A Korean Grapheme-to-Phoneme Conversion System
Using Selection Procedure for Exceptions


*Speech Information Technology Research Center, Kwangwoon University,
** Department of Computer Engineering, Kwangwoon University
*** Department of Computer Science, Dongduk Women’s University

sunhkim@daisy.kw.ac.kr, dkswendms@hanmail.net, kimsh@daisy.kw.ac.kr, yhlee@dongduk.ac.kr

Abstract
Cultural, social, economic and other various environmental factors affect our language, and different words and terminology are used and coined for different contexts, which triggers quantitative change of vocabulary of a language. This paper presents a Korean grapheme-to-phoneme conversion system using a selection procedure for exceptions from added text corpus, which reflects such dynamic nature of the Korean language. For our experiment, we used the text corpus released by the Electronics and Telecommunications Research Institute (ETRI) for speech recognition, consisting of 53,750 sentences (740,497 Eojols), and obtained a 100% performance level of the proposed grapheme-to-phoneme conversion system.

1. Introduction
Being an essential part of an automatic speech recognition (ASR) system and a Text-To-Speech (TTS) system, a grapheme-to-phoneme (GTP) conversion is a system for assigning phonemic/phonetic transcriptions to graphemic word forms. Generally, a GTP consists of an extensive word list, a set of rules, or a hybrid combining a word list (exceptions) and a set of rules. The Korean language provides a good example to which a hybrid strategy applies; that is, a Korean GTP is composed of a set of regular rules and an exceptions dictionary [1, 2, 3].

While the regular rules reflect the systematic phonological aspect of the language, the exceptions are related to its quantitative change of vocabulary. This dynamic aspect of language is due to the cultural, social, economic and other various environmental factors, and different words and terminology are used and coined for different contexts every day. Once the regular rules are established, the performance of the GTP system depends on the exceptions dictionary, which shows the dynamic aspect of the language.

This paper aims to present a Korean GTP conversion system using selection procedure for exceptions from added text corpus based on [4, 5]. This study will help to improve the performance of the GTP system and, finally, that of the ASR and TTS in Korean.

The paper is organized as follows. The following section describes a Korean GTP system. In section 3, we propose a method of compiling an exceptions dictionary in two steps: compiling an exceptions dictionary using a general dictionary and compiling an exceptions dictionary using added text corpus. Section 4 presents the results of the experiment using the text corpus released by ETRI for speech recognition, consisting of 53,750 sentences (740,497 Eojols). And in the last section we provide the conclusion of this paper.

2. A Korean Grapheme-to-Phoneme Conversion System
Adopting a hybrid strategy for the GTP, we propose a Korean GTP conversion system as shown in FIG. 1 below. This system does not deal with the procedure beyond the morphological analysis, such as the preprocessing or the text normalization. Once the sentences are preprocessed, they are analyzed based on their part-of-speech information through morphological analysis.

![Figure 1: GTP conversion system](image)

Only after the morphological analysis, will the GTP rules take effect. The GTP rules in FIG 1 are a combination of a set of regular rules and an exceptions dictionary, for which a set of rules for the exceptions take effect. The regular rules are divided into two groups: the morphological rules and the phonological rules. As is shown in Figure 1, the morphological rules precede the rules for the exceptions, which in turn are followed by the phonological rules.

As we have shown in [5], the exceptions for the Korean GTP are characterized by certain phonological alternations in their limited contexts. Accordingly, once the exceptions are extracted, the rules for the exceptions convert the phonemes in the given phonetic contexts into the corresponding pronunciations. This process for the exceptions is possible.
because regularity of the phonological phenomena related to
the exceptions is defined as in [5].

The morphological rules affect the words containing
certain morphological information after the
morphological analysis. In Korean, a lenis consonant in
the beginning of the verb endings are realized as a fortis
consonant when it is preceded by the verb stem that
ends with a nasal consonant (Morphological Tensification). The phonological rules apply without
any exceptions to the words in the given context.

Following the research on the phonological alternations
for the standard Korean pronunciation within Eojels [4], this
proposed system includes a set of rules of phoneme change. Table I shows the observed alternations and the number of
their rules when they are implemented in the system. The regular alternations are for the morphological and
phonological rules and the irregular alternations are those for
the exceptions.

<table>
<thead>
<tr>
<th>Regular phonological</th>
<th>Observed alternations</th>
<th># of rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>Morphological</td>
<td>Morphological Tensification</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>Neutralization in Coda</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Neutralization of consonant clusters</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Nasalization of liquids</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Phonological Tensification</td>
<td>59</td>
</tr>
<tr>
<td></td>
<td>Aspiration</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>Nasalization of obstruent</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>Lateralization</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Merge (8)</td>
<td>69</td>
</tr>
<tr>
<td></td>
<td>Double Nasalization</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>h-Deletion</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>h-Nasalization</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Liaison</td>
<td>437</td>
</tr>
<tr>
<td></td>
<td>Palatalization</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Vowel Shortening</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Lexical Tensification</td>
<td>116</td>
</tr>
<tr>
<td></td>
<td>Nasalization of laterals</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>n-Insertion</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>Neturalization/Simplification+Liaison</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 1: Alternations and the number of their corresponding rules

3. Exceptions Dictionaries

In this section, we provide a method of generating exceptions
dictionaries for a Korean GTP system based on the study of
exceptions [5]. All possible combinations of consonants and
vowels and their realizations are examined through the
analysis of [6] in order to consider the phonetic contexts and
the phonological phenomena related to the exceptions.

The phonetic contexts and the phonological phenomena of the exceptions are summarized in Table 2. According to [5],
there are three types of the phonetic contexts in which the
exceptions appear: (I) when a sonorant is followed by a lenis
obstruent; (II) when a nasal consonant is followed by a lateral
consonant; and (III) when a consonant in the coda position is
followed by the vowel which is the nucleus of the following
syllable containing a silent consonant (‘○’). And the

phonological phenomena in each context correspond to the
irregular alternations in Table 1: the Lexical Tensification in
(I); the Nasalization of the lateral consonant in (II); and,
finally, the /n/-Epenthesis and the Liaison following the Neutralization in the coda or the Simplification of the
c consonant clusters in (III).

<table>
<thead>
<tr>
<th>p</th>
<th>t</th>
<th>s</th>
<th>c</th>
<th>k</th>
<th>l</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td>m</td>
<td>n</td>
<td>N</td>
<td>l</td>
<td>V</td>
<td>C</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Exceptions and their phonetic contexts

(C: a consonant of a consonant cluster; V: a vowel or
diphthong)

(I) Lexical Tensification
(II) Nasalization of the lateral
(III) /n/-Epenthesis, Neutralization/ Simplification + Liaison

3.1. Exceptions dictionary using a general dictionary

The first step to compile an exceptions dictionary involves the
generation of an exceptions dictionary by extracting the words
having the exceptional pronunciations among the entries of a
general Korean dictionary. By using a general Korean
dictionary, the repetition of vocabulary can be minimized and
also various kinds of vocabulary can be included in the exceptions dictionary. The general Korean dictionary used in
this research is [6], which has a record of about 50,000 entries
of high frequency.

To select the exceptions among the entries of [6], we
begin with extracting the words in the phonetic contexts in
which the irregular alternations are observed as was shown
above in Table 2. The words resulting from this process,
called the Reference Dictionary 1 in this study, consist of a
group of words showing exceptional pronunciations and a
group without any alternations. We extract the words of
exceptional pronunciations by manual review of the words
listed in the Reference Dictionary 1 based on [7, 8]. And the
resulting words are compiled as the basic exceptions
dictionary.

FIG. 2 describes the method of compiling a basic exceptions dictionary using a general dictionary.
Figure 2: Method of compiling basic exceptional pronunciation dictionary

3.2. Exceptions dictionary using added text corpus

In our actual life, we use a lot of words that do not appear in a general dictionary. Here, we propose a method of generating an exceptions dictionary to complement the pronunciation dictionary when the text corpus is added to an ASR or a TTS system.

The text corpus is basically an assemblage of sentences, which need to be analyzed, pre-processed, and divided into Eojols (units surrounded by space). As the words having exceptional pronunciations are nouns and their derivatives, the Eojols are sufficient units for the extraction of the exceptions. Figure 3 describes the method of compiling the exceptions dictionary using added text corpus.

First, the Eojols in the exceptional contexts are gathered, which we call the Vocabulary in the Exceptional Contexts 1. They are then compared to the words in the Reference Dictionary 1 that was generated through the process described in Figure 2. As a result of the comparison and after removing all repeated words, we obtain the Eojols that are not included in the reference dictionary 1. We call them Vocabulary in the Exceptional Contexts 2. After reviewing the entries of the Vocabulary in the Exceptional Contexts 2, we obtain newly added Eojols of exceptions, which we call the Added Exceptions Dictionary. Finally, the exceptions dictionary 1 compiled from the general dictionary, together with the Added Exceptions Dictionary, constitute the Exceptions Dictionary 2, the exceptions dictionary using the added text corpus.

In order to extract the exceptions dictionary using added text corpus, we do not exploit the existing exceptions dictionary but we take the Reference Dictionary which includes all the words in the exceptional contexts. The Reference Dictionary 2 in Figure 3, which is a combination of the Vocabulary in the Exceptional Contexts 2 and the Reference Dictionary 1, will be used as the reference dictionary, when an exceptions dictionary is to be edited from a new text corpus.

4. Results

The experiments were conducted in two steps: (1) the performance test of the GTP with the basic exceptions dictionary according to the proposed method in 3.1.; and (2) the performance test of the GTP with the extended exceptions dictionary according to the proposed method in 3.2.

First, for the basic exceptions dictionary, we obtained 2,855 words through the analysis of [6], which has 49,561 entries. There were 29,983 words in the exceptional contexts, from which 2,855 words of exceptional pronunciation were extracted. The performance of the GTP system with this exceptions dictionary of 2,855 entries was evaluated using the text corpus for ASR designed by ETRI. We used the text corpus, which is composed of a month newspapers, which consisted of 53,750 sentences (740,497 Eojols). After removing the repeated words, we had 189,481 Eojols for the experiment.

The following Table presents the results of the experiment using the proposed GPT system. The performance of the system is given in the following table 3 according to the specific cases where each set of rules takes effect: the number of Eojols to which the morphological rules applied and its rate was presented under the column M; the number of Eojols to which the rules for exceptions applied and its rate was presented under the column E; the number of Eojols to which the phonological rules applied and its rate was presented under the column P; and the number of unchanged Eojols U and its rate was presented under the column U.

<table>
<thead>
<tr>
<th></th>
<th>M</th>
<th>E</th>
<th>P</th>
<th>U</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td># of Eojols</td>
<td>19</td>
<td>11,765</td>
<td>112,471</td>
<td>64,360</td>
<td>189,481</td>
</tr>
<tr>
<td>Percentage (%)</td>
<td>0.01</td>
<td>6.21</td>
<td>59.36</td>
<td>33.96</td>
<td>99.54</td>
</tr>
</tbody>
</table>

Table 3: Performance rate of the GTP with the basic exceptions dictionary

Before conducting the GTP performance test with the extended exceptions dictionary, we extracted the exceptions dictionary using added text corpus, which consisted of 53,750 sentences (740,497 Eojols). After removing the repeated words, we were left with 189,481 Eojols for the experiment. We used the text corpus of ETRI, which consisted of 53,750 sentences (740,497 Eojols). As the input text corpus, 189,481 Eojols were used after removing the repeated words in the exceptional contexts. Among these 189,481 Eojols, the 159,975 Eojols were in the exceptional contexts, which constitute the vocabulary in the exceptional contexts 1 in Figure 3. When this vocabulary in the exceptional contexts 1 is compared with the Reference Dictionary of 29,983 words, 91,226 Eojols are obtained as the Vocabulary in the exceptional contexts 2. Through a manual review of these 91,226 Eojols of the Vocabulary in the Exceptional Contexts 2, we obtained 69 Eojols for the Added Exceptions Dictionary. As these 91,226 Eojols can belong to more than two exceptional contexts, the number of Eojols according the
exceptional contexts was 149,219. The following Table presents the distribution of the resulting exceptions of the Vocabulary of in the exceptional contexts 2.

<table>
<thead>
<tr>
<th># of Vocabulary</th>
<th># of exceptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexical Tensification</td>
<td>107,294 68</td>
</tr>
<tr>
<td>Nasalization of laterals</td>
<td>267</td>
</tr>
<tr>
<td>Neteuralization/Simplification + Liaison</td>
<td>41,658 1</td>
</tr>
<tr>
<td>Total</td>
<td>149,219 (91,226) 69</td>
</tr>
</tbody>
</table>

Table 4: Number of the vocabulary in the exceptional contexts and the number of the exceptions in each context

Next, we conducted the GTP performance test with the extended exceptions dictionary, which consisted of 2,924 words. We used the same text corpus of 53,750 sentences (740,497 Eojols), which was reduced to 189,481 Eojols after removing the repeated words. We obtained 100% of the performance level with the extended exceptions dictionary as follows.

<table>
<thead>
<tr>
<th># of Eojols</th>
<th>M</th>
<th>E</th>
<th>P</th>
<th>U</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage (%)</td>
<td>0.01</td>
<td>6.67</td>
<td>59.36</td>
<td>33.96</td>
<td>99.54</td>
</tr>
</tbody>
</table>

Table 5: Result derived from selection procedure

5. Conclusions

In this paper we have presented a GTP system using the selection procedure for exceptions from added text corpus, which reflects the dynamic nature of the Korean language. As for the selection procedure for exceptions, we proposed a method of generating exceptions dictionaries for a Korean GTP system based on the study of exceptions [5]. The process of generating the exceptions dictionary was conducted in two steps. The first step was to compile an exceptions dictionary of 2,855 words by extracting the exceptions among the entries of a general Korean dictionary. Next, we obtained 69 words for the Added Exceptions Dictionary according to the proposed the method of generating an exceptions dictionary using the selection procedure to complement the exceptions dictionary. For our experiment, we used the text corpus released by ETRI for speech recognition, consisting of 53,750 sentences (740,497 Eojols), and obtained a 100% performance level of the proposed GTP system. This study will contribute the performance improvement of an ASR or a TTS system.

6. References


