

# A linguistically-motivated approach to grapheme-to-phoneme conversion for Korean

Kyuchul Yoon and Chris Brew

*Department of Linguistics, The Ohio State University, Columbus, Ohio, USA*

---

## Abstract

This paper describes a hand-written rule-based grapheme-to-phoneme (GTP) conversion system for Korean built within the Festival text-to-speech (TTS) synthesis framework. The core of the GTP conversion system is a simple implementation of nine linguistically-motivated morphophonological rules. These rules, which are well known to students of Korean linguistics, were implemented in Festival rewrite formalism, and were applied to 1.3 million distinct orthographic words (space-delimited *eojeols*) from the Korean Newswire corpus. The outputs were evaluated against a representative subset of *eojeols*. The subset was examined by three native speakers of Korean, who judged 91.17% of the word types in a stratified sample of Korean *eojeols* to be acceptable pronunciations, which means that our system converted 99.63% of the grapheme tokens correctly. This performance is comparable to that obtained from earlier studies such as Kim et al. [2002] which, contrary to our system, used an elaborate morphological analysis module. This is evidence of the potential benefit of well-abstracted linguistic knowledge. In addition, because our approach is based on well-known linguistic principles, error analysis is fairly straightforward. Straightforward error analysis is an essential step in knowing what features are likely to be informative in training a hybrid system where exceptions to rules are handled by a machine-learning component.

*Key words:* Korean, grapheme-to-phoneme, rule-based, Festival TTS framework

---

## 1 Introduction

### 1.1 Building GTP rules

Grapheme-to-phoneme conversion (the task of assigning a phonemic representation of an orthographic string) can rely either on hand-written rules

or on learning from data. Both methodologies are reasonable, because morphophonology is typically shot through with subregularities and partially applicable rules. For example, in English, many nouns ending in “f” (including “thief”) have their plural in “ves”, but others (including “chief”) have their plural in “fs”<sup>1</sup>. Because of such phenomena we do not expect a rule-based account to be entirely elegant and exception-free, and we do not expect the models created by a machine learning approach to be completely compact and perspicuous. For Korean, where, as we will see, there are phenomena analogous to the “chief”/“thief” dichotomy, we chose to first explore the potential of the rule-based approach. We have achieved state of the art performance by this route, although we see this only as the first step toward a better-than-state-of-the-art solution that will be some form of hybrid.

The Festival rewrite formalism is loosely based on classic Generative Phonology, with ordered rules and the possibility of rule-rule interactions. This places the obligation of managing rule ordering squarely on the designer of the rules. Because the relevant rules of Korean are relatively regular, well-understood and few in number, at least when compared to the corresponding rules of languages like English, we did not find this obligation too onerous.

Kim et al. [2002] suggest a two stage process which combines a relatively complex treatment of graphemes occurring at morpheme boundaries with a simpler one for morpheme-internal graphemes. The latter component is learnt from data, but the former consists of a database of hand-written “phonetic patterns”. The idea is that each phonetic pattern indicates a potential conversion of a boundary grapheme, but that these conversions come with side conditions (“morphophonemic connectivity rules”) on the conversion of adjacent segments. The conversion rules are initially applied non-deterministically, then a constraint satisfaction process is applied to ensure that a globally consistent interpretation is chosen.

To our understanding, what these “morphophonemic connectivity rules” are doing is to apply Korean morphophonological rules to each item in the phonetic patterns dictionary. Our system does similar things, but the main difference between our approach and that of Kim et al. [2002] is that they provide large numbers (1992 if we understand correctly) of highly specific rules for morpheme connectivity, whereas we have rules that are stated in more abstract terms and that better reflect the body of detailed knowledge that has been accumulated about the morphophonology of Korean.

It is not clear how costly Kim et al. [2002]’s database of phonemic conversions was to develop. The development strategy that they seem to have adopted was to match the prescription of the guidelines provided by the Korean Ministry of

---

<sup>1</sup> Example due to Mark Liberman, posted to <http://www.languagelog.org/> post of 2 Jan 2004, title “The Theology of Phonology”

Education [Korea Ministry of Education and Human Resources Development, 1995] on a case-by-case basis. This is an ad-hoc response to the observation that the desired behavior at morpheme boundaries is more complex than that which occurs morpheme internally.

By contrast, the traditional account in Korean linguistics identifies a much smaller number of more general rules. We have implemented nine of the most important of these rules in the Festival rule formalism and achieve performance comparable to that of Kim et al. [2002].

We believe it is a strength of our rule-based approach and an argument against the machine learning approach that the former is more modular and uses different types of information in a more systematic, disciplined and interpretable way. Nevertheless, as described in section 3.4, there are aspects of the problem which do seem to merit a machine learning approach and we see our approach as just one part of developing a hybrid system that we are confident can cover more aspects than either approach can cover by itself. Certain properties of the Korean writing system justifies this confidence, in our view.

## 1.2 *Hangul and the Festival GTP rule system*

Korean has an alphabetic writing system, called *hangul*. For many languages with writing systems, a systematic relationship exists between the orthographic form of a word and its pronunciation. The relationship between the phonology and the orthography of a language depends on the type of writing system. In a phonographic writing system such as the one used for Finnish, the orthography represents the output of morphophonological processing, whereas in a morphographic writing system such as Korean, the orthography represents only the input to morphophonology. It is only in the latter case that the implementation of morphophonology is part of the job at the GTP module.

The Korean hangul writing system is an “alpha-syllabary” [Taylor, 1979]. It has symbols for consonants and vowels, which are grouped into orthographic syllables. At its inception in 1446, hangul appears to have been fully “phonographic”: words were written as they were pronounced, and pronounced as they were written. If a morpheme appeared in different contexts, and was realized differently, it would also have been written differently, as is the case for the English “in-” prefix, which surfaces in different forms (“*impossible*”, “*illegible*” and “*inconsistent*”, for example) according to its phonetic environment. However, the current hangul is “morphographic”: verb and noun bases are written in one constant shape, ignoring morphophonological sound changes [King, 1996], which the reader is now expected to reconstruct. The regularization of spelling that was adopted in 1933 as the Unified Orthography made

hangul somewhat less phonographic but because this spelling reform was so recent, the morphophonological processes that the GTP system must implement are fairly simple, and many orthographic texts can be translated into a reasonably accurate phonemic encoding (which is also, by grace of the elegant writing system, likely to be accurately syllabified without substantial further effort).

Traditional Korean texts interspersed hanja (Chinese characters) for Sino-Korean words with hangul for native Korean words. However, most Korean texts today are written almost entirely in hangul, with arabic numerals and some strings (e.g., URLs and such) in the roman alphabet interspersed. Therefore, hangul conversion covers nearly the whole task of grapheme-to-phoneme conversion. Although the task of providing Sino-Korean readings for hanja is of independent interest, and will possibly prove relevant to our interests in understanding the mental representation of different lexical strata, we do not address it here.

We believe that a proper consideration of prosody will be essential to high-quality Korean GTP. Experimental studies have shown that many aspects of speech are affected by a kind of “prosody-driven” allophony. Yoon [2003], for example (see also Jun [1998]), examined the same syllable, i.e., two Korean coronal fricatives followed by /a/, in different prosodic positions and found that with the segmental context held constant, prosody had a consistent effect on the allophonic realization of the target segments, but this variation is not something which can be captured by the usual manipulations of pitch and duration common in TTS systems. Therefore, since we eventually plan to experiment with prosodically conditioned unit selection as a strategy for speech synthesis, it is important to us to organize the GTP system in such a way that information about the prosodic hierarchy can be coherently integrated with the rest of the information that TTS needs to infer from the orthography.

This paper describes the design of a Korean GTP subsystem built within the Festival TTS framework, along with the testing of the GTP rules against a wordlist derived from a large corpus. Our implemented system has two components. In the first step, a Korean text in the morphophonemic representation inherent in the native orthography is deterministically mapped into the orthography required by the Festival GTP rules, an ASCII transliteration which preserves the syllable boundaries represented in the native orthography.

The second step is an implementation of the most important morphophonemic rules of Korean. This second component needs to be sensitive to segmental context, syllable position, and morpheme boundaries. It adjusts both segmental composition and syllable boundaries to produce a phonemic representation of the utterance that the synthesizer should actually speak. We implement and test the second component, and present a formal evaluation of its performance,

which is (as far as we are able to determine) state of the art. A contribution of our work is a careful and detailed evaluation of what currently works and what doesn't in Korean GTP.

The rest of this paper is structured as follows. Following an introduction to the Korean segmental repertoire and the ways in which relevant aspects of phonological structure are encoded in the orthography, the design of the system is motivated and the GTP rules are described and evaluated against a large wordlist. We then draw consequences from the evaluation, leading to suggestions for future work.

## 2 Methods

### 2.1 Korean phonology relative to orthography

Modern Seoul Korean has seven<sup>2</sup> vowel segments and twenty one consonant segments, which are listed and described in Table 1.

In the hangul writing system, sets of jamo (orthographic phoneme segments) are grouped into eumjeol (orthographic syllables), as shown in Figure 1. Sequences of eumjeols are grouped into eojeol (space-delimited orthographic “words”). In careful styles (such as reading newspapers aloud), each eojeol corresponds roughly to a single accentual phrase (AP), a tonally demarcated unit in Korean prosody<sup>3</sup>. Therefore, the GTP rules and the definitions of concatenative units can refer to syllable boundaries. (Recall however that the syllable boundaries presented in the orthography may not match those of the final output, see Figures 1 & 3.) In the same vein, GTP rules and the definitions of concatenative units can also refer to the preliminary AP boundaries produced by our heuristic assumption that orthographic words correspond to APs (see Figures 2 & 3).

Figure 3 shows 2 eojeols (7 eumjeols) extracted from the text in Figure 2, with the syllabified transliteration in ISO TR 11941:1996 (i.e., the input to the GTP rules proper), the pronounced syllables (i.e., output of the GTP rules), and an IPA transcription of this pronounced phoneme sequence. Note the “tensification” of the lenis stop onset in syllable 6 and the “resyllabification” of

---

<sup>2</sup> We collapsed the mid-front and low-front vowel into /e/ reflecting the tendency that most younger Seoul speakers do not distinguish them any more.

<sup>3</sup> As in most languages whose prosody has been studied, Korean has several levels of prosodic domain. AP is the accentual phrase. It is smaller than the IP (intonational phrase) and larger than the prosodic word (PW). See [Jun, 2000] for further details of Korean prosody.

**밤** /bam/ 'night'

Here the *jamo* ㅂ /b/ is the onset of the syllable (a stop released with lenis burst).

**삽** /sab/ 'shovel'

Here the *jamo* ㅂ /b/ is the coda of the syllable (voiceless closure with no release).

Fig. 1. The phonetic realization of /b/ in different syllabic positions.

*Original hangul text:*

정통부는 이와함께 초고속인터넷 장애발생시 손해배상 요건을 현재 4시간 이상에서 3시간 이상으로 강화해 해당 시간요금의 3배 이상 손해배상을...

*Text transliterated in ordinary ISO TR 11941:1996:*

jeongtongbuneun iwahamgge cogosoginteones jangebalsengsi **sonhebesang yogeoneul** hyeonje 4sigan isangeseo 3sigan isangeulo ganghwahe hedang siganyogeumeui 3be isang sonhebesangeul...

*Text transliterated with eumjeol and eojeol boundaries marked with - and ( ) respectively:*

(jeong-tong-bu-neun)(i-wa-ham-gge)(co-go-sog-in-teo-nes)(jang-e-bal-seng-si)(**son-he-be-sang**)(**yo-geon-eul**)(hyeon-je)(4-si-gan)(i-sang-e-seo)(3-si-gan)(i-sang-eu-lo)(gang-hwa-he)(he-dang)(si-gan-yo-geum-eui)(3-be)(i-sang)(son-he-be-sang-eul)...

Fig. 2. Example text (7-4-2002, 인터넷 한겨레신문 Newspaper Internet Hankyoreh).

손	해	배	상	요	건	을	
(son-	he-	be-	sang)	(yo-	geon-	eul)	<b>GTP input</b>
(son-	he-	be-	sang)	(yo-	ggeo-	neul)	<b>GTP output</b>
[s <sup>h</sup>	he	pe	s <sup>h</sup> aŋ]	[jo	k'ɾ	ni]	<b>IPA</b>

Fig. 3. The section highlighted in Figure 2

its coda nasal. However, the GTP rules do not encode sub-phonemic variation, such as the “deaspiration” of the AP-medial [s<sup>h</sup>] in syllable 4, which typically has a shorter aspiration interval than an AP-initial [s<sup>h</sup>], as in syllable 1.

Once the initial set of morphophonological GTP rules is in place, it can be used to transliterate any large on-line text database, such as the Korean NewsWire [Linguistic Data Consortium, LDC]. It is then possible

- (1) to iteratively test, explore and improve the GTP rules (for example, by finding statistical patterns that distinguish different etymological strata).
- (2) to choose a subset of the corpus to record and transcribe with the K-ToBI prosodic labeling conventions [Jun, 2000] in order to test hypothe-

ses about the effects of syllable and higher-level prosodic position on segments.

- (3) to design a small corpus of reasonably familiar phrases that can be recorded to extract positionally-specific diphones and larger concatenative units.

## 2.2 Morphophonology of Korean

An example of a morphologically conditioned alternation is that between the alveolar fricative [s] and the lenis stop [d] in the word /os/ ‘garment’ in the following example: (hyphen in the transcription makes explicit orthographic syllable boundaries that are built in to the hangul writing system.)

/os-e/ 옷에	[ose] ‘on the garment’ (/e/ locative particle is a bound morpheme)
/os-an/ 옷안	[odan] ‘inside of garment’ (a compound noun. /an/ ‘inner’ is a free morpheme)

The alternation between [s] and [d] for the final segment of the morpheme /os/ is comparable to the alternation between [d] and [t] for orthographic *ed* in English *mobbed* versus *mopped*. That is, it involves not only a change in sound but a change in the name of the segment. If Korean had adopted a writing system similar to that of Finnish, this difference would have been encoded in the orthography. Korean orthography neutralizes the distinction between [s] and [d] realizations of the final /s/ of /os/. The reader is expected to recover the distinction from implicit knowledge of the general principle of coda neutralization. In just the same way, English orthography neutralizes the distinction between [d] and [t] realizations of the past tense morpheme /d/. Here the general principle is that tautosyllabic [pd] sequences are never allowed. Once again, this is part of the implicit knowledge of the competent speaker. For the GTP system to be effective, this knowledge must be made available to it in some form. A Korean system that wants to get the [s] [d] distinction right has to be sensitive to the fact that coda neutralization applies to /os-an/ but not to /os-e/. It can do this if it “knows” that /an/ is a free morpheme, whereas /e/ is bound.

## 2.3 The “phonetizer” (GTP preprocessor module)

Because hangul is a fairly transparent alpha-syllabary, any hangul text can be transliterated into the roman alphabet in a straightforward way. We use the ISO TR 11941:1996 romanization standard. There is an almost one-to-one

correspondence between jamo in the hangul text and roman letters in this transliteration.

We adapted a transliteration program (written by Nick Cipollone) that produces output appropriate for use with the GTP rule programming language of Festival. This “phonetizer” is a table-driven program that maps the two-byte code for each hangul syllable to a romanized representation. The output of the phonetizer is the input to the GTP rules proper. The same segmental repertoire is used for both the input and the output of the GTP rules. This phonetizer could serve as the core of a more elaborate text analysis module that also deals with abbreviations, and such non-hangul characters as numerals, foreign names and other words in the roman alphabet, the odd hanja. For present purposes we don’t need that.

Many of the GTP rules refer to a phone’s position in the (orthographic or “underlying” morphophonemic) syllable, as well as to the surrounding segments and the morphological structure. Since the orthography directly encodes syllabification, it was easy to adjust Cipollone’s code so as to output that information along with the segments. We would not object if writing systems for other languages made syllabification this easy.

The domain of application of most of the morphophonemic rules that we implemented is the “surface” AP. For example, the post-obstruent tensing rule is known to apply only within APs [Jun, 1998]. But some rules such as the aspiration of lenis obstruents adjacent to /h/, nasalization of obstruents, and some varieties of tensification of lenis obstruents are also known to apply in larger scale units, certainly larger than the space delimited orthographic word. Then again, palatalization is sensitive to word boundaries [Lee, 1996]. Therefore in an ideal situation, the GTP rules need to be aware of many details of prosodic grouping at a variety of scales. However, the current version of our system does not have a full model of phrasing, so we need, for the moment, a simpler model to serve as the basis for the GTP module. In the deliberate style that we judge appropriate for reading newspapers, the AP corresponds roughly to the grouping of eumjeols into eojeols as illustrated in Figure 4. This means that we are able to build a simple but workable phrasing model by referring only to the punctuation marks and spaces that delimit the eojeol. Future work will include the creation of a more elaborate phrasing model, which will be particularly necessary once we move beyond the deliberate newspaper style.

#### *2.4 Implemented morphophonological rules (the GTP module proper)*

The basic form of a grapheme-to-phoneme rule in the Festival system is designed to capture the segmental context. That format is:

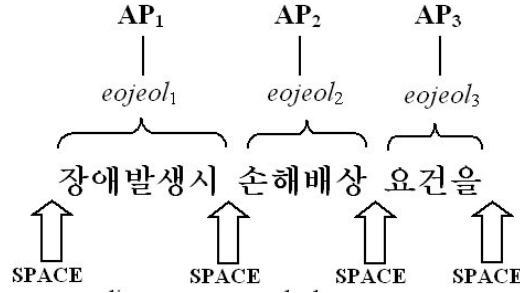


Fig. 4. Eojeols corresponding to accentual phrases.

( LEFTCONTEXT [ ITEMS ] RIGHTCONTEXT = NEWITEMS )

For example, consider the rule of palatalization that changes /t/ into [c] in the following example,

/sot-i/ 솔이                      →            [so-ci]  
 transliterated form                      pronounced form

This rule would be encoded in the Festival rule language as:

( o [ t - ] i = - c )

where ‘-’ is a representation of syllable boundary in Festival. We use a variety of such “pseudo-phones” to encode the boundaries of prosodic constituents in the Festival input and output languages. This is an *ad hoc* device to get the information that we want to express into the form Festival can use, but it does the job.

Some rules (e.g. neutralization of coda obstruents) are sensitive also to the bound/non-bound status of the following morpheme. Eventually, we plan to make the rules refer to this kind of morphological context directly by developing a morphological parsing module that will analyze forms and mark morpheme boundaries and types before they are input to the morphophonological rules. The *ad hoc* solution adopted in our system is to enumerate bound morphemes, such as nominative and topic markers ‘-i’ (이), ‘-ga’ (가), ‘-eun’ (은) and ‘-neun’ (는), at the beginning of each rule set to capture this sensitivity to the 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.

The following is a list of morphophonological rules [Lee, 1996] that we have implemented. Although this is not an exhaustive list of all the morphophonemic rules of Korean, it covers most of the rules that involve allomorphic variation. We excluded allophonic rules, such as lenis stop voicing, because these

would force us to complicate the phonemic repertoire. Instead, the segmental variations caused by such allophonic rules will be handled by appropriate diphone selection (this is future work). In the examples below, the free morphemes of the input are separated by a plus sign, and a bound morpheme is separated from its host by a period, and a morpheme-internal orthographic syllable boundary is indicated by a hyphen. In the output, hyphens stand for syllable boundaries in the pronounced form.

### Palatalization

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

### Neutralization of coda obstruents

/bu-eok/ 부엌 ‘kitchen’ → [bu-eog]

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

contrast) /os.e/ 옷에 ‘garment + on’ → [o-se]

(See section 2.1)

### Liaison

/ap.i/ 앞이 ‘front + nom.’ → [a-pi]

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

### Aspiration of lenis obstruents adjacent to /h/

/beob-hag/ 법학 ‘study of law’ → [beo-pag]

/nah-da/ 낳다 ‘give birth’ → [na-ta]

### Tensification of lenis obstruents

/beob-de/ 법대 ‘law school’ → [beob-dde]

/bat/ + /gil/ 밭길 ‘farm-trail’ → [bad-ggil]

### Nasalization of obstruents

/gug-min/ 국민 ‘citizens’ → [gung-min]

/gabs.man/ 값만 ‘price alone’ → [gam-man]

### Simplification of coda clusters

/gabs-jin/ 값진 ‘precious’ → [gab-jjin]

/gabs.i/ 값이 ‘price + nom.’ → [gab-ssi]

/gabs/ + /in-sang/ ‘price + hike’ → [ga-bin-sang]

값인상

### Nasalization of /l/

/sib.li/ 십리 ‘ten + unit of distance’ → [sim-ni]

### Lateralization of /n/

/ceon.li/ 천리 ‘thousand + unit of distance’ → [ceol-li]

/kal/ + /nal/ 칼날 ‘knife-blade’ → [kal-lal]

The Festival GTP framework is an engineering solution rather than a state-of-the-art phonological formalism. One consequence of this is that rule ordering becomes an important part of the grammar designer’s task. Some rules interact with each other, e.g. the liaison should apply after the rules such as the palatalization and the neutralization of coda obstruents (see Table 2) and the palatalization should apply before the coda neutralization. The GTP module

proper applies one rule after the other to the input eojeol to reflect the rule ordering. Rules such as the consonant cluster simplification and coda neutralization were split to subcomponents to deal with cases of similar conditioning environments, making the maintenance of the rules easier, e.g. one subcomponent of the coda cluster simplification rule deals with eojeols ending with consonant clusters (subcomponent 3 in Table 2) whereas another subcomponent deals with eojeols whose output involves resyllabification (subcomponent 2 in Table 2).

### 2.5 *Sino-Korean words*

We suspected ahead of time that words of Sino-Korean would be an important test case for our system. It is well known that the pronunciation of Sino-Korean words is sensitive not only to the segmental context, but also to other factors. There is a rough parallel between the status of Sino-Korean words and the status of English compounds of Latinate origin: in both cases it is reasonable to suppose that speakers are explicitly or implicitly aware of the existence of the different lexical strata, and that there may be important differences in the way in which GTP rules apply. Relevant factors such as the frequency of the compound relative to the morphemes, and semantic transparency. For example, the orthographic syllable /byeong/ (병) can be either of the homographic morphemes ‘illness’ and ‘bottle’ and both have two pronunciations. In older, less transparent compound words /hwa-byeong/ (화병) ‘vase’, /ji-byeong/ (지병) ‘chronic terminal illness’ and /ho-li-byeong/ (호리병) ‘genie’s bottle’, the eumjeol is pronounced [byeong] whereas in more transparent compounds such as /kol-la-byeong/ (콜라병) ‘coke bottle’, /hwa-byeong/ (화병) ‘disease caused by anger’ and /heo-li-byeong/ (허리병) ‘backache’, the initial segment is tensified and the eumjeol is pronounced [bbyeong].

Our system does not have access to the relevant distinctions of etymology and frequency profile, so is certainly going to fail in some of these cases. It is a matter for further research to arrive at a deeper understanding of exactly which distinctions are relevant, and to encode them into an improved GTP system.

## 3 Evaluation

We evaluated the GTP rules by applying them to a large list of attested eojeols extracted from a online text corpus in the style that we are initially targeting for the TTS system (namely, newspaper reading) and then selecting a representative subset of eojeols for examination by three native speakers of

Seoul Korean, including the first author. All were graduate students at the linguistics department of the Ohio State University.

### 3.1 Materials

We first extracted and counted the number of occurrences of each of the 1.3 million distinct eojeols in the more than 143,000 articles of the Korean Newswire corpus. Note that this corpus is not a lexicon, but a list of eojeols, corresponding approximately to accentual phrases. Thus, it tests the GTP rules for different pronunciations of the same syllable in different lexical and phrasal contexts. For example, the lexical item /beob/ ‘law’ occurs in the following eojeol (among others):

<i>Input to GTP</i>		<i>Output from GTP</i>	<i>Rules applied</i>
/sa-beob-bu/	사법부 ‘Department of Justice’	→ [sa-beob-bbu]	tensification
/sa-beob-bang-he/	사법 방해 ‘Obstruction of justice’	→ [sa-beob-bbang-he]	tensification
/bang-beob-i/	방법이 ‘method -nom.’	→ [bang-beo-bi]	resyllabification

We first applied the GTP rules to all of these eojeols and tabulated the first stage results by the number of GTP rules that had applied to produce the output, by the length (in number of syllables) of the output “phonemicized” spelling, and by the frequency in the Newswire corpus of the (input) eojeol (see Table 3). We tabulated the eojeols in this way, because (1) the output is likely to be most reliable for eojeols in which morphemes are pronounced exactly as they are spelled, and less reliable when rules interact and (2) long, low frequency eojeols are more likely to be phrases or loosely conjoined compounds, where the tensification rule will be misapplied.

We selected a stratified random sample of 792 words, 66 from each of the twelve cells in Table 3, to evaluate the GTP rules efficiently across the full range of lengths, frequencies, and number of rule applications. Each of the 792 output forms was converted back into hangul and presented to three native speakers of Seoul Korean along with the original hangul spelling of the eojeol, to judge the GTP output form as a pronunciation of the input. We asked the speakers to mark unacceptable syllables.

### 3.2 Quantitative results

The results of the evaluation by three native speakers of Korean are given in Table 5. The percentage of eojeol tokens rated as fully correct in all respects was 91.17%.

Previous work on Korean [Kim et al., 2002] reports the performance on a corpus of sentences from the MBC NewsDB database. The results are shown in Table 4.

Their evaluation differs from ours in the following ways:

- They rely on an extensive exceptions dictionary
- We don't have access to a morphological analysis.
- Their test set is a large corpus in which eojeols may be repeated, whereas ours is a list of eojeols.

It is future work to address the first two differences, but the third difference can be addressed now. We estimated the performance that we would have seen by testing on the whole corpus by weighting each of our 792 test eojeols by the number of occurrences of that eojeol in the whole corpus. This gives a frequency-weighted word error rate of 97.10% correct. If we measure the error rate by individual graphemes (roughly comparable to the phoneme correctness rate above) we find that 99.16% of the graphemes in the list of eojeols are correctly converted. The corresponding frequency-weighted figure for the whole corpus is 99.63%.

Our evaluation has both strengths and weaknesses compared to that of Kim et al. [2002]: a strength is that our gold standard is based in the very secure intuitions of three trained Korean linguists; a potential weakness that we evaluated only on a representative sample of the eojeols in the corpus. To the extent that the sample is representative (and we have no reason to think otherwise) it seems a fair conclusion that Kim et al. [2002]'s system and ours are performing at comparable levels. A complete listing of eojeols that were classified as errors by the evaluators is given in the Appendix. As expected, Sino-Korean compound words were the primary source of errors in our GTP output.

Our performance figures validate the design choices that we made. The difficulties that the system has with Sino-Korean words and recent borrowings with /s/, while expected, do show the need to develop an independent morphological module and to refine our GTP rules. They also tell us where we should focus our efforts in these two future tasks. A morphological module will give us more control over the application of some of the morphophonological rules that were implemented in our system. We know from phonology texts

that such rules as the tensification of lenis obstruents, liaison, neutralization of coda obstruents, etc. are sensitive to the morphological status of the input words. The full details of how this should work in practice are as yet unknown.

### 3.3 Error analysis

Although the performance of our system is generally good, achieving this low overall error rate is not our only goal. An important benefit of our linguistically-motivated approach is that the simple structure of the rule system gives us a natural taxonomy for the errors made by the system. The following sections use this taxonomy to offer analyses and make suggestions for future improvement.

#### 3.3.1 Tensification

Here are some of the problematic sets of (completely or partially) homographic *eojeol* from the corpus that our GTP module does not yet handle correctly. We will see that these errors are partly to be ascribed to the effects of factors that our system does not yet consider, but perhaps could.

Set (1)			
sam-gweon	삼권	→	[sam-gweon] ‘book three’
		→	[sam-ggweon] ‘three rights’

The problem here is that the tensification rule should not apply in ‘book three’, where /gweon/ ‘book’ acts as a classifier (i.e., a quasi-particle).

Set (2)			
gwan-lyeon-ju-ga		→	[gwal-lyeon-ju-ga] ‘related bond + nom.’
관련주가		→	[gwal-lyeon-ju-gga] ‘related bond price’

For this example, the tensification rule should apply in the compound ‘related bond price’, but it will be blocked by the homophony with /-ga/ (nominative particle).

Set (3)			
min-sa-beob	민사법	→	[min-sa-bbeob] ‘civil law’
sa-beob	사법	→	[sa-beob] ‘judicial’

Here the tensification rule should apply in the semantically transparent compound (where /min-sa/ itself is a compound, meaning ‘people’), but not in /sa-beob/ ‘judicial’.

Set (4)

seo-ul-beob-de → [seo-ul-**beob**-dde] ‘Law School of Seoul Nat’l Univ.’

서울법대

The tensification rule should not apply for this word at the boundary between the proper name Seoul and the compound meaning ‘law school’. One thing to note here is the role played by the morphosyntactic structure of the Sino-Korean compound words. The word-internal morphosyntactic structure is /seo-ul # beob-de/ (# for morpheme boundary) and the first segment of the second component morpheme does not tensify, whereas in such compounds as /seo-ul # beob/ ‘law of Seoul’ the same segment would be tensified as [seo-ul-bbeob]. This suggests that the level of morphosyntactic grouping within Sino-Korean compounds interacts with the rule of tensification.

Note that while forms such as /sa-beob/ and /min-sa-beob/ might both be listed in a large online dictionary, it is not practical to list all possible Sino-Korean words, including highly technical or newly-coined compounds, because the creation of new Sino-Korean compounds is a highly productive process, so the list would never be comprehensive. Some of the example words from the corpus are given below.

gun-sa-beob-jeong → [gun-sa-**beob**-jjeong] ‘military court’

군사법정

su-sa-beob-gwan → [su-sa-**beob**-ggwan] ‘judicial officer of investigation’

수사법관

hwe-sa-beob → [hwe-sa-**bbeob**] ‘company law’

회사법

yag-sa-beob → [yag-ssa-**bbeob**] ‘law of pharmacists’

약사법

an-lag-sa-beob → [al-lag-ssa-**bbeob**] ‘law of euthanasia’

안락사법

Dependent nouns such as /geos/ (것) and /su/ (수) cannot occur without the support of a preceding clause or a demonstrative. The initial segment of these two dependent nouns should be tensified in order to sound natural if preceded by an eojeol that ends with an /l/. Despite their dependency on the preceding phrases, they are not written as part of the preceding word, but they start their own eojeol with a preceding space. They can form part of an accentual phrase with the preceding eojeol. They can also start their own

phrase. This is an impossible challenge for the current formulation of our GTP rule of tensification, which takes an *ojeol* as the domain of rule application, because there is no conditioning segmental context which can trigger the rule. This is another indication that a morphological module should really precede our GTP module.

### 3.3.2 /s/-borrowings

The next major source of errors was recent borrowings from other languages. The words contain orthographic /s/'s corresponding to 's' in the loan source. In forms such as /nyus/ (뉴스) 'news', where there is a voiceless alveolar fricative in the loan source, the fricative can be nativized either with /s/ (the aspirated [s<sup>h</sup>]) or (more typically) with /ss/ (the tense [s']). Problems arise because the fricative is always rendered in *hangeul* as aspirated /s/ rather than as tense /ss/, although in many loanwords with a following /i/ or /e/, it is the tense that is pronounced as some actual examples show.

seu-po-ceu	스포츠	→	[seu-po-ceu]	'sports'
po-seu-teu-men	포스트맨	→	[po-seu-teu-men]	'postman'
seul-lo-u-peu	슬로우프	→	[seul-lo-u-peu]	'slope'
peo-sen-teu	퍼센트	→	[peo-sen-teu]	'percent'
si-seu-tem	시스템	→	[ssi-seu-tem]	'system'
se-il	세일	→	[sse-il]	'sale'

Preceding segments may affect the identity of the fricative.

peol-seu	펄스	→	[peol-sseu]	'pulse'
teu-len-seu-a-si-a	트랜스아시아	→	[teu-len-sseu-a-si-a]	'trans-Asia'

A closer look at some native words suggests that an etymological factor is also at play. For example, the same syllable /se/ is pronounced [se] in /se-in/ 'people' whereas it is pronounced [sse] in /se-il/ 'sale'.

se-in	세인	→	[se-in]	'people'
se-myeon-sil	세면실	→	[se-myeon-sil]	'washroom'
si-seung	시승	→	[si-seung]	'test drive'

Morpheme boundary information also seems to be playing a role.

geol-seu-ka-u-teu	걸스카우트	→	[geol#seu-ka-u-teu]	'girl scouts'
peol-seu	펄스	→	[peol-sseu]	'pulse'

The above examples appear to suggest that the /s/-borrowings are formally similar to the Sino-Korean words. That is, segmental context is not sufficient to distinguish the correct utterance form and an etymological factor contributes to the output pronunciation.

### 3.3.3 /n/-insertion

The rule of /n/-insertion, which was not implemented in our system because of its sensitivity to factors other than contextual segments, was another source of errors. /n/-insertion is sensitive to the status of the component morphemes in a compound noun. In general, the second morpheme should be a free morpheme that begins with a vowel or approximant (shown as underlined below).

bun-dam# <u>yag-sog</u>	→	[bun-dam-nyag-ssog]	‘promise of (e.g., work) division’
분담약속			
nam# <u>yu-leob</u>	→	[nam-nyu-reob]	‘south Europe’
남유럽			
ab-seung# <u>ye-sang</u>	→	[ab-sseung-nye-sang]	‘expectation of a landslide victory’
압승예상			
bi-nan# <u>yeo-lon</u>	→	[bi-nan-nyeo-ron]	‘public opinion of criticism’
비난여론			

As the following examples show, if the second morpheme is not a free morpheme or the word is not a compound word, /n/-insertion does not apply [Lee, 1996].

eo-lin#i	어린이	→	[eo-li-ni]	‘child’
jeolm-eun#i	젊은이	→	[jeol-meu-ni]	‘young man’
seom-yu	섬유	→	[seo-myu]	‘fiber’
cam-yeo	참여	→	[ca-myeo]	‘participation’

As the rule formulation shows, an obligatorily free morpheme that starts with a /y/ should be preceded by a morpheme boundary, which is in turn preceded either by any of the three nasal sound or by any of the three voiced lax stops. In addition, any vowel can follow /y/ except for /i/. The rule can be formulated as below (# indicates a morpheme boundary).

$$\phi \rightarrow n / m|n|ng \# \text{---} yV, \text{ where } V \neq i \quad \text{--- (1)}$$

$$\phi \rightarrow n / b|d|g \# \text{---} i \text{ or } yV, \text{ where } V \neq i \quad \text{--- (2)}$$

Rule (2) feeds into the rule of nasalization so that the pre-morphemic boundary segments [b, d, g] turn into homorganic nasals [m, n, ng]. Here are some

examples.

de-hag#ya-gu 대학야구	→	[de-hang- <b>nya</b> -gu]	‘college baseball’
cu-seog#yeon-hyu 추석연휴	→	[cu-seong- <b>nyeon</b> -hyu]	‘Cu-seog holidays’
a-lab#yeon-meng 아랍연맹	→	[a-lam- <b>nyeon</b> -meng]	‘Arabic alliance’
ggoc#yeo-leum 꽃여름	→	[ggon- <b>nyeo</b> -reum]	‘flowering Summer’

The last example /ggoc-yeo-leum/ (꽃여름) tells us that the /n/-insertion rule interacts with the rule of coda neutralization. The first component free morpheme /ggoc/ is followed by another free morpheme, and this triggers coda obstruent neutralization so that the final /c/ first changes into /d/, which provides the appropriate input to the /n/-insertion. Examples are given below.

<u>ggoc</u> #ip	꽃잎	→	[ggon-nib]	‘flower petal’
<u>bat</u> #il	밭일	→	[ban-nil]	‘farm work’

### 3.3.4 /n/-lateralization and /l/-nasalization

The lateralization of /n/ and nasalization of /l/ also require a morphological analysis for correct application of the rule. In the first example /yeon-lyo/ (연료), the /n/ should be lateralized when followed by an /l/, but not across the morpheme boundary. The /lyo/ (료) in /gang-yeon-lyo/ (강연료) is a bound morpheme meaning ‘fee’. However, when the /n/ is preceded by an /l/ across the morpheme boundary as in /kal-nal/ (칼날) and /mul-nan-li/ (물난리), lateralization should occur.

yeon-lyo	연료	→	[yeol-lyo]	‘fuel’
<u>gang-yeon</u> #lyo	강연료	→	[gang-yeon-nyo]	‘lecture fee’
<u>kal</u> #nal	칼날	→	[kal-lal]	‘knife-blade’
<u>mul</u> #nan-li	물난리	→	[mul-lal-li]	‘(water) flood’

### 3.4 The need for morphological analysis

In presenting these examples of tensification, /s/-borrowings, /n/-insertion, /n/-lateralization and /l/-nasalization, we have identified several places in which correct morphological information would be useful to the GTP mod-

ule, but it remains an open question whether such information is currently available in practice. Many decisions made by the morphological analyser will be irrelevant to GTP, so good or bad performance in terms of global error rate will not necessarily translate into corresponding effects on the error rate of GTP. It nevertheless seems useful to briefly review the state of the art in morphological analysis and assess its potential for our purposes.

Cha et al. [1997] did a study on statistical and rule based hybrid POS tagging and Hong et al. [1996] tried a dictionary/rule-based morphological analyzer as a part of their speech translation system. In both cases we may assume that the morphological component was of some benefit, or it would not have been used. Kim et al. [2002] combined dictionaries and statistically learned rules to provide a morphological component for a Korean grapheme-to-phoneme converter. This certainly captures a degree of implicit morphological knowledge. However, they relied upon an dictionary with 2,894 entries to cover cases that their rules would otherwise fail to handle. Given the high productivity of Sino-Korean words in forming compound words and the more or less unlimited number of potential foreign words with /s/ that could be borrowed in the future, we feel that an exception dictionary can be no more than a temporary solution. We believe that a better approach is a hybrid system in which a rule-based component similar to the one that we have developed complements a fall-back system based on machine learning.

Less directly, Sarker and Han [2002] implemented an LTAG (lexicalized tree adjoining grammar) based statistical morphological analysis module trained on the Penn Korean TreeBank [Han et al., 2002]. This achieves 95.78% precision and 95.39% recall. This is an encouraging result for us, since it suggests that fairly accurate morphological analysis is available, but, as indicated above, we should probably curb our enthusiasm until we know just how well it is performing on the small subset of decisions which are actually needed for correct GTP. This is a matter for further research and the error analysis just shown allows us to approach that research efficiently. In comparing different morphological parsers, we can focus on the parses of types of words that are problematic from our GTP rules.

## 4 Summary

We have presented how Korean morphophonological rules can be implemented as hand-written GTP rules in the Festival Speech Synthesis System. We have defined a phoneset and used it to transliterate a large text corpus, which was subsequently processed through our GTP module. As the results indicated, Festival is a viable framework for implementing Korean morphophonology. Hand-written GTP rules were often (91.17% of the time) able to guess cor-

rectly how a word from a large text corpus would be pronounced in context. This is useful in itself, but also provides an empirical basis for decisions about what to record in service of our longer-term goal of building a synthesizer.

Our system’s handling of Sino-Korean and borrowed words is currently not on a par with the rest of the system, and needs further work. Although our hand-written rule-based GTP module cannot currently be said to outperform those learnt from data, we believe that our approach is a viable option given that it does not need any training materials. If we had used a machine-learning approach that lumped all the features together, we would not have been able to identify the common sources of all of the error types – namely, that these are all cases where morphological structure and not just segmental context are relevant. This analysis is implicit in the part-of-speech based approach of Kim et al. [2002], but has not, to our knowledge, previously been stated explicitly. An added benefit of our work was the insight to the nature of the GTP conversion task in Korean: we have shown that you can go most of the way with well-abstracted linguistic rules, but that in the end further components will be necessary to smoothly handle subregularities and exceptions. At some stage we will introduce an exception dictionary, but we are not ready to do so yet.

## A Appendix

A complete listing of eojeols that were classified as errors by the three evaluators<sup>4</sup>.

### 1. Sino-Korean words (involving tensification)

<u>Orthographic text</u> <u>Unacceptable GTP output</u> <u>(GTP input if different)</u>	<u>Corrected (underlined) forms</u> <u>&amp; glosses</u>
토지임대법은 to-ji-im-de-beo-beun (to-ji-im-de-beob-eun)	to-ji-im-de- <u>b</u> beo-beun <i>‘land-leasing law -top.’</i>
액션법 eg-ssyeon-beob (eg-syeon-beob)	eg-ssyeon- <u>b</u> beob <i>‘action law’</i>
개선법	

<sup>4</sup> *-nom.* for nominative case marker, *-top.* for topic marker, *-gen.* for genitive case marker, *-par.* for particle, *-acc.* for accusative case marker

ge-seon-beob	ge-seon-bbeob 'reform law'
불법폭력 bul-beob-pong-nyeog (bul-beob-pog-lyeog)	bul-bbeob-pong-nyeog 'illegal violence'
실시중인 sil-si-jung-in	sil-ssi-jung-in 'in progress'
실시한다면 sil-si-han-da-myeon	sil-ssi-han-da-myeon 'start-par.(if)'
양쯔강가 yang-jjeu-gang-ga	yang-jjeu-gang-gga 'yang-jjeu river shore'
국장급의 gug-jjang-geu-beui (gug-jang-geub-eui)	gug-jjang-ggeu-beui 'gug-jang rank -gen.'
시점에서의 si-jeo-me-seo-eui (si-jeom-e-seo-eui)	si-jjeo-me-seo-eui 'viewpoint -loc.-gen.'
평화중재활동에 pyeong-hwa-jung-je-hwal-dong-e	pyeong-hwa-jung-je-hwal-ddong-e 'peace intervention activity -loc.'
결승에는 gyeol-seung-e-neun	gyeol-sseung-e-neun 'final match -top.'
종합증권업을 jong-hab-jjeung-gweo-neo-beul (jong-hab-jeung-gweon-eob-eul)	jong-hab-jjeung-ggweo-neo-beul 'comprehensive security business -acc.'
정권부터 jeong-gweon-bu-teo	jeong-ggweon-bu-teo 'regime -par.'
증권거래 jeung-gweon-geo-le	jeung-ggweon-geo-le 'stock exchange'
주채권은행간에 ju-ce-gweo-neun-heng-ga-ne (ju-ce-gweon-eun-heng-gan-e)	ju-ce-ggweo-neun-heng-ga-ne 'between major creditor banks'
대우채권단	

de-u-ce-gweon-dan	de-u-ce-ggweon-dan 'Daewoo creditor group'
선수권대회는 seon-su-gweon-de-hwe-neun	seon-su-ggweon-de-hwe-neun 'championship tournament -top.'
생물권 seng-mul-gweon	seng-mul-ggweon 'biosphere'
관련주가 gwal-lyeon-ju-ga (gwan-lyeon-ju-ga)	gwal-lyeon-ju-ga or -gga 'related bond -nom. or related bond price'
폭탄사건은 pog-tan-sa-geo-neun (pog-tan-sa-geon-eun)	pog-tan-sa-ggeo-neun 'bomb accident -top.'
종합주가지 jong-hab-jju-ga-ji (jong-hab-ju-ga-ji)	jong-hab-jju-gga-ji 'overall stock value'
재평가차액은 je-pyeong-ga-ca-e-geun (je-pyeong-ga-ca-eg-eun)	je-pyeong-gga-ca-e-geun 'overall estimated margin -top.'
화염병과 hwa-yeom-byeong-gwa	hwa-yeom-bbyeong-gwa 'fire-bottle -par.'
고속철도 go-sog-ceol-do	go-sog-ceol-ddo 'high speed railway'
몰수됐다 mol-su-dwed-dda (mol-su-dwess-da)	mol-ssu-dwed-dda 'got confiscated'
몰두해 mol-du-he	mol-ddu-he 'preoccupied'
울상의 ul-sang-eui	ul-ssang-eui 'tearful face -gen.'
결성됐다는 gyeol-seong-dwed-dda-neun (gyeol-seong-dwess-da-neun)	gyeol-sseong-dwed-dda-neun 'organized -par.'

결손이어서 gyeol-so-ni-eo-seo (gyeol-son-i-eo-seo)	gyeol- <u>sso</u> -ni-eo-seo 'loss -par.'
활성화조치 hwal-seong-hwa-jo-ci	hwal- <u>sseong</u> -hwa-jo-ci 'measures of activation'
물가 mul-ga	mul- <u>gga</u> 'prices'
일대지진현장 il-de-ji-jin-hyeon-jang	il- <u>dde</u> -ji-jin-hyeon-jang 'site of earthquake'
과대평가하는 gwa-de-pyeong-ga-ha-neun	gwa-de-pyeong- <u>gga</u> -ha-neun 'overestimate -par'
사기사건 sa-gi-sa-geon	sa-gi-sa- <u>ggeon</u> 'fraud incident'
피살사건 pi-sal-sa-geon	pi-sal-sa- <u>ggeon</u> 'homicide'
통화안정증 tong-hwa-an-jeong-jeung	tong-hwa-an-jeong- <u>jjeung</u> 'currency stabilization symptom'
고의성이 go-eui-seong-i	go-eui- <u>sseong</u> -i 'intention -nom.'
방문성과 bang-mun-seong-gwa	bang-mun-seong- <u>gwa</u> 'result of a visit'
수출증가율 su-cul-jjeung-ga-yul (su-cul-jeung-ga-yul)	su-cul- <u>jeung</u> -ga-yul 'rate of export increase'
성토장 seong-to-jang	seong-to- <u>jjang</u> or seong-to-jang 'stage of debate'

## 2. Dependent Nouns (DN) (involving tensification)

될수있도록 dwel-su-id-ddo-log (dwel-su-iss-do-log)	dwel- <u>ssu</u> -id-ddo-log 'do DN be -par.'
---	--

않을것이라고  
a-neul-geo-si-la-go  
(anh-eul-geos-i-la-go)

a-neul-ggeo-si-la-go  
'not DN be -par.'

퍼센트대에서  
peo-sen-teu-de-e-seo

peo-sen-teu-dde-e-seo  
'percent DN -loc.'

### 3. Borrowed words with /s/

뉴스서비스가  
nyu-seu-seo-bi-seu-ga

nyu-sseu-sseo-bi-sseu-ga  
'news service -nom.'

펄스는  
peol-seu-neun

peol-sseu-neun  
'pulse -top'

시스템구현이  
si-seu-tem-gu-hyeo-ni  
(si-seu-tem-gu-hyeon-i)

ssi-seu-tem-gu-hyeon-i  
'system implementation -nom.'

로스엔젤레스로  
lo-seu-en-jel-le-seu-lo

lo-sseu-en-jel-le-sseu-lo  
'Los Angeles -loc.'

세이브로  
se-i-beu-lo

sse-i-beu-lo  
'save -loc.'

트렌스덤  
teu-len-seu-deom

teu-len-sseu-deom  
'part of a product name'

샘플을  
sem-peu-leul  
(sem-peul-eul)

ssem-peu-leul  
'sample -acc.'

루이스였다  
lu-i-seu-yeod-dda  
(lu-i-seu-yeoss-da)

lu-i-sseu-yeod-dda  
'Luis be -past-par.'

서스데이  
seo-seu-de-i

sseo-seu-de-i  
'Thursday'

종합스포츠단지  
jong-hab-sseu-po-ceu-dan-ji  
(jong-hab-seu-po-ceu-dan-ji)

jong-hab-seu-po-ceu-dan-ji  
'sports complex'

아나로그시스템은  
a-na-lo-geu-si-seu-te-meun  
(a-na-lo-geu-si-seu-tem-eun)

a-na-lo-geu-ssi-seu-te-meun  
'*analog system -top.*'

자바프로세서의  
ja-ba-peu-lo-se-seo-eui

ja-ba-peu-lo-sse-sseo-eui  
'*Java processor -gen.*'

#### 4. /n/-insertion

통근열차  
tong-geu-nyeol-ca  
(tong-geun-yeol-ca)

tong-geun-nyeol-ca  
'*commuter train*'

태평양  
te-pyeong-yang

te-pyeong-nyang  
'*Pacific ocean*'

국민여론의  
gug-mi-nyeo-lo-neui  
(gug-min-yeo-lon-eui)

gug-min-nyeo-lo-neui  
'*public opinion -gen.*'

조선여성과  
jo-seo-nyeo-seong-gwa  
(jo-seon-yeo-seong-gwa)

jo-seon-nyeo-seong-gwa  
'*Chosun woman -par.*'

미화약품  
mi-hwa-ha-gyag-pum  
(mi-hwa-hag-yag-pum)

mi-hwa-hang-nyag-pum  
'*American chemical drug*'

#### 5. Lateralization of /n/ & nasalization of /l/

자금동원력  
ja-geum-dong-weol-lyeog  
(ja-geum-dong-weon-lyeog)

ja-geum-dong-weon-nyeog  
'*ability to gather money*'

미관리들을  
mi-gwan-ni-deu-leul  
(mi-gwan-li-deul-eul)

mi-gwal-li-deu-leul  
'*American officials -acc.*'

## References

- Jeongwon Cha, Geunbae Lee, and Jong-Hyeok Lee. Hybrid POS tagging with generalized unknown-word handling. In *Proceedings of the 2nd International Workshop on Information Retrieval with Asian Languages*, pages 43–50, 1997.
- Chunghye Han, Eon-Suk Ko, Heejong Yi, and Martha Palmer. Penn Korean Treebank: Development and evaluation. In *Proceedings of the 16th PacificAsian Conference on Language and Computation*. Korean Society for Language and Information, 2002.
- Yongkuk Hong, Myoung-Wan Koo, and Gijoo Yang. A Korean morphological analyzer for speech translation system. In *Proceedings of the 4th International Conference on Spoken Language Processing*, 1996.
- Sun-Ah Jun. The accentual phrase in the Korean prosodic hierarchy. *Phonology*, 15(2):189–226, 1998.
- Sun-Ah Jun. K-ToBI (Korean ToBI) labelling conventions. Version 3.1, 2000. <http://www.linguistics.ucla.edu/people/jun/ktobi/K-tobi.html>.
- Byeongchang Kim, Geunbae Lee, and Jong-Hyeok Lee. Morpheme-based grapheme to phoneme conversion using phonetic patterns and morphophonemic connectivity information. *ACM Transactions on Asian Language Information Processing*, 1(1):65–82, 2002.
- Ross King. Korean Writing. In Peter T. Daniels and William Bright, editors, *The World's Writing Systems*, pages 218–227. Oxford University Press, 1996.
- Korea Ministry of Education and Human Resources Development. *국어 어문 규정 (Korean Standard Rule Collections)*. 대한교과서 (Taehan Publishing), Korea, 1995.
- Hoyoung Lee. *국어음성학 (Korean Phonetics)*. 태학사, Korea, 1996.
- Linguistic Data Consortium (LDC). Korean Newswire Text Corpus, 2000. catalog number LDC2000T45, ISBN 1-58563-168-X.
- Anoop Sarker and Chunghye Han. Statistical morphological tagging and parsing of Korean with an LTAG grammar. In *Proceedings of the 6th International Workshop on Tree Adjoining Grammars and Related Formalisms*, 2002.
- Insup Taylor. The Korean writing system: An alphabet? A syllabary? A logography? In *Proceedings of Visible Language*, volume 2, pages 67–82, New York, 1979. Plenum.
- Kyuchul Yoon. The effects of prosody on segmental variation. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2003)*, Borovets, Bulgaria, 2003.

VOWELS					
Romanization	Hangul	IPA	Height	Front/Back	
/a/	ㅏ	[a]	low	back	
/i/	ㅣ	[i]	high	front	
/eu/	ㅡ	[ɨ]	high	central	
/u/	ㅜ	[u]	high	back	
/e/	ㅓ	[e]	mid	front	
/eo/	ㅛ	[ɤ]	mid	central	
/o/	ㅝ	[o]	mid	back	
CONSONANTS					
Romanization	Hangul	IPA	Manner	Place	Phonation
/b/	ㅂ	[p]	stop	labial	lenis
/bb/	ㅃ	[pʰ]	stop	labial	fortis
/p/	ㅍ	[pʰ]	stop	labial	aspirated
/d/	ㄷ	[t]	stop	alveolar	lenis
/dd/	ㄸ	[tʰ]	stop	alveolar	fortis
/t/	ㅌ	[tʰ]	stop	alveolar	aspirated
/g/	ㄱ	[k]	stop	velar	lenis
/gg/	ㄲ	[kʰ]	stop	velar	fortis
/k/	ㅋ	[kʰ]	stop	velar	aspirated
/ss/	ㅆ	[sʰ]	fricative	alveolar	fortis
/s/	ㅅ	[sʰ]	fricative	alveolar	aspirated
/h/	ㅎ	[h]	fricative	glottal	aspirated
/j/	ㅈ	[tʃ]	affricate	palatal	lenis
/jj/	ㅉ	[tʃʰ]	affricate	palatal	fortis
/c/	ㅊ	[tʃʰ]	affricate	palatal	aspirated
/m/	ㅁ	[m]	nasal	labial	lenis
/n/	ㄴ	[n]	nasal	alveolar	lenis
/ng/	ㅇ	[ŋ]	nasal	velar	lenis
/l/	ㄹ	[l]	lateral	alveolar	lenis
/w/	n/a	[w]	approximant	labial	lenis
/y/	n/a	[j]	approximant	palatal	lenis

Table 1

Korean segments represented using jamo (hangul subsyllabic letters), the ISO TR 119 41:1996 romanization, the International Phonetic Alphabet, and the features that define the phonemes in the TTS phoneset. There are no separate symbols for glides.

Rule order	Rule name
1	Palatalization
2	Simplification of coda clusters (subcomponent 1: word-final)
3	Simplification of coda clusters (subcomponent 2: resyllabification)
4	Simplification of coda clusters (subcomponent 3: consonant cluster)
5	Simplification of coda clusters (subcomponent 4: aspiration)
6	Simplification of coda clusters (subcomponent 5: sonorant)
7	Neutralization of coda obstruents (subcomponent 1: word-final)
8	Neutralization of coda obstruents (subcomponent 2: plain)
9	Neutralization of coda obstruents (subcomponent 3: aspiration)
10	Tensification of lenis obstruents
11	Nasalization of obstruents
12	Lateralization of /n/ & nasalization of /l/
13	Liaison

Table 2

Ordering of rule application. See section 2.3 for details.

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

Table 3

Tabulation of 1.3 million eojeol by the frequency in the Newswire, the number of rules applied, and the number of syllables

	correct	total	% correct
sentences	534	621	86.6%
morphemes	14929	15039	99.27%
phonemes	58419	58583	99.72%

Table 4

Performance figures of Kim et al. [2002]

Evaluators	KY	EK	HS
Sino-Korean words (involving tensification)	40	40	40
Dependent nouns (involving tensification)	3	3	3
Borrowed words with /s/	12	12	12
/n/-insertion	5	5	5
Lateralization of /n/, Nasalization of /l/	2	2	2
<b>Total unacceptable forms</b>	62	62	62
<b>ojeol accuracy rate</b>	91.17%	91.17%	91.17%

Table 5

Evaluation of GTP output ojeol by three native speakers of Korean