# KWGInterpreter: A Lemmatizing POS Tagger for the Korean Language

Yongkyoon No
(Chungnam National University)

## 1 Why another Tagger?

Several taggers have been developed for the Korean language over the last decade. Research leading to one or more taggers includes Im, Kim, and Rim (1997), Shin, Lee, and Lee (1997), and Shim and Yang (2002). Some of these are being serviced on the Internet.[1] These taggers view the language from a specific point of view. The grammar they assume is basically the one proposed by Nam and Koh (1987). As descriptively more adequate grammars of the language have been made available since Nam and Koh (1987)[2], it is high time for there to be a new tagger which is based on them.

The existing taggers all analyze a form of the verb into the verb base and a string of suffixes. A verb form is typically analyzed into two or more parts: `s1/VB+s2/EP+s3/EF`. While this analysis does convey some syntactic information realized in the verb form, it fails to convey the information to a full extent. The subset of tags covering verb suffixes used by the above mentioned taggers is motivated by morphotactics only.

[1]We could collect the following URLs which are in operation at the time of this writing.

`http://cl.korea.ac.kr/~dglee/komatag/`

`http://bertha.postech.ac.kr/koma_and_tagger.html`

`http://cs.sungshin.ac.kr/~shim/demo/machdemo.htm`

[2]Chae and No (1998) present a critical survey of the more traditional grammar and offer alternative analyses informed by a sophisticated model of morphology and a monostratal syntactic theory.

This means that the output of the tagger is not particularly geared toward syntactic processing. The fact that a certain suffix follows a certain other suffix is, albeit morphotactically significant, not relevant to syntax. Rather, what syntactic consequence a suffix has is of far greater significance. It is in this respect that the distinctions presupposed by existing taggers of the Korean language have been drawn wrongly.

Described in this paper is a new tagger for Korean, *KWGInterpreter*, I developed for the past couple of years. Section 2 lists POS tags and morphological properties it uses. Section 3 describes how the tagger represents lexemes and what particularly important lexemes there are in the language. The last section gives an example output of a typical run of the lemmatizing tagger on a passage.

# 2 Its Set of Tags

The tags KWGInterpreter uses can be divided into five subgroups. They are described in 2.1 through 2.5.

## 2.1 Nouns and Noun-like categories

Words that may be followed by a postposition fall under one of the following eleven categories:

(1) Classifier CountNoun GridNoun MassNoun Measurer ProcessNoun Pronoun ProperNoun UnmodifiableNoun VAdjunctNoun StateNoun

The distinctions between MassNoun and CountNoun, between Classifier, GridNoun, and Measurer, and between UnmodifiableNoun and others, and between VAdjunctNoun and others are all new. CountNouns may precede a "Numeral + {Classifier, Measurer, GridNoun}" sequence. MassNouns may not. UnmodifiableNouns may not be modified by a relative clause or an adjective. They typically modify a following noun. VAdjunctNouns can be the head of a phrase that functions as an adjunct of a verb without a postposition following them. Noun phrases that are neither the subject of a verb nor a complement of a verb or postposition can be formed when one of VAdjunctNouns is used as the head.

GridNouns, Classifiers, and Measurers are preceded by Numerals or NumeralAdjectives. Measurers denote units of measure, often defined consciously by a group of people. Classifiers denote units of counting and have no origin of their meanings except for the language itself. GridNouns denote spans or areas as they are seen against the background of other spans or areas ordered around them. Numerals preceding GridNouns are ordinal in meaning, while ones preceding a Classifier or Measurer are cardinal in meaning.

ProcessNouns and StateNouns behave similarly to MassNouns, except that they have the priviledge of combining with semantically underspecified words which are often called "light verbs". ProcessNouns come comfortably before a nonstative *ha*, with or without an accusative postposition. StateNouns sit comfortably before a stative *ha*, typically unseparated.[3] These nouns are more similar to UnmodifiableNouns than to MassNouns, in that they occur mostly unmodified.

When a noun belongs to more than one subclass of nouns, KWGInterpreter simplifies its lexicon according to the following rules of precedence.

- If a word belongs to CountNoun and to VAdjunctNoun, do not enter it as a CountNoun.
- If a word belongs to MassNoun and to VAdjunctNoun, do not enter it as a MassNoun.
- If a word belongs to VAdjunctNoun and to UnmodifiableNoun, do not enter it as a UnmodifiableNoun.

A characteristic of KWGInterpreter, that sets it apart from all other taggers for the language is that it does not have the category "bound noun". So-called bound nouns comprise a very heterogeneous group and its members are distributed, in KWGInterpreter, in a wide range of subcategories of nouns: MassNoun, VAdjunctNoun, UnmodifiableNoun, and GridNoun.

[3]Delimiters, *tul*, which marks the plurality of the subject, and *yo*, a marker of moderate deference to the hearer, can intervene.

## 2.2 Morphosyntactic properties marked on the verb

An inflected form of a verb is associated with one of the seven features, which constitute what No (2004) calls MajorClass: RootClauseForm, AdverbialClauseForm, RelativeClauseForm, NounComplementClauseForm, NominalClauseForm, IndirectQuot, and GovernedForm.

(2) Obligatory features from which every verb form has to choose one:

AdverbialClauseForm
IndirectQuot
NominalClauseForm
NounComplementClauseForm
GovernedForm
RelativeClauseForm
RootClauseForm

Those verb forms which realize IndirectQuot or RootClauseForm have also to realize a sentence type, i.e. one of Declarative, Interrogative, Imperative, and Propositive. Certain verb forms are specified for tense, i.e. for Past, RemotePast, or Nonpast, for modality, i.e. for Realis or Irrealis, and evidential, i.e. for Restrospective or Nonretrospective. All forms are specified with respect to honorificity, i.e. for SubjectHonorific or nonhonorific.

Thus, in addition to the features in (2), those in (3) are all in KWGInterpreter's tag set.

(3) Additional features that verb forms may realize:
[ Setence type ] Declarative Imperative Interrogative Propositive
[ Honorificity ] SubjectHonorific Nonhonorific
[ Tense ] Past, RemotePast, Nonpast
[ Modality ] Realis Irrealis Retrospective Nonretrospective

## 2.3 Modifiers

KWGInterpreter recognize the following five categories of modifiers. Adjectives are a closed class and it is different from what the existing

taggers call “adjectives”. [4] A special feature of KWGInterpreter regarding the grammar of Korean it bases itself on is: *hi* is seen as a word of category Adverb. For motivation, see No (2003)

| | |
|---|---|
| Adjective | 새, 무슨, 딴, 어느 |
| Adverb | 자주, 오래, 많이, 꼭, 히 |
| DemonstrativeAdjective | 이, 그, 저, 요, 고 |
| Numeral | 일, 이, 삼, 사 |
| NumeralAdjective | 한, 두, 세, 네 |

## 2.4 Punctuation marks

Punctuation and quotation marks are divided into the following seven categories.

(4) COMMA CParen CQ FPunct IPunct OParen OQ

Parentheses and quotation marks come in two varieties: the opening ones and the closing ones. The punctuation marks which typically come at the end of a sentence are taken to belong to FPunct. The others belong to IPunct.

## 2.5 Other closed classes

Each of the remaining parts of speech consists of a couple of words that play important syntactic roles in the language. Some of these consist of just one word.

| | |
|---|---|
| DM | 요 |
| NM | 여 |
| PM | 들 |
| Complementizer | 고 |
| Delimiter | 은, 는, 도, 만 |
| Interjection | 어, 아이, 참 |

[4] KWGInterpreter does not divide verbs into subclasses. Thus, the traditional distinction between stative and nonstative verbs is not made. “Adjectives” of the existing taggers, which are better called “stative verbs”, correspond to a subset of “Verbs” in KWGInterpreter.

| | |
|---|---|
| PA | 을, 를 |
| PC | 과, 와, 하고 |
| PG | 의 |
| PN | 이, 가, 께서 |
| PO | 에, 한테, 로, 으로, 로서 |
| SS | ㅅ |
| Unknown | |

# 3 Identification of lexemes

## 3.1 Terms of representation

KWGInterpreter is a lemmatizing tagger. It identifies lexemes in the input sentences. The terms of representing a lexeme is its orthographic representation, its POS (or omission of its POS in the case of verbs), and its meaning represented simply as a word in English. While the first two terms are sufficient for distinguishing between most homonymous words, some bad cases of homonymy involve the same phonological words in the same part of speech: the CountNoun *mos* is ambiguous between a 'pond' reading and a 'nail' reading; the verb *ket* is ambiguous between a 'walk' reading and a 'remove from top' reading. The last term, the English word, is chosen from English analogs of the word being represented.

The meanings of some Korean words are so tenuous that English fails to provide their analogs. We take the liberty of representing them with the general-purpose markers GRAMMAR, GRAMMARN, and GRAMMARS. The last two are reserved for what might be called, respectively, nonstative and stative "light verbs".

## 3.2 Verb lexemes with an empty stem

KWGInterpreter recognizes three verb lexemes whose stems are empty. These are exemplified by the following sentences, respectively.

(5) 영희도 떠난답니다
Younghee also leave Nonpast DECL EMPTY ‘say’ Nonpast DECL
‘It is said that Younghee leaves as well’

(6) 영희가 떠나려다가 말았어
Younghee NOM leave ADV EMPTY ‘intend’ ADV quit PAST DECL
‘Younghee intended to leave and then changed her mind’

(7) 영희도 음악가지
Younghee also musician EMPTY ‘be’ Nonpast DECL
‘Younghee, too, is a musician’

The first zero-stemmed verb has a meaning analogous to that of *say* and it occurs only after an IndirectQuot form of another verb. The second has a nearly vacuous meaning, except that it is some sort of dynamic action. The meaning of the sentence relies crucially on the inflection of the verb of the subordinate clause. Thus, the suffix *-려* determines much of the meaning of the empty-stemmed verb in the main clause in (6). KWGInterpreter represents this particular lexeme as `EMPTYSTEM(GRAMMARN)` as opposed to the first, which is `EMPTYSTEM(say)`.

The last of empty-stemmed verb is the copula. This verb has a nonempty stem *이* in some environments. KWGInterpreter represents this verb as `EMPTYSTEM(COPULA)`.

# 4 An example

The tagger is serviced at `http://linguist.cnu.ac.kr:8080/servlets/KWGInterpreter`, and it gives an output when fed with a passage of Korean. An example follows.

| | |
|---|---|
| 물론 | {물론 (VAdjunctNoun NA)} |
| 지방에 | {지방 (MassNoun NA)} {에 (PO GRAMMAR)} |
| 거주하는 | {거주 (ProcessNoun NA)} |
| | {하 (GRAMMARN)[ Realis RelativeClauseForm Nonpast ]} |
| 백성들 | {백성 (CountNoun NA)} {들 (PM GRAMMAR)} |
| 중에는 | {중 (VAdjunctNoun among_during)} {에 (PO GRAMMAR)} |
| | {는 (Delimiter GRAMMAR)} |

국왕의 {국왕 (CountNoun NA)} {의 (PG GRAMMAR)}
행차가 {행차 (ProcessNoun NA)} {가 (PN GRAMMAR)}
있다는 {있 (exist)[ Realis Nonpast Declarative IndirectQuot ]}
{EMPTYSTEM (say)[ Nonpast NounComplementClauseForm ]}
소문을 {소문 (CountNoun NA)} {을 (PA GRAMMAR)}
듣고 {듣 (hear)[ AdverbialClauseForm ]}
부지런히 {부지런 (StateNoun diligent)} {히 (Adverb GRAMMAR)}
경기도로 {경기 (ProperNoun Kyeonggi)} {도 (CountNoun administrative_unit)}
{로(PO to_towards_as)}
올라와 {오르 (go_up)[ AdverbialClauseForm ]}
{오 (come)[ AdverbialClauseForm ]}
민원을 {민원 (CountNoun citizen's_concern)} {을 (PA GRAMMAR)}
접수시키는 {접수 (ProcessNoun NA)}
{시키 (make_let) [ Realis Nonpast RelativeClauseForm ]}
사람도 {사람 (CountNoun person)} {도 (Delimiter GRAMMAR)}
있었다. {있 (exist)[ RootClauseForm Declarative Past ]} {. (FPunct period)}
그렇지만 {그렇 (be_the_case)[ AdverbialClauseForm ]}
지방의 {지방 (MassNoun NA)} {의 (PG GRAMMAR)}
백성들이 {백성 (CountNoun NA)} {들 (PM GRAMMAR)} {이 (PN GRAMMAR)}
국왕의 {국왕 (CountNoun NA)} {의 (PG GRAMMAR)}
행차 {행차 (ProcessNoun NA)}
일정을 {일정 (MassNoun NA)} {을 (PA GRAMMAR)}
맞추어 {맞추 (get_right)[ AdverbialClauseForm ]}
민원을 {민원 (CountNoun citizen's_concern)} {을 (PA GRAMMAR)}
접수시키는 {접수 (ProcessNoun NA)}
{시키 (make_let) [ Realis Nonpast NounComplementClauseForm ]}
것은 {것 (MassNoun GRAMMAR)} {은 (Delimiter GRAMMAR)}
결코 {결코 (Adverb NA)}
쉬운 {쉽 (be_easy)[ Realis RelativeClauseForm Nonpast ]}
일이 {일 (MassNoun task)} {이 (PN GRAMMAR)}
아니었다. {아니 (be_not)[ RootClauseForm Declarative Past ]}
{. (FPunct period)}

**References**

Chae, Hee-Rahk and Yongkyoon No. 1998. A survey of morphological issues in korean: Focusing on syntactically relevant phenomena. *Korean Linguistics*, 9:65–109.

Im, Heeseok, Jindong Kim, and Haechang Rim. 1997. ecel thayku pyenhyeng kyuchikul iyonghan hankwuke phumsa thayke [a tagger for korean using tag transformation rules]. *cengpokwahakhoy nonmunci [Journal of Information Science Society]*, 24(6):673–684.

Nam, Kishim and Yeunggun Koh. 1987. *Phyocwun kwuke munpeplon [A standard grammatical theory of Korean]*. thap chwulphansa.

No, Yongkyoon. 2003. 'nwuna mollay tolul tencica'uy kwusengsengpun kwucowa phasayng pusauy hawipemcwuhwa [the constituent structure of 'nwuna mollay tolul tencica' and subcategorization properties of derived adverbs]. Paper presented at 2003 Winter Meeting of the Linguistic Society of Korea, February.

No, Yongkyoon. 2004. Nominative/accusative adpositions in negative auxiliary constructions. *Enewa cengpo [Language and Information]*, 8(2):73–91.

Shim, Kwangseob and Jaehyung Yang. 2002. Mach: A supersonic korean morphological analyzer. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING-2002)*, pages 939–945.

Shin, Sanghyeon, Geunbae Lee, and Jonghyeok Lee. 1997. thongkyeywa kyuchikey kipanhan 2tankyey hankwuke phumsa thayking sisutheym [a two-level pos tagging system based on statistics and rules]. *cengpokwahakhoy nonmunci [Journal of Information Science Society]*, 24(2):160–169.