that is our primary database. Also, the LTS rules can refer to the two types of prosodic constituents that are encoded in the way that hangul elements are organized on the page. The smallest elements are the *jamo*, or alphabetic letters. The *jamo* are arranged into *eumjeol*, or 'orthographic syllables', making it possible to distinguish onset consonants from coda consonants. This is important for the implementation of some morphophonological rules as LTS rules. Also, sequences of *eumjeol* are grouped into *eojeol*, or 'orthographic words', corresponding roughly to accentual phrases (AP) [4], a tonally demarcated unit that can include one or more lexical items. This is important in designing a corpus of materials for recording diphones, since many allophonic processes refer to position within the AP.

Starting from a romanization program written by Nick Cipollone for the ISO TR 11941:1996 transcription scheme, we created a phonetization program that produces output appropriate for use with the LTS rule programming language of Festival. The phonetizer is a straightforward table-driven program that maps the two-byte code for each hangul syllable to an appropriate romanized representation (for details of how Korean text is encoded, see [2], especially appendix N). The output of the phonetizer is the input to the LTS rules proper. The input and output of the LTS rules is the same phone set, which is listed in Table 1. Note that, with the exception of [l] vs [r], the phone set corresponds to the *jamo* set, which differentiates only phonemic contrasts, not major allophones. We chose not to extend the phones beyond those differentiated by the transcription scheme, because we plan to record separate diphones to take account of effects of prosodic position at the levels of the syllable and AP. That is, our working hypothesis is that that the ef-

fects of prosodic position are the primary type of allophonic variation that we need to model, and we can encode that variation more naturally by recording different diphones for each different relevant prosodic position.

VOWELS			
Segment	Height	Front/Back	
/a/ [a]	low	back	
/i/ [i]	high	front	
/eu/ [ɨ]	high	central	
/u/ [u]	high	back	
/e/ [e]	mid	front	
/eo/ [ʏ]	mid	central	
/o/ [o]	mid	back	
CONSONANTS			
Segment	Manner	Place	Phonation
/b/ [p]	stop	labial	lenis
/bb/ [p']	stop	labial	fortis
/p/ [p^h]	stop	labial	aspirated
/d/ [t]	stop	alveolar	lenis
/dd/ [t']	stop	alveolar	fortis
/t/ [t^h]	stop	alveolar	aspirated
/g/ [k]	stop	velar	lenis
/gg/ [k']	stop	velar	fortis
/k/ [k^h]	stop	velar	aspirated
/ss/ [s']	fricative	alveolar	fortis
/s/ [s^h]	fricative	alveolar	aspirated
/h/ [h]	fricative	glottal	aspirated
/j/ [tʃ]	affricate	palatal	lenis
/jj/ [tʃ']	affricate	palatal	fortis
/c/ [$tʃ^h$]	affricate	palatal	aspirated
/m/ [m]	nasal	labial	lenis
/n/ [n]	nasal	alveolar	lenis
/ng/ [ŋ]	nasal	velar	lenis
/w/ [w]	approximant	labial	lenis
/l/ [l]	lateral	alveolar	lenis
/r/ [ɾ]	rhotic	alveolar	lenis
/y/ [j]	approximant	palatal	lenis

Table 1. The phoneme inventory for Korean TTS, with IPA for common allophone, and defining features.

The features for the phone set classify the vowels by height and front/backness, and the consonants by manner, place, and phonation type. There is no feature of vowel length, because there are no distinctive differences in vowel length for most younger speakers of the standard dialect, including the speaker who will record our diphone set. Also, [lenis] is used as a cover term for both the (canonically voiceless) lenis obstruents /b/, /d/, /g/, and /j/, and the inherently voiced sonorant consonants /m/, /n/, /ng/, /w/, /l/, /r/, and /y/.

3. MORPHO-PHONOLOGICAL RULES

An obvious alternative to calling sonorants [lenis] would have been to include the term [voiced] to describe the phonation type of these consonants and the relevant allophones of /b/, /d/, /g/, and /j/. That is, we could have included a postlexical phonological rule of 'lenis-stop voicing' to encode the difference between the typically longer voiceless closure and rather breathy vowel onset in AP-initial position and the typically very short closure with no clear interruption of voicing in AP-medial position. However, this treatment would miss the generalization that the two other obstruent types also have different allophones in these different positions. For example, in AP-initial position, the non-fortis fricative has a noticeable interval of aspiration, and the aspirated stops and affricates have longer voice onset time than they do in AP-medial position. And there are other phenomena, such as high vowel devoicing, that tend to interact with these positional effects. Therefore it makes sense to keep the phoneme inventory simple and instead handle the all of the AP-related allophony by recording different diphones for AP-initial and medial positions.

This contrasts to our treatment of the 'flapping rule', which describes the alternation between [l] and [ɾ] in forms such as /mul/ [mul] 'water' as opposed to /mul/+/i/ [muɾi] 'water (nom.)'. This alternation is encoded explicitly in the ISO TR 11941:1996 romanization, which represents the relevant (single) *jamo* as /r/ in syllable-initial position, but /l/ in coda position. In the native vocabulary, the conditioning of the allophone by syllable position is fairly straightforward. The *jamo* is pronounced as [ɾ] in syllable-onset position between vowels, but as [l] in coda position and in onset position after a coda [l] in the preceding syllable. In word-intial position in the various foreign strata of the vocabulary, however, the same *jamo* can stand for other sounds. In Sino-Korean forms, it represents a contentless onset or a glide variant of the following high vowel (as in the normal pronunciation [ji] of the common surname that is romanized as *Lee* or *Rhee*). In more recent loanwords, it can also represent [ɾ], as in the loanword for *radio*). This conditioning by etymology in the various foreign strata of the language argues against a treatment comparable to our treatment of 'lenis stop voicing' and related AP-conditioned allophonic effects.

Ranging between these extremes of regular prosodically-conditioned allophony and word-specific pronunciations are a host of more or less regular morphophonological processes that reduce the phonological transparency of hangul spellings. Our TTS system for Korean explicitly encodes the following subset of such processes as LTS rules.

Palatalization. The two coronal stops /d,t/ in coda position are pronounced as the palatal affricates /j,c/ when followed by /i/- or /y/-initial bound morphemes such as the

nominative case marker. For example:

/sot/ ‘cauldron’ + /i/ ‘nom.’ → /so-ci/

Simplification of coda clusters. The first or the second consonant in a coda cluster is deleted, or the second consonant is resyllabified as the onset of the next syllable, depending on such contextual facts as whether the next syllable starts with a vowel and whether the next syllable is part of a bound or a free morpheme. In many cases, the result of this process becomes the input to other rules such as ‘tensification’ or ‘nasalization of obstruents’.

/gabs-jin/ ‘precious’ → /gab-jjin/

/gabs-man/ ‘price alone’ → /gam-man/

Aspiration of lenis obstruents adjacent to /h/. A coda lenis obstruent before an onset /h/ or an onset lenis obstruent after a coda /h/ is fused with the /h/ to result in an open syllable followed by an aspirated obstruent onset.

/beob-hag/ ‘study of law’ → /beo-pag/

/nah-da/ ‘give birth’ → /na-ta/

Neutralization of coda obstruents. All coda obstruents are realized as one of the lenis stops /b,d,g/ unless they precede vowel-initial bound morphemes. This rule provides the input to the tensification rule and liason.

/bu-eok/ ’kitchen’ → /bu-eog/

/ggoc-byeong/ ‘vase’ → /ggod-bbyeong/

/os/ ’garment’ + /an/ ‘inside’ → /o-dan/

Tensification of lenis obstruents. A lenis obstruent onset following any obstruent coda at the morpheme boundary in a compound word is pronounced as a fortis obstruent.

/beob-de/ ‘law school’ → /beob-dde/

/bat-gil/ ‘farm-trail’ → /bad-ggil/

Nasalization of obstruents. An obstruent coda is nasalized when followed by an onset nasal.

/gug-min/ ‘citizens’ → /gung-min/

Nasalization of l (l-to-n). When /l/ is preceded by any consonant other than /n,l/, it is pronounced /n/. This rule provides some of the input contexts for the nasalization of obstruents rule.

/sib/ ‘ten’ + /li/ (a unit of distance) → /sim-ni/

Lateralization of n (n-to-l). The dental nasal /n/ is pronounced as /l/ before or after /l/.

/ceon/ ‘thousand’ /li/ → /ceol-li/

/kal-nal/ ‘knife-blade’ → /kal-lal/

Liason. Like the second consonant in a cluster, a singleton coda consonant is resyllabified as the onset of the next syllable before a vowel-initial syllable. If this syllable is initial to a free morpheme, the rule of coda neutralization applies first.

/ap/ ‘front’ + /i/ ‘nom.’ → /a-pi/

/geot/ ‘outer’ + /os/ ‘garment’ → /geo-dod/

/mul/ ‘water’ + /i/ ‘nom’ → /mu-ri/

Unlike the other rules, liason typically involves only movement of sounds across syllabic boundaries. That is, there is symbol rewrite only when the resyllabified coda was /l/. In other cases, this rule acts as the input to something like our treatment of lenis-stop voicing. That is we capture the relevant allophonic variation by recording different diphones. For example, there are separate /od-/ and /o-d/ diphones to capture the alternation between voiceless unreleased coda /d/ in /ggod-bbyeong/ ‘vase’ and voiced released onset /d/ in /geo-dod/ ‘outer garment’.

Many of the rules refer to morphological context. For example, coda simplification and aspiration next to /h/ are relevant only for syllables in sequence across a morpheme boundary, and their application is sensitive to the status of the second syllable as a bound morpheme or free morpheme. The bound morphemes are enumerated at the beginning of each rule set to capture this sensitivity to morphological status. That is, the rule set scans the input for bound morphemes and if it finds no bound morphemes, it proceeds with the rest of the rule set. The list of bound morphemes can be extended as necessary.

Another aspect of morphological context is trickier, and has to do with the degree of conjoining in a morphologically transparent compound form. This aspect is most relevant for the application of the tensification rule, which was described as applying to an lenis obstruent “after any obstruent coda at the morpheme boundary in a compound word”. That is, tensification is somewhat like the compound stress pattern in English. The application of tensification at a morpheme boundary glues together the components in sequence, signalling to the hearer that this is a compound word. (Note the initial stress in the English translations of /beob-dde/ ‘law school’ and /bat-gil/ ‘farm trail’, which are also compound words.) Tensification differs from the analogous English phenomenon, however, in that applies only to compounds that are relatively transparent semantically. Thus, where English compound stress applies in both ‘blueberry’ and ‘cranberry’, Korean tensification applies in /heo-li/ ‘back’ + /byeong/ ‘disease’ → /ho-li-bbyeong/ ‘backache’, but not in /ji-byeong/ ‘chronic terminal illness’ (where the meaning of /ji/ is opaque in the same way as English ‘cran-’). Since tensification also does not apply in less tightly conjoined compounds that are comparable to English ‘law degree requirements’, the rule probably must overgeneralize when applied to unlimited domain text.

In most rules, the domain of rule application refers to the larger prosodic context of the AP, as well to syllable position and morphological context. In a deliberate or didactic style that is appropriate for reading newspapers, the AP corresponds roughly to the grouping of *eumjeol* into *eojeol*. This means that our phrasing rules can be very simple, referring only to the punctuation marks and spaces that delimit the *eojeol*. (To build a more general-purpose TTS system, of course, we would need more elaborate heuristics to determine the boundaries of accentual phrases.) The rough correspondence between accentual phrases and *eojeol* also

high freq. (≥ 10)				
#rules	1	2	3	4 or more
2-syll.	6015	1639	266	38
3-syll.	20693	9629	2165	326
4-syll.	19029	11574	4266	927
5-syll.	10197	8651	3933	854
6-more	4023	4059	1971	473
med freq. ($2 \geq freq > 10$)				
2-syll.	6914	4624	1775	431
3-syll.	36503	25290	9931	2522
4-syll.	59337	42090	17008	4015
5-syll.	58677	41977	17237	4175
6-more	48747	35848	14966	3683
low freq. (= 1)				
2-syll.	3755	2624	1121	256
3-syll.	26981	19650	8280	2148
4-syll.	51507	38466	16709	4148
5-syll.	60831	47116	20338	5338
6-more	68623	54287	24143	6762

Table 2. Distribution of polysyllabic *eojeol* in the Korean Newswire corpus, by the number of syllables, number of occurences in the Newswire Corpus, and the number of rules applied by the LTS rules.

makes it fairly straightforward to evaluate the LTS rules on a large text corpus.

4. EVALUATION AND FUTURE USES

To evaluate our LTS rules, we first extracted and counted the number of occurrences of each of the 1.3million *eojeol* in the more than 143,000 articles of the Korean Newswire corpus. We applied the LTS rules to all of these *eojeol* and tabulated the results by the length (in number of syllables) of the output "phonemicized" spelling, the frequency in the Newswire corpus of the (input) *eojeol*, and the number of LTS rules that had applied to produce the output. There were about 1500 monosyllabic *eojeol* ranging in frequency from 1 occurrence in the corpus to more than 100,000 occurrences (for the morphemes /deung/ 'endlist', /ggeut/ 'end', /go/ 'indirect quote', /i/ 'this', and /il/ 'nth day'). The numbers for polysyllabic *eojeol* are shown in Table 2.

We had several reasons for tabulating the *eojeol* by the three factors in this table. First, we were interested in number of rule applications, because we expected the output to be most reliable for *eojeol* in which all of the words that are pronounced just as they are spelled, and least reliable for *eojeol* where many rules interact, potentially reflecting irregular diachronic development. Second, we are interested in *eojeol* length and frequency, because long low-frequency *eojeol* are more likely to be phrases or loosely conjoined compounds, where the tensification rule will be mis-applied. Grouping the outcome *eojeol* as in Table 2 allows us to evaluate the LTS rules efficiently across the full range of lengths, frequencies, and number of rule applications.

We selected a random 66 *eojeol* from each of the 12 cells in Table 2. These output 'phonemicized spellings' were paired with their corresponding input orthographic forms and converted to back into hangul, and presented to a native speaker of Seoul Korean to judge the acceptability of the output as the pronunciation of the orthographic form. Of the 792 output forms, all but 55 were judged as acceptable. That is, on these forms, the LTS rules achieved 93.05% accuracy. Of the 55 unacceptable forms, 49 were Sino-Korean words in which correct pronunciation involves either over-application of tensification, or exceptional morphological rules such as 'n-insertion' (which we had not implemented in our LTS module). The other 6 forms were recent loan words with 's' in the loan source. The spelling of words such as /nyu-seu/ (from English *news* ignores the phonological rules governing which of the two fricatives is used to pronounce the form. Thus, *news* is spelled /nyu-seu/, with an aspirated /s/, but pronounced as /nyu-sseu/, with a fortis /ss/. (The spelling conventions probably reflect the fact that the *jamo* for the fortis fricative is a doubling of the *jamo* for the aspirated fricative.)

In future work, we plan to augment our existing rules with rules for Sino-Korean words, and expect to use machine learning techniques for this part of the problem. For example, we suspect that many of *eojeol* to which tensification should not apply are lower-frequency sequences of several shorter Sino-Korean compounds (comparable to such English compounds as *law degree requirement changes*. If this guess is correct, then specifying the domain of application could be a nice case study for the use of statistical models in large corpora, such as our list of 1.3 million *eojeol* and the context sentences from which they were extracted.

5. REFERENCES

[1] Paul Taylor, Alan W. Black, and Richard Caley, "The architecture of the festival speech synthesis system," in *3rd ESCA Workshop on Speech Synthesis*, Jenolan Caves, Australia, 1998, pp. 147–151.

[2] Ken Lunde, *CJKV Information Processing*, 0'Reilly, 1999.

[3] Linguistic Data Consortium (LDC), *Korean Newswire Text Corpus*, 2000, catalog number LDC2000T45,ISBN 1-58563-168-X.

[4] Sun-Ah Jun, "The accentual phrase in the korean prosodic hierarchy," *Phonology*, vol. 15, no. 2, pp. 189–226, 1998.