Sejong Korean Corpora in the Making

Beom-mo Kang and Hunggyu Kim

Korea University
Seoul, 136-701 Korea
bmkang@korea.ac.kr, kimhg@ikc.korea.ac.kr

Abstract

We introduce a set of Korean corpora in the making. One of them is a corpus consisting of morphologically analyzed Korean words and it is called "Sejong Morph Tagged Corpus". It is a part of Sejong Corpora, which are the results of a government-sponsored language resources compiling project in Korea. We give an outline of the corpus building component of the project and describe in some detail "Sejong Morph Tagged Corpus". The latter is being further processed for disambiguation to be turned into "Sejong Morph Sense Tagged Corpus" and into a Korean Treebank of syntactically parsed sentences.

Corpora in the 21st Century Sejong Project

The 21st Century Sejong Project is a comprehensive project aiming to build various kinds of language resources including Korean corpora, comparable to BNC (Aston & Burnard, 1998), and Korean electronic dictionaries. The project was conceived of in 1997 and started in 1998 as a 10-year long-term project. By 2003, we completed 6 years of our work.

The Sejong Corpora are a collection of raw corpora of modern Korean (written and spoken), North Korean, Korean used abroad, old Korean, and oral folklore literature. They also include parallel corpora consisting of Korean and other languages such as English and Japanese. Among these, a morph tagged corpus is a central part. In the process of compiling these corpora we followed suggestion from Text Encoding Initiative (TEI, Sperberg-McQueen & Burnard, 1994) to a certain degree.

By 2003, we compiled a modern Korean raw corpus of 57 million words. We have additional 75 million words of already existing electronic texts which were processed and standardized in the first year of the Sejong project. These raw texts are mostly written Korean. We have relatively small amount, around 3 million, of spoken words. The morph tagged corpus is morphologically analyzed written Korean, around 10 million words by the end of 2003. The morph sense tagged corpus, which is the result of disambiguation of morphs, has 5.5 million words. From 2002 we started to build a treebank, i.e. syntactically analyzed Korean sentences on the basis of simple phrase structure grammar rules. Currently, we only have 0.15 million words being part of syntactic trees.

Written corpora, i.e. a raw corpus of modern Korean, a morph tagged corpus, a morph sense tagged corpus, and a treebank, have been compiled at Center for Electronic Texts of Korea University. The following table is a summary.

(1) Written Parts of Sejong Corpora by 2003

raw corpus/written	57.0	million
	(+75.0	million)
morph tagged corpus	10.0	million
morph sense tagged corpus	5.5	million
treebank	0.15	million

Sejong Morph Tagged Corpus

At the first stage of morphological analysis and tagging, we tagged only written texts. Later, Yonsei university, another project participant, stated to work on spoken texts and produced some 30 thousand morphologically analyzed words. We, Korea University, have been working on written texts and in this paper we have little to say about the spoken part, except that they adopted the same tags and added some more in consideration of characteristics of spoken texts.

English POS tagged corpora such as LOB Corpus and Brown Corpus (Francis and Kucera, 1982) have tags for the whole word-forms (e.g. talked_VVD). This method is understandable since English has a simple inflectional system. In contrast, Korean POS tagged corpora need to have words morphologically analyzed because of many inflectional morphemes. For example, 'geoleosseo' ("walked") has three parts: a verb stem (VV), a prefinal ending (EP) and a word-final ending (EF).

(2) geoleosseo ("walked") :

geod_VV +	eoss_EP +	eo_EF
walk	PAST	DECLARATIVE

Notice that the verb stem undergoes a phonological change: d → l.

Here are the tags we used in the project. The tags were prepared in the first year of the 21C Sejong Project by Im & Song (1998).

(3) List of Tags for Morph Tagged Corpus

category	-subcategory	tag
noun	-common noun	NNG
	-proper noun	NNP
	-bound noun	NNB
pronoun		NP
numeral		NR
verb		VV
adjective		VA
auxiliary		VX

"be"	-positive	VCP
	-negative	VCN
determiner		MM
adverb	-general	MAG
	-conjunctive	MAJ
interjection		IC
case marker	-subject	JKS
	-complement	JKC
	-genitive	JKG
	-object	JKO
	-adverbial	JKB
	-vocative	JKV
	-quotation	JKQ
discourse particle		JX
conjuctive particle		JC
ending	-prefinal	EP
	-final	EF
	-connective	EC
	-nominal	ETN
	-modificational	ETM
prefix		XP
suffix		XS
base (root)		XR
miscellaneous symbols including		
	-foreign alphabet	SL
	-Chinese character	SH
	-many others	SF, SP, etc.

Some of the POS tags (morpheme categories) that we used are on the level of parts of speech in school grammar (verb, adjective), and some are more detailed than parts of speech (common noun, proper noun, bound non, etc.). Nominal case markers and verbal endings are classified rather in detail since these are the most important elements in Korean morphology and grammar. For example, case markers are differentiated into subject, complement, genitive, object, adverbial, and vocative case markers and endings are classified into prefinal, final, connective, nominal, and modificational endings. Very productive derivational morphemes, i.e. prefixes and suffixes, are analyzed, too. Among these are several kinds of suffixes which turn some nouns into verbs and adjectives.

Sample data are give in Figure 1.

이용자들은	이용자/NNG + 들/XSN + 은/JX
컴퓨터를	컴퓨터/NNG + 를/JKO
통하여	통하/VV + 아/EC
전자정보	전자/NNG + 정보/NNG
시스템에	시스템/NNG + 에/JKB
연결하여	연결/NNG + 하/XSV + 아/EC
필요한	필요/NNG + 하/XSA + ㄴ/ETM
정보를	정보/NNG + 를/JKO
즉시	즉시/MAG
받아볼	받/VV + 아/EC + 보/VX + ㄹ/ETM
수	수/NNB
있다.	있/VV + 다/EF + ./SF

미래의	미래/NNG + 의/JKG
학습환경,	학습/NNG + 환경/NNG + ,/SP
그	그/MM
선택과	선택/NNG + 과/JC
책임	책임/NNG
현대의	현대/NNG + 의/JKG
테크놀로지는	테크놀로지/NNG + 는/JX
분명	분명/MAG
우리에게	우리/NP + 에게/JKB
학습환경의	학습/NNG + 환경/NNG + 의/JKG
신세계를	신/XPN + 세계/NNG + 를/JKO
열어주고	열/VV + 어/EC + 주/VX + 고/EC
있다.	있/VV + 다/EF + ./SF

Figure 1: Morph Tagged Corpus Data

The first column contains a word-form and the rest is a sequence of "morph/TAG" pairs. Except for adverbs (MAG), conjunctions (MAJ) and other independent morphs, most of word-forms are composed of a root (noun NNG, verb VV) followed by one or more affixes (case markers JK, endings E) and possibly a punctuation mark such as a comma and a period. Since case markers and endings are identified on the level of (allo)morphs rather than morphemes, the corpus is called a "morph" tagged corpus rather than a "morpheme" tagged corpus. For example, the subject marker has two allomorphs '-ga' and '-i' according as the preceding sound is a vowel or a consonant. In the corpus, morphological analysis preserve these two forms, which can be automatically transformed into a single morpheme when needs arise.

Sejong Morph Sense Tagged Corpus

Sejong Morph Tagged Corpus described above has the problem of ambiguity. Only grammatical or morphological categories, not meanings, are considered. Of course, there are many homonymous words in Korean with the same part of speech, like two English nouns of 'bank'. For example, Korean word-form 'eunhaeng' means either "a bank" or "a ginko (nut)". Since Korean has a relatively simple syllable structure of (C)V(C) and most Korean nominals are composed of two syllables, Korean has more nominal homonyms than English. But unlike English, nouns and verbs/adjectives have different inflections and cause little N/V ambiguity prevalent in English (e.g. convict n / convict v).

Certainly we need to disambiguate the tagged corpus for correct word frequency data and for other purposes. Sejong Morph Sense Tagged Corpus is such a disambiguated corpus, with word-forms disambiguated on the dictionary entry level. That is, words which are listed as separate entries in the Standard Korean Dictionary are distinguished and identified by the entry number in the dictionary in the case of homonyms. For example, 'mal' in the sense of "language" is marked as "mal_01" and 'mal' in the sense of "horse" is marked as 'mal_05'. (Incidentally, there are 12 entries with the form of 'mal', some of which are scarcely used.)

This kind of disambiguation is done for words of major lexical categories: nouns (NNG), verbs (VV), adjectives (VA), adverbs (MAG), determiners (MM),

and noun-like roots (XR). The procedure of disambiguation is mostly manual work of examining concordance lines of potentially ambiguous word forms. Before examining each instance of a word-form, concordance lines are sorted according the word-forms of keyword and forms of adjacent words. Because collocational patterns tend to be different for each word (lexical entry), instances of a word (lexical entry) flock together, which makes the manual disambiguation task much easier. For example, homonymous word-form 'eunhaeng' is to be identified as a word for "jinko" rather than a word for "bank" when used with a verb 'simda'("to plant").

Sample data is given in Figure 2.

수백만.	수/MM + 백만/NR + ./SF
수천만의	수/MM + 천만/NR + 의/JKG
코를	코__01/NNG + 를/JKO
자극	자극__01/NNG
시킨	시키/VV + ㄴ/ETM
그	그/MM
원흉에	원흉/NNG + 에/JKB
이제와서	이제__01/NNG + 와서/NNG
새삼	새삼/MAG
법석을	법석__01/NNG + 을/JKO
떠는	떨__01/VV + 는/ETM
당국의	당국__02/NNG + 의/JKG
태도에	태도__03/NNG + 에/JKB
시각을	시각__04/NNG + 을/JKO
잡아보고	잡__01/VV + 아/EC + 보/VX + 고/EC
싶은	싶/VX + 은/ETM
것이다.	것/NNB + 이/VCP + 다/EF + ./SF

Figure 2: Morph Sense Tagged Corpus

Notice that homonymous words are disambiguated by the entry number attached to the right of a morph. Also note that out of 17 word-forms in the above example, we have as many as 9 potentially ambiguous words.

Now that we have a disambiguated corpus of more than 5 million words, we are able to compile a frequency list of lemmas (Kang & Kim, 2004), much more valuable data than a frequency list based on a (ambiguous) morph tagged corpus (Kim & Kang, 2000).

Sejong Treebank

In 2002, when we started to build Sejong Treebank, we parsed sentences composed of some 30,000 thousand words in total. On average a sentence has about 10 words. Now that headings of one or two words are included in the calculus, many sentences are over 10 words and some are very long.

In 2003, the number of words grew up to 150, 000. In our project we adopted the following analysis methods.

1) Only surface sentence structures are considered. Namely, transformations do not play a role.

2) Empty elements such as traces and null pronouns are not identified.

3) Only binary branching is allowed, so that no node is composed of three or more nodes in the tree.

4) complements and adjuncts are partially distinguished in the sense that only subjects, objects, and complements of verbs 'doeda' (become) and 'anida' (not be) are clearly marked.

5) The parsed tree is annotated with tags which show both categories and (grammatical) functions.

Let us elaborate on the last point. Mostly, a tag for a node is composed of two parts, showing its syntactic category and its grammatical function (relation). Here are the list of major structural tags and the list of major functional tags.

(4) structural tags

S	sentence
NP	noun phrase
VP	predicate (verb, adjective) phrase
AP	adverbial phrase
DP	determiner phrase
IP	interjection phrase

(5) functional tags

SBJ	subject
OBJ	object
CMP	complement (of verbs of "be, become")
MOD	modifier
AJT	adjunct

For example "NP_SBJ" stands for a node of noun phrase functioning as subject, and "VP_MOD" stands for a node of predicate phrase modifying another expression (noun). Some node is marked only by a structural tag because the function is predictable. For example, a VP without any other functional tag is a predicate from the viewpoint of grammatical function.

The analysis tree of a simple sentence in (6) with a subject, an object, and a transitive verb is given in (7) (SBJM: subject marker, OBJM: object marker).

(6) John-i Mary-leul mannassta.
J-SBJM M-OBJM met
'John met Mary.'

(7)
```
(S (NP_SBJ  John/NNP + i/JKS)
   (VP (NP_OBJ  Mary/NNP + leul/JKO)
       (VP  manna/VV + eoss/EP + da/EF + ./SF)))
```

The parsing is based on morph tagged texts, which are part of Sejong Morph Tagged Corpus mentioned above. The parsing and annotating procedure is a mixture of manual and automatic methods. A computer program offers a parsing when possible and the annotator checks if it is correct.

The parsed sentences are stored in the form shown in Figure 3. The whole sentences is given first and then the result of the syntactic analysis.

Because the annotation includes both syntactic categories and grammatical functions, the parsed trees can be easily converted into dependency structures of dependency grammar. As a matter of fact a computer program exists which achieves this task automatically. In principle, converting from dependency structures into constituent structures is not possible but the other

```
공동체를 결속시키는 불의 기능은 고대 희랍이나 로마에서도 예외는
아니었다.

(S (NP_SBJ   (VP_MOD   (NP_OBJ 공동체/NNG + 를/JKO)
                       (VP_MOD 결속/NNG + 시키/XSV + 는/ETM))
           (NP_SBJ     (NP_MOD 불/NNG + 의/JKG)
                       (NP_SBJ 기능/NNG + 은/JX)))
   (VP   (NP_AJT   (NP 고대/NNG)
                   (NP_AJT   (NP_CNJ 희랍/NNP + 이나/JC)
                             (NP_AJT 로마/NNP + 에서/JKB + 도/JX)))
         (VP   (NP_CMP 예외/NNG + 는/JX)
               (VP 아니/VCN + 었/EP + 다/EF + ./SF))))
```

Figure 3: Treebank Data

direction is possible when proper information about grammatical functions are provided for unclear cases. This is why we chose the current way of annotation instead of adopting dependency structure annotation.

Korean, like any other languages, have various kinds of grammatical structures and constructions, including arguments, adjuncts, modifiers, auxiliaries, causatives, and displaced elements. How sentences with these constructions are to be syntactically analyzed under the current annotation scheme is not always clear. We have been working hard to provide some workable guidelines, the discussion of which is beyond the scope of this paper.

Acknowledgments

This work is supported by the 21C Sejong Project sponsored by The Ministry of Culture and Tourism of Korean Government. We thank the student assistants of Center for Electronic Texts, Korea University, who have been working in the making of Sejong corpora.

References

Aston, G. & Burnard L. (1998) The BNC handbook: Exploring the British National Corpus with SARA, Edinburgh: Edinburgh University Press.

Francis, W. N. & Kucera, H. (1982) Frequency analysis of English usage: Lexicon and grammar, Boston: Houghton Mifflin Co.

Im, H. & Song, C. (1998) Tags for morphological analysis. Report of the 21C Sejong Project - 1st year, Ministry of Culture and Tourism. [written in Korean]

Kang, B. & Kim, H. (2004) Frequency analysis of the use of Korean morphemes and words 2, Seoul: Institute of Korean Culture, Korea University. [written in Korean]

Kim, H. & Kang, B. (2000) Frequency analysis of the use of Korean morphemes and words 1, Seoul: Institute of Korean Culture, Korea University. [written in Korean]

Kim, H. & Kang, B. (1996) Korea-1 Corpus: design and composition. Korean Linguistics. [written in Korean]

Sperberg-McQueen, C.M. & Burnard L. (eds.) (1994) Guidelines for electronic text encoding and interchange, Chicago: TEI.