Language, Information and Computation(PACLIC 11), 1996, 229-236

# A Proposal of Korean Conjugation System and its Application to Morphological Analysis

Yoshitaka Hirano Yuji Matsumoto
Nara Institute of Science and Technology
Takayama, Ikoma, Nara 630-01 Japan
{yosita-h,matsu}@is.aist-nara.ac.jp

**Abstract**

This paper presents a new Korean verb conjugation system, which enables an easy treatment of Korean morphological phenomena such as contraction. This makes the size of the dictionary for ending forms to be small.

We also introduce a Korean morphological analysis system. Korean morphological analysis system generally analyzes sentences within the segments(a part between spaces). We propose a system that considers the information beyond segmentation.

## 1 Introduction

Korean has many irregular transformation such as contraction. Korean morphological analysis system MoA treats contraction within its system (J.-H.Kim 1994). We propose a method to treat such phenomena by means of a verb conjugation system. Korean verbs generally have many ending forms in conjugation. For example, 가다(go) has 가 as its stem, and takes a variety of conjugated endings such as ㅂ니다, 면, 지만, ㄴ and so on. In this way each verb requires a distinct set of ending forms. When we connect the ending 습니다 with a verb, 가다 becomes 갑니다 (stem:가 + ending:ㅂ니다), but 먹다(eat) becomes 먹습니다 (stem:먹 + ending:습니다). The ending 습니다 takes the form ㅂ니다 or 습니다 according to the verb. The conjugated form of the ending depends on the verb to which it is connected. When we compile a dictionary, we have to include all possible words with possible endings. However, this method is not practical.

We propose a method in which all surface variations are explained by verb conjugation. For example, as for the nonconjugational ending ㅂ니다, the verb 가다 conjugates to 가 and the verb 먹다 conjugates to 먹스. Then, 가다 becomes 갑니다 (stem:가 + conjugational ending:none + nonconjugational ending:ㅂ니다), and 먹다 becomes 먹습니다 (stem:먹 + conjugational ending:스 + nonconjugational ending:ㅂ니다).

After proposing a verb conjugation system, we describe a Korean morphological analysis system as a direct application of it.

In Korean morphological analysis, methods to reduce ambiguities have been studied. However, most of the systems analyze sentences only within segments (Eojeol, i.e., a sequence of morphemes surrounded by spaces)(J.-H.Kim 1995). We propose a method to reduce some ambiguities by means of using an information over segment boundary.

## 2 Verb conjugation

We prepare 24 conjugation types and seven conjugation forms for each of them. The conjugation types consists of five vowel stem verbs, two regular consonant stem verbs and nineteen irregular verbs. All verbs are classified into 24 types.

Table 1 shows seven conjugation forms and Table 2 shows some examples of verb conjugation. Table 3 shows a list of vowel stem verbs and their suffix vowels. In Table 2, '+' indicates a positive vowel stem verb and '−' indicates a negative vowel stem verb. 'reg' means a regular conjugation verb, and 'irg' means an irregular conjugation verb. 'C' indicates a consonant. And ㅏ and ㄴ specify unit letters. A unit letter means a constituent that constructs a hangul. Usually unit letters do not exist on the conjugational ending, however, we consider 보ㅏ means 봐 as shown in Table 4. We will describe the details on section 3. As a result we only need five conjugation types for each vowel stem verbs.

Because of the verb conjugation, we do not need to include the morphemes such as 면 으면 and ㄴ다 는다 into the dictionary. Therefore, it is possible to make the numbers of nonconjugational endings and prefinal endings be small. Furthermore, the adjective conjugation can be classified in a similar way as verbs. Note that only adjectives have ㅎ-irregular conjugation, and they do not have conjugation forms 2 and 6.

| form | the nonconjugational ending | example |
|---|---|---|
| form1 | not conjugates any verbs | 지만 |
| form2 | beginning with ㄴ, ㅅ, 오, ㅂ | 는군요 |
| form3 | connecting to 으 and not beginning with ㄴ, ㅅ, 오, ㅂ, ㄹ | 면 |
| form4 | connecting to 으 and beginning with ㄴ, ㅅ, 오, ㅂ, ㄹ | ㄹ까 |
| form5 | connecting to 아, 어, 여 | 요 |
| form6 | connecting to ㄴ, 는 | 다면 |
| form7 | kinds of 습니다 | 습니다 |

Table 1: Conjugation forms and possible nonconjugation endings

## 3 Korean character coding

We have built a Korean morphological dictionary with unit letters. When we analyze a sentence, the original sentence is transformed into unit letters.

| type | basic | stem | form1 | form2 | form3 | form4 | form5 | form6 | form7 |
|---|---|---|---|---|---|---|---|---|---|
| vowel1 | 만나다 | 만나 | | | | | | ㄴ | |
| vowel2 | 보내다 | 보내 | | | | | nothing ,어 | ㄴ | |
| vowel3 | 보다 | 보 | | | | | 아,ㅏ | ㄴ | |
| vowel4 | 주다 | 주 | | | | | 어,ㅓ | ㄴ | |
| vowel5 | 뛰다 | 뛰 | | | | | 어 | ㄴ | |
| reg.C+ | 받다 | 받 | | | 으 | 으 | 아 | 는 | 스 |
| reg.C− | 먹다 | 먹 | | | 으 | 으 | 어 | 는 | 스 |
| 으irg+ | 따르다 | 따ㄹ | ㅡ | ㅡ | ㅡ | ㅡ | ㅏ | ㄴ | ㅡ |
| 르irg+ | 오르다 | 오 | 르 | 르 | 르 | 르 | ㄹ라 | 른 | 르 |
| 러irg+ | 이르다 | 이르 | | | | | 러 | ㄴ | |
| 우irg | 푸다 | ㅍ | ㅜ | ㅜ | ㅜ | ㅜ | ㅓ | ㄴ | ㅜ |
| 하irg | 하다 | | 하 | 하 | 하 | 하여,해 | 한 | 하 | |
| 어irg | 그러다 | 그ㄹ | ㅓ | ㅓ | ㅓ | ㅓ | ㅐ | 근 | ㅓ |
| ㅂirg1 | 돕다 | 도 | ㅂ | ㅂ | 우 | 우 | 와 | ㅂ는 | ㅂ스 |
| ㅂirg2 | 눕다 | 누 | ㅂ | ㅂ | 우 | 우 | 워 | ㅂ는 | ㅂ스 |
| ㅂirg3 | 뵙다 | 뵈 | ㅂ | ㅂ | | | 어,ㅓ | ㅂ는 | ㅂ스 |
| ㄷirg− | 듣다 | 드 | ㄷ | ㄷ | ㄹ으 | ㄹ으 | ㄹ어 | ㄷ는 | ㄷ스 |
| ㅅirg+ | 낫다 | 나 | ㅅ | ㅅ | 으 | 으 | 아 | ㅅ는 | ㅅ스 |
| ㄹirg+ | 알다 | 아 | ㄹ | | ㄹ | | ㄹ아 | ㄴ | |

Table 2: example of verb conjugation

| type | suffix vowels |
|---|---|
| Vowel1 | ㅏ, ㅓ |
| Vowel2 | ㅐ, ㅔ, ㅕ, ㅖ |
| Vowel3 | ㅗ |
| Vowel4 | ㅚ, ㅜ, ㅣ |
| Vowel5 | others |

Table 3: Vowel stem verb types

| | | | | |
|---|---|---|---|---|
| ㅘ | ⟶ | ㅗ | + | ㅏ |
| ㅝ | ⟶ | ㅜ | + | ㅓ |
| ㅕ | ⟶ | ㅣ | + | ㅓ |
| ㅙ | ⟶ | ㅚ | + | ㅓ |

Table 4: Unit letter translation rules

Korean coding systems for morphological analysis often use an alphabet coding system. This has two consonant letters to specify initial consonant or final consonant(D.B.Kim 1994). However we use letters from the KS C5601 line4 as the unit letters. Using this character coding achieves better reading than the alphabet coding. The unit letters we proposed include the double final consonants such as ㄹㄱ and ㄹㅎ. This makes it easier to treat consonant stem verbs. Table 4 shows some special unit letter translation rules. This can treat the contraction. We also include ㅇ(ieung, a letter to show an empty consonant) as a unit letter. Using double consonants and ieung we can distinguish whether a consonant letter is a final or initial consonant. Thus, we need not prepare two sets of consonant letters for initial and final consonants. For example, suppose we analyze the sentence "그녀도 왔다." "왔다" is more accurately written "오았다". Such phenomenon that "오아" becomes "와" is called contraction. Contraction is easily handled in our translation rules. First we transform the original sentence into unit letters, "ㄱㅡㄴㅣㅕㄷㅗㅇㅗㅏㅆㄷㅏ

." For instance, the original form of "ㅇㅗㅏㅆ" is "왔". In our dictionary 오다 is written as "ㅇㅗㄷㅏ" by the unit letters. The conjugational form5 of this verb whose conjugational type is vowel3 has the ending "ㅇㅏ" and "ㅏ". So "오다" conjugates "ㅇㅗㅇㅏ" and "ㅇㅗㅏ" as the conjugational form5. Now "ㅇㅗㅏ" is included in the original sentence, so we can conclude that "왔" is composed of "와", which is the conjugational form5 of the verb "오다" and the morpheme "ㅆ".

By decomposing a hangul to unit letters, a word is treated as a sequence of unit letters. For example, we consider the prefinal ending of past tense 'ㅆ다' as a word. This connects with the conjugation form5 of verbs. Therefore we do not need to include two prefinal endings, '았다' and '었다', in the dictionary.

## 4 Morphological analysis over segments

Korean sentences are separated by spaces into phrasal segments. Generally morphological analysis is done only within the segments. However, suppose that we analyze the following sentences.

(a) 이박사의 수 많은 작품들

(b) 미국에 갈 수 있다

We cannot decide whether 수 is a common noun or a bound noun. In such a case, we, therefore, have to take the outside information of a segment into consideration.

Look at the morpheme on the left side of '수' over the segmentation. Sentence (a) has 의/adnominal case particle(ACP), and sentence (b) has ㄹ/adnominal ending(AE) to the left. Due to Korean grammar, the bound noun '수' cannot take ACP to the left. Thus it is clear that the POS(part-of-speech) of '수' in sentence (a) is a common noun. Although the correct POS of '수' cannot be decided unambiguously in many cases like the one in (b), frequent occurrences of the pattern 'ㄹ/AE + 수/BN' strongly suggests that '수' in the sentence (b) is a bound noun.

As is seen in the example, using outside information of segments reduces ambiguities. In our system, connection rules take both segment boundary and morpheme information beyond the segments into account to cope with this ambiguity.

## 5 Analysis method

The Korean morphological analysis system we are developing is called Kocha. The algorithm is based on the minimal cost analysis, where a cost is allocated to each morpheme and connection of morphemes. The lower the cost is, the

more the plausible morphemes or the connections are to occur. When some morphemes may not occur adjacently, the cost of the connection is undefined.

The words in the dictionary and the possible connection of POSs are defined with costs. We consider that a space for segmentation is special kind of word. Thus, it is possible to describe a connection rule between a word and a space. For example, particles cannot locate right after a space. Therefore in the connection dictionary the cost between space and a particle is undefined. Conversely a particle and a space appear frequently in this order, so the cost between a particle and a space is set to be low. We also consider the connectability of the morphemes which appear at the opposite side of a space. We use both connectability of adjacent morpheme and the morpheme over a space.

Samples of a word dictionary, a connection dictionary and a connection dictionary for over segmentation are shown in Tables 5, 6 and 7. Suppose analyzing the sentence "철수는 가지 않는다". '가지' is analyzed either as "가지/noun" or as "가/conjugational form of verb 가다 + 지/conjugational ending". We cannot decide which is better without further information. Also we cannot decide if '않다' is a verb or an auxiliary verb. Our method estimates the plausibility of those possibilities by referring to the costs to each morpheme and to each connection of the morphemes. Both a word and a connection have its defined cost. A connection over segmentation also has a cost. We sum up all the costs of the morphemes and connections for every path. In the calculation of the cost over segmentation, the connection costs defined over segmentation are also added. For instance, in the calculation of the connection cost of '지' and '않는' in Figure 1, the cost sums up 63 (30+30+3). When any of the costs is not defined, the cost is regarded as an infinity. The path with the minimal cost is regarded as the most suitable result.

| POS | word | cost of word |
|---|---|---|
| noun | 가지 | 100 |
| verb | 가다 | 100 |
| | 않다 | 100 |
| aux.verb | 않다 | 100 |
| conj.ending | 지 | 10 |
| space | | 100 |

Table 5: Word dictionary

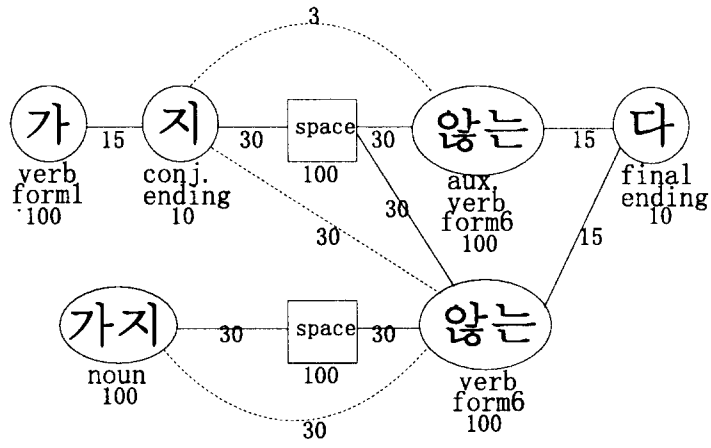


Figure 1: possible connection and costs

| POS | POS | cost of connection |
|---|---|---|
| verb(form1) | conjunctive ending(to form1) | 15 |
| space | verb | 30 |
| space | noun | 30 |
| space | auxiliary verb | 30 |
| noun | space | 30 |
| conjugative ending | space | 30 |

Table 6: normal connection rules

| word(POS) | word(POS) | cost for connection |
|---|---|---|
| 지/conjunctive ending | 않다/auxiliary verb | 3 |
| noun | verb | 30 |

Table 7: Connection rules over segmentation

Figure 1 shows a sample of possible connections and costs in the analysis of "가지 않는다". The dotted lines show that they are connectable with the indicated cost beyond the segmentation. There are three possible paths. Table 8 shows all possible paths and their total costs. Now path (a) has the lowest cost of the three. Thus path (a) is taken as the most suitable path, and actually is a correct result.

Table 8: possible paths and their total costs

| | path | total cost |
|---|---|---|
| path (a) | 가/V + 지/CE + SP + 않는/AV + 다/FE | 413 |
| path (b) | 가/V + 지/CE + SP + 않는/V + 다/FE | 440 |
| path (c) | 가지/N + SP + 않는/V + 다/FE | 415 |

# 6 Application of the rules

We implemented Korean morphological analyzer KoCha by changing the dictionary of our Japanese morphological analyzer, ChaSen(formerly called JUMAN)(Matsumoto 1994) and by modifying the system so as to consider the costs over segmentation. We use Korean grammar and connection rules over segmentation as noted above. Our dictionary has now about 20,000 words (16,000 nouns, 4,000 others). All the dictionary entries are written in unit letters.

We performed an experiment using 50 Korean sentences taken from news paper articles. In this experiment, the system correctly analyzes the POS of Korean sentences in the precision of 97.2% (1242 correct POS tags over 1278 morphemes). When the connection rule over segmentation is not used, precision went down to 96.6% (1235 correct tags).

Our group has developed a visualization tool for ChaSen, called ViCha (Yamashita 1996). We modified it for KoCha. Figure 2 depicts KoCha system running on ViCha. The system show all the possible paths of morphemes in a graph structure, in which the most plausible path is high-lighted.

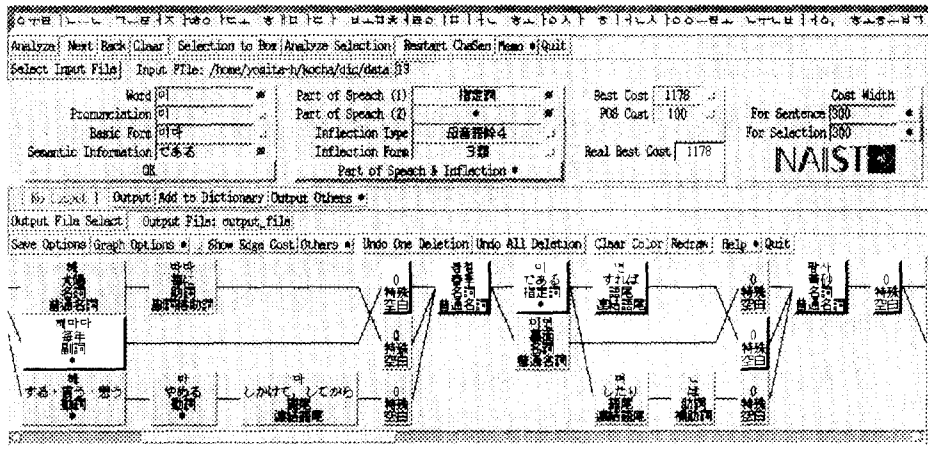


Figure 2: KoCha implemented on ViCha

# 7 Conclusion

We have shown a new verb conjugation system for Korean verbs and its application to morphological analysis. In this grammar, we could make the dictionary of endings be small and could treat the contraction easily. Then we reported the advantage of using the information over segmentation in a Korean morphological analysis.


## Acknowledgement

We are very grateful to Sunao Ueda, Yukitoshi Yutani for providing Korean noun and verb.


## References


Jae-Hoon Kim, Jungyun Seo. 1994. A Korean Part-of-Speach Tag Set for Natural Language Processing (in Korean). KAIST.

Jae-Hoon Kim, Jungyun Seo. 1994. A Practical Morphological Analysis of Korean (in Korean). KAIST

Deok-Bong Kim, Sung-Jin Lee, Key-Sun Choi and Gil-Chang Kim. 1994. A TWO-LEVEL MORPHOLOGICAL ANALYSIS OF KOREAN. *COLING-94* Vol.1.

Jae-Hoon Kim, Byung-Gyu Jang, Gil Chang Kim, Jungyun Seo. 1995. Morphological Ambiguity Reduction Using Subsumption Relation in Koreanf. *NLPRS '95* Vol.1.

Y.Yutani, et al. 1993. KOREAN-JAPANESE DICTIONARY. *Shogakukan, Keumsun*:1986-1993

H.Kanno. Chousengo no Nyuumon (in Japanese). 1993. *Hakusuisha.*

한국산업표준심의회.1992. Code for Information Interchange KS C 5601-1992 (in Korean)

Y.Matsumoto, et al. 1994. Japanese Morphological Analysis System JUMAN Manual (in Japanese). NAIST Technical Report, NAIST-IS-TR94025

T.Yamashita, Y.Matsumoto. 1996. Visual Interface for Morphological Analysis System **ViJUMAN** Manual (in Japanese). NAIST Technical Report, NAIST-IS-TR96005

Y.Yutani. 1990. Hanguru no Kiso (in Japanese). *Taishuukan Shorin*

任瑚彬, et al. 1995. Gaikokujin no tameno Kankokugo Bunpou (in Japanese). Yensei Univ,

R.Kouno. 1979. Kouno Rokurou Chosaku Shuu1 (in Japanese). *Heibonsha*