# A TWO-LEVEL MORPHOLOGICAL ANALYSIS OF KOREAN

Deok-Bong Kim, Sung-Jin Lee, Key-Sun Choi, and Gil-Chang Kim
Center for Artificial Intelligence Research
Computer Science Department, KAIST
373-1 Kusong-Dong, Yusong-Ku, Taejon 305-701, Korea
E-mail: {dbkim, kschoi}@csking.kaist.ac.kr

## ABSTRACT

*The two-level morphology model has received a great deal of attention and has been implemented for languages like Finnish, English, Japanese, Russian, French, and so on. However, this model has been claimed to be inappropriate for Korean morphological analysis, because the complex conjugation (inflection) and agglutination in word formation, and the syllable-based representation of words may lead to a huge number of two-level morphological rules. In this paper, we show that the two-level model can be successfully applied to Korean and its rule size is limited to only 52. An extension of two-level morphology is described for Korean language.*

## INTRODUCTION

The two-level morphology model (Koskenniemi, 1983; Antworth, 1990; Barton, 1986; Ritchie, 1991; Sproat, 1992) is a well-known computational model of morphology, which has adaptability as well as simplicity. In practice, this model has been successfully applied to several languages including Finnish, English, Japanese, Russian, and French. However, the two-level model has been considered to be inappropriate for Korean (Kang, 1992; Kwon, 1991). That is, the two-level morphological analysis of Korean is believed to be difficult and infeasible because the complex conjugation (inflection) and agglutination in word formation, and the syllable-based representation of words may lead to a huge number of two-level morphological rules. In this paper, we show that the two-level model can be successfully applied to Korean and its rule size is limited to only 52.

This paper presents a successful two-level system for Korean morphological analysis. The system was based on a shareware PC-KIMMO (Antworth, 1990); however, we extended the I/O component of PC-KIMMO to handle Korean alphabet *HANGUL*; we constructed a Korean dictionary and a Korean morphological grammar (i.e., morphotactics and spelling rules) for the PC-KIMMO; we also used a shareware KGEN (Miles, 1991) to translate the linguistic spelling rules into the executable automata (i.e., finite state transducers (FSTs)). This paper focuses on the dictionary and the morphological grammar for Korean.

## TWO-LEVEL REPRESENTATION OF KOREAN WORDS

The two level model is concerned with directly mapping between two representations of a word: (1) the *surface form* (SF) as it appears in the text, and (2) the *lexical form* (LF) which is represented as a sequence of basic morphs and diacritics (e.g., '+' to mark morpheme boundary and '#' for word boundary). As a result, an input word in the two-level model is analyzed by mapping the word itself (SF) to a sequence of lexical forms in dictionary without intermediate stages. In this section, we present a two-level representation of Korean words.

To understand the two-level description for Korean morphology, one should be properly familiar with Korean alphabet and their transcription system. So we first describe them. For ordinary writing system, the Korean alphabet consists of 40 letters: 10 pure vowels, 11 compound vowels, 14 basic consonants and 5 double consonants. A Korean word is represented with a sequence of syllables; a syllable can be made up of a consonant, a vowel, and a consonant; there are several forms of syllables (e.g., *CV*, *CVC*, *VC*, *V*, and *C* forms); and initial consonant letter may not be distinguished from final consonant letter. However, the initial consonant and the final consonant must be distinguished from each other for successful two-level

Table 1: The transcription of Korean alphabet (HANGUL).

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Pure Vowels | HANGUL | 아 | 어 | 오 | 우 | 애 | 에 | 으 | 이 | 위 | 외 | | | | |
| | IPA | a | ə | o | u | ɛ | e | ɨ | i | ü | ö | | | | |
| | MYCODE | a | e | o | u | 8 | 9 | _ | i | wu | wi | | | | |
| Compound Vowels | HANGUL | 야 | 여 | 요 | 유 | 얘 | 예 | 워 | 웨 | 와 | 왜 | 의 | | | |
| | IPA | ya | yə | yo | yu | yɛ | ye | wə | we | wa | wɛ | ɨy | | | |
| | MYCODE | ya | ye | yo | yu | y8 | y9 | we | w9 | wa | w8 | yi | | | |
| Basic Consonants | HANGUL | ㄱ | ㄴ | ㄷ | ㄹ | ㅁ | ㅂ | ㅅ | ㅇ | ㅈ | ㅊ | ㅋ | ㅌ | ㅍ | ㅎ |
| | IPA | k | n | t | l | m | p | s | ŋ | č | $\check{c}^h$ | $k^h$ | $t^h$ | $p^h$ | h |
| | MYCODE(I) | g | n | d | l | m | b | s | | j | c | k | t | p | h |
| | MYCODE(F) | G | N | D | L | M | B | S | * | J | C | K | T | P | H |
| Double Consonants | HANGUL | ㄲ | ㄸ | ㅃ | ㅆ | ㅉ | | | | | | | | | |
| | IPA | k' | t' | p' | s' | č' | | | | | | | | | |
| | MYCODE(I) | q | f | r | v | z | | | | | | | | | |
| | MYCODE(F) | Q | | | V | | | | | | | | | | |

system; if not, it might cause a lot of useless work (i.e., invalid mapping) and incorrect results because i-*th* consonant in a word is not clear whether it is an initial consonant or a final consonant. Furthermore, to write two-level spelling rules for PC-KIMMO, each of Korean alphabet must be mapped to ASCII character on the keyboard. Therefore, we devised a transcription system for Korean alphabet as shown in Table 1, which has the following features:

- There is no letter corresponding to the initial consonant 'ㅇ'. We did not consider the letter because it is a sort of an orthographic filler for the ordinary writing system and is not pronounced.
- The initial consonant letters are not the same as the final consonant letters. (To see this, compare the initial consonants MYCODE(I) with the final consonants MYCODE(F) in Table 1.)
- Each of compound vowels is represented by a pair of two letters: a semi-vowel letter (i.e., *y* or *w*) and one of pure vowel letters excluding '위' /*ü*/ and '외' /*ö*/; here '위' and '외' are treated as the compound vowels.
- There are two archiphoneme letters: (1) the archiphoneme *A* for the proper treatment of *vowel harmony*[1], which can be changed into *NULL* symbol 0, a vowel letter *a*, or a vowel letter *9* by context; and (2) the archiphoneme *I* for the proper treatment of predicative postposition '이' /*i*/, which can be changed into either *0* or a vowel letter *i* by context.

We believe that our transcription system makes it simple and clear to describe two-level spelling rules of Korean, and it enables the two-level processor to handle efficiently the complex spelling changes.

Here, three special symbols are used properly to treat lexical irregularities of Korean verbal morphology: *+* for regularity, *X* for '*le*'-irregularity, and *$* for all irregularities excluding the '*le*'-irregularity; *X* must be differentiated from *$* because of the following reasons. In Korean morphology, most of verbal stems ending in the syllable '르' /*lɨ*/ are irregular. The final syllable '르' /*lɨ*/ of the stem, when followed by the vowel '어' /ɔ/ and preceded by any vowel other than the light vowels ('아' /*a*/ and '오' /*o*/), is changed into '러' /*lɔ*/ and the consonant 'ㄹ' /*l*/ is added to the preceding syllable. We call it '*L*'-irregularity. For example, the verb stem '흐르' /*hɨ-lɨ*/ (to flow) plus the suffix '어' /ɔ/ (INFINITIVE) becomes the verbal word '흘러' /*hɨl-lɔ*/. However, there is '*le*'-irregularity which oc-

[1] Modern Korean has a "diagonal" vowel harmony (Ahn, 1985) kept in only one area of word formation, that is, between the final vowel of a verbal stem and the following ə-initial suffix. This system works in the ə-initial suffix harmony where ə has an alternation *a* if the final vowel of a verbal stem is a light vowel *a* or *o*. For example, the verb stem '보' /*bo*/ (to see) plus the suffix '어' /ə/ (INFINITIVE) becomes the verbal word '보아' /*bo-a*/. However, the verb stem '주' /*cu*/ (to give) plus the suffix '어' /ə/ (INFINITIVE) becomes the verbal word '주어' /*cu-ə*/. As a result, the archiphoneme *A* is used for the initial vowel ə of suffixes, which is to distinguish it from ə elsewhere.

curs in the same context as '*L*'-irregularity: it causes only to be changed the following vowel '어' /ə/ into '러' /lə/; for example, the verb stem '이르' /*i-lɨ*/ (to arrive) plus the suffix '어' /ə/ (INFINITIVE) becomes the verbal word '이르러' /*i-lɨ-lə*/. Therefore, a mechanism is needed to treat them properly.

One of the special symbols is used to represent a specific lexical form, and is almost placed at the end of the lexical form. For example, the verbal stem *guB* has two meanings, i.e., "curved" as an adjective and "grill" as a verb. In this case, the problem is on the difference between the variation forms for adjective and those for verb; when it is combined with a suffix *A*, the surface form becomes either the *guBe* as adjective, or the *guwe* as verb. To distinguish between them, the following lexical forms can be listed in dictionary: *guB+* for regular adjective, and *guB$* for '*B*'-irregular verb.

## WORD STRUCTURE AND LEXICONS

The word structure in general denotes knowledge of the internal morpheme combinations of known words. As a result, it shows how morphemes can combine to form valid words; it is important to a proper word recognition. In the two-level model it is represented with linked lexicons, i.e., with *continuation classes* of morphemes.

The continuation classes used in our lexicons are as follows: interjection (IJ), prenoun (PR), adverb (AB), noun (NN), pronoun (PN), numeral (NU), verb (VB), adjective (AJ), verbalizer (VR), postposition (PP), I-postposition (IP), nominal-prefix (NF), verbal-prefix (VF), prefinal-ending (PE), final-ending (FE), nominal-ending (NE)[2], *Begin*, and *End*. Every class indicates a lexicon. However, the *Begin* and *End* are some special lexicons; *Begin* amounts to the initial state in automata, and *End* has the same role as the final state; in fact, there is no lexical entry. The following shows our linked lexicons.

```
Begin -> interjection | prenoun | adverb
        | noun | pronoun | numeral | verb
        | adjective | nominal-prefix
        | verbal-prefix
interjection -> End
prenoun -> End
adverb -> End | postposition
nominal-prefix -> noun
verbal-prefix -> verb | adjective
noun -> End | postposition
       | I-postposition | verbalizer
pronoun -> End | postposition
       | I-postposition
numeral -> End | postposition
       | I-postposition
verb -> prefinal-ending | final-ending
       | nominal-ending
adjective -> prefinal-ending
       | final-ending | nominal-ending
verbalizer -> prefinal-ending
       | final-ending | nominal-ending
I-postposition -> prefinal-ending
       | final-ending | nominal-ending
postposition -> End
prefinal-ending -> final-ending
       | nominal-ending
final-ending -> End
nominal-ending -> End | postposition
       | I-postposition
```

The right arrow '->' indicates that a class on its left side can continue with one of classes on its right side; a vertical bar '|' indicates OR.

## TWO-LEVEL RULES AND FINITE STATE AUTOMATA

Based on the work of Korean morphology by Lee (1991), 52 two-level rules has been developed for the Korean morphological alternations. By way of an example, we explain the following Korean morphological alternation in the two-level framework.

In Korean, some verbals ending in the final consonant *B* are irregular. The final consonant *B* of the stem, when followed by a vowel, is changed into *w*. But it is not changed when followed by a consonant. For example, when an irregular verb *doB* (to help) is combined with the suffix *A*, it is changed into *dowa*. In the two-level system, it is represented as follows:

*Lexical Representation: d o B $ + A*
*Surface Representation: d o w 0 0 a*

[2] The *nominal-ending* belongs to final-ending; it consists of nominal endings, sentential endings, and connective endings.

This shows a correspondence between lexical representation and surface representation. In PC-KIMMO, such a correspondence is represented with the notation *lexical-character:surface-character* like *d:d*, *o:o*, *B:w*, *$:0*, *+:0*, and *A:a*. Here the lexical character *$* is a signal indicating that a basic word or stem followed by it is irregular, and it corresponds to a surface *0* (the NULL symbol) which is not printed in the output form. The lexical *+* (a morpheme boundary symbol) also corresponds to a surface *0*.

The above alternation may be described as the following two-level rule:

*B:w* ⇔ ___ *$:0 +:0 A:@ (B Variation Rule)*

This rule states that a lexical *B* is realized as a surface *w* if and only if it is followed by the conjugation information *$*, the morpheme boundary *+*, and a linking suffix *A*. A surface *@* in the above rule stands for any alphabetic character that constitutes a feasible pair with a lexical *A*. For example, the surface *@* may be realized as *a*, *e*, or *0* when all feasible pairs with lexical *A* are like *A:a*, *A:e*, and *A:0*.

The two-level rules can be automatically translated into the state transition tables by using a rule compiler such as TWOL (Karttunen, 1987) and KGEN (Miles, 1991). The tables built by KGEN may be actually used in PC-KIMMO. The above rule is translated by KGEN into the state transition table below:

| | *B* | *B* | *$* | *+* | *A* | *@* | *(lexical characters)* |
|---|---|---|---|---|---|---|---|
| | *w* | *@* | *0* | *0* | *@* | *@* | *(surface characters)* |
| *1:* | *2* | *5* | *1* | *1* | *1* | *1* | |
| *2.* | *0* | *0* | *3* | *0* | *0* | *0* | |
| *3.* | *0* | *0* | *0* | *4* | *0* | *0* | |
| *4.* | *0* | *0* | *0* | *0* | *1* | *0* | |
| *5:* | *2* | *5* | *6* | *1* | *1* | *1* | |
| *6:* | *2* | *5* | *1* | *7* | *1* | *1* | |
| *7:* | *2* | *5* | *1* | *1* | *0* | *1* | |

The rows of the table represent the seven states, in which final states are marked with colons and nonfinal states are marked with periods. The columns represent arcs from one state to another. A zero transition indicates that there is no valid transition from that state for that input symbol.

## CONCLUSION

We have shown that the two-level morphology model, which has been claimed to be inappropriate for Korean, can be successfully applied to Korean. That is, we have implemented a successful two-level morphology system for Korean (see APPENDIX). This system was based on PC-KIMMO which is a shareware. However, we modified the I/O component of PC-KIMMO to handle Korean alphabet *HANGUL*; we have constructed a Korean dictionary for the PC-KIMMO, which contains about 12,000 entries; we represented a Korean morphotactics for the PC-KIMMO, which indicates the morphological structures of known words; we wrote 52 two-level spelling rules for the PC-KIMMO, which recover almost all spelling alternations in Korean morphology.

Our two-level system has been experimented with 2,172 randomly words selected from Korean textbooks (413,975 words) for elementary education. For this test set, the system produces the correct outputs although it includes about 5% extra incorrect analyses (i.e., *overgeneration*). Here the overgeneration is ascribed to the fact that it results from the weak expressive power of morphotactic information in PC-KIMMO.

## REFERENCES

Ahn, S. C. (1985). *The Interplay of Phonology and Morphology in Korean.* Ph.D. Thesis, Univ. of Illinois.

Antworth, E. L. (1990). *PC-KIMMO: A Two-Level Processor for Morphological Analysis.* Summer Institute of Linguistics.

Barton, G. E. (1986). Computational Complexity in Two-Level Morphology. In *Proceedings of the 24th Annual Meeting of Association for Computational Linguistics*, pp. 53-59.

Kang, S. S. and Y. T. Kim (1992). A Computational Analysis Model of Irregular Verbs in Korean Morphological Analyzer. *Journal of Korea Information Science Society, 19:2*, pp. 151-164. (in Korean)

Karttunen, L., K. Koskenniemi, and R. M. Kaplan (1987). *A Compiler for Two-Level Phonological Rules.* Xerox Palo Alto Research Center and Center for the Study of Language and Information.

Koskenniemi, K. (1983). *Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production.* Ph.D. Thesis, Univ. of Helsinki.

Kwon, H. C. and Y. S. Chae (1991). A Dictionary-based Morphological Analysis. In *Proceedings of Natural Language Processing Pacific Rim Symposium*, pp. 178-185.

Lee, H. S. and B. H. Ahn (1991). *Lecture on HANGUL Orthography.* Shin-Koo Press, Seoul. (in Korean)

Miles, N. and E. Antworth (1991). *Preliminary Documentation for KGEN – a rule compiler for PC-KIMMO –.* Summer Institute of Linguistics.

Ritchie, G. D., G. J. Russell, A. W. Black, and S. G. Pulman (1991). *Computational Morphology: Practical Mechanisms for the English Lexicon.* MIT Press, Cambridge.

Sproat, R. (1992). *Morphology and Computation.* MIT Press, Cambridge.

## APPENDIX: Running Examples

```
dbkim/csking> pckimmo
PC-KIMMO TWO-LEVEL PROCESSOR
 Version 1.0.5, Copyright 1992 SIL
Type ? for hel

PC-KIMMO> load rule kor.rul
Rules being loaded from kor.rul
 52 Rules Loaded

PC-KIMMO> load lexicon kor.lex
Lexicons being loaded from kor.lex
 Lexicon Start                  1 entries
 Lexicon Nominal             7973 entries
 Lexicon Adverb                20 entries
 Lexicon Verbal              2784 entries
 Lexicon Ending                94 entries
 Lexicon Postposition        1443 entries
 Lexicon Others                32 entries
 Lexicon End                    1 entries

PC-KIMMO> recognize

recognizer≫ 도와

doB$+A          돕$+A            [VB+FE]

recognizer≫ 일러

il_$+A          이르$+A          [VB+FE]

recognizer≫ 하였다

ha$+AV++da      하$+Aㅆ++다      [VB+PE+FE]

recognizer≫ 했다

ha$+AV++da      하$+Aㅆ++다      [VB+PE+FE]

recognizer≫ 학교에서

haGgyo+9se      학교+에서        [NN+PP]

recognizer≫ 준

juN             준               [NN]
ju++N           주++ㄴ           [VB + FE]
juL++N          줄++ㄴ           [VB + FE]

recognizer≫ 피하다

* pi+ha$+da     피+하$+다        [NN+VR+FE]
```