KAIST Tree Bank Project for Korean: Present and Future Development

Key-Sun Choi*, Young S. Han*, Young G. Han†, and Oh W. Kwon*

* *Center for Artificial Intelligence Research*
Computer Science Department
Korea Advanced Institute of Science and Technology
Taejon 305-701 Korea
{*kschoi, yshan, yghan, ohwoog*}*@csking.kaist.ac.kr*
† *Department of Koean Language and Literature*
Ulsan University Ulsan 680-749 Korea

Abstract

In this paper, we introduce the on-going project for building a large annotated corpus of Korean written texts undertaken by KAIST[1] since 1992. At present, the corpus consists of over 5 million word units of Korean covering 13 subject fields. The POS tagset used to annotate the corpus contains 74 tags. The tagset is designed to provide sufficient recoverability even for high level linguistic processing. Current efforts are put mostly on bracketing the corpus. The corpus is expected to be distributed through the Korean Linguistic Data Consortium before the end of 1994.

Summary of The Resource

Name	KAIST Tree Bank
Documents	Project Reports in Korean
Price	Not decided yet
Limitation	Academic use only
FTP site	Not set up yet
Media	8mm, MO, Floppy Disk
Format	Unix tar, ms dos text file
Contact Person	Key-Sun Choi
Style of Demonstration	video, or PC-to-OHP
Language	Korean
Data Type	corpus, tagged corpus
Field	Balanced
Size (Mbyte)	10 Mbytes (first release)

[1] Korea Advanced Institute of Science and Technology

1. Introduction

Recently there have been intense interests in building large corpora of Korean texts among several academic institutions. The projects may differ in the application goals. Some are pursued to be used in lexicographic study which will eventually lead to the construction of dictionaries either for human or for machine. Some are more linguistically oriented and designed for the modeling of various levels of language. The aim of KAIST tree bank project has been set up with both the lexicographic and analytic studies in mind since the project took off at 1992. To capture the two aspects implies the corpora should be sufficiently large and carefully balanced to reflect the general usage of the language if general domain is adopted.

The KAIST tree bank project[2] is initiated with the goal of constructing more than 10 million word units of Korean that includes at least 5 million bracketed units to be made public. The public introduction of the products will be made incrementally for the academic use as the work progresses. For now, the annotations are limited to tagging and bracketing. Current status of the project is still in the stage of collecting texts with about 6 million word units accumulated and about 1 million units tagged. The first release of 1 million units is expected in the near future. In the following sections, the categories of texts and the annotation schemes are explained.

2. Composition of the Corpus

The modern Korean texts are classified into five groups, and each group is further divided into several categories. Note the corpus as we defined is supposed to contain not so much knowledge about the world as linguistic information. Our philosophy of balancing the corpus is to reflect the distribution of each category as observed in the modern Korean texts. We also want the distribution should not be very different from the linguistic competency of an average Korean. This is because the use of the corpus is intended not only for the linguistic study but also for the cognitive study. Different goals may suggest different composition of corpus. For the purpose of linguistic study, the bigger the corpus is, the better is the chance of getting the linguistic patterns from written texts of general domain. An individual account of language usage, however, may not necessarily follow the global distribution. An individual is at best skewed with respect to the maximal linguistic explanation. This implies that the statistical information from the large general corpora may not be accurate in modeling the linguistic behaviour of individuals.

In practice, developing intelligent systems using natural language techniques will almost be based on a particular domain. It seems reasonable that the construction of corpus should be coupled with well defined goals, and in the long run corpora will become more specific than we find them to be today. The goal of KAIST initiative is

[2]This project is a joint effort with Automatic Interpretation Laboratory of Electronic and Telecommunication Research Institute, and Center for Artificial Intelligence Research.

set to general texts by which we mean not something conclusive, but something basic to most applications. How to implement the notion of generality into the balance of the corpus will be a good socio-linguistic problem. Though little is known as to this problem in the literature, there can be at least two approaches. The first option is to make assumption on how people are exposed to texts, and the second is to observe the real world of texts.

The first option makes sense in that no theoretical principle exists for the balancing. In the second option, we examine how an individual or population is exposed to the printed texts. For example, an observation of text printings by press and publishers each year gives a distribution of the texts, but strong dependency on this indication can be misleading because most of the individuals are biased to their environment and only weakly affected by the total distribution of texts. The proportion of daily or weekly papers will easily outdo other categories. On the other hand, the fact makes the proportion less convincing that a good sector of population dispense with the contemporary papers and our sophisticated linguistic skills are mostly the results of the secondary or higher education.

The balance for KAIST tree bank is determined by assuming that literature, news paper, and academics are the most important segments of printed materials that affect the linguistic contents of most Koreans, and they are equally influential. Table 1 shows the composition of the 1 million units of the corpus soon to be released. A word unit is a unit of spacing in Korean and is usually a combination of words and functional morphemes.

Another problem follows after the decision of balancing that is collecting and selecting texts. For the clarity of the process, the category is further divided into subcategories. Selecting articles is still left to the human discretion. A linguist and several well trained assistants were involved in selecting the articles representative of the category in our project. No less trivial is the problem of clearing the copyrights of the selected texts, and there being no short cut, it needs to be solved with much patience.

3. Annotations

The annotations of large corpora aim at extracting linguistic information, thus they are as diverse as the applications. In general, POS tagging and bracketing are the two most popular annotations. The set of POS tags is composed of 6 classes that are as in table 2.

As a bootstrap, 0.1 million units were manually tagged, then the results were fed into an automatic tagger based on HMM (Lee, Choi, Lim, and Lee 94). The accuracy of 94% is reached at present, but is not sufficient enough to be useful. The cycle of manual correction and automatic tagging is still in process. We expect 1 million units to be tagged with about 95% accuracy before the first release. Figure 1 shows the tagged example. The tag set initially consisted of 51 tags, and is trimmed to a set of 51 tags. As to the coding format of the corpus, presently texts are stored in plain

category	word units	ratio %
1. Korean Literature		
poetry	80,000	8
fiction, essay	180,000	18
comedy, news broadcast	50,000	5
total	*310,000*	*31*
2. News Paper		
editorial	100,000	10
review	100,000	10
article	100,000	10
total	*300,000*	*30*
3. Academics		
history, philosophy	60,000	6
social science, economics, politics	60,000	6
natural science	60,000	6
fine arts	50,000	5
total	*230,000*	*23*
4. Text Books	70,000	7
5. Religion	70,000	7
6. Unpublished Writings	20,000	2
total sum	1,000, 000	*100*

Table 1: Balance Sheet for the 1M unit Corpus.

text files with athographic information and linguistic markings, and Korean Standard Codes (KSC5601) are used for the portability of code transformation to other character codes. It has been strongly proposed that the files be coded using markup language such as SGML and put into a database system.

Tagging Korean texts is different from tagging English in the following points making it more difficult.

- Heavy morphological analysis is needed so that some morphemes may be recovered to their original forms without which tags cannot be determined.
- There are two levels of tag sequences: within a word unit and between word units.

Bracketing still in its design stage is one of the major works left for the rest of the project. Bootstrap corpus is bracketed first, then CFG rules are extracted from the samples. Using the rules, more samples are bracketed followed by manual correction. The procedure will repeat till reasonably accurate rules are acquired. It turned out that CFG rules of dependency relation are readily available since dependency between words is relatively easy to identify. It is expected that the 1M unit corpus can be bracketed once current bootstrapping stage is completed. From the bracketed corpus, Korean

Tag name	Full Name	Examples
Nouns		
nc	[common noun]	
nct	[time]	
nca	[action]	사랑(하) (sarang(ha)), 공부(하) (kongpwu(ha))
ncs	[stative]	청결(하) (chengkyel(ha)), 미안(하) (mian(ha))
nq	[proper noun]	
nb	[bound noun]	(-한) 것 ((han) kus), (-할) 수 ((hal) swu)
nbu	[unit]	(세) 개 ((sey) kae), (일곱) 마리 ((ilkop) mari)
npp	[personal pronoun]	저희 (cehui), 우리 (wuri), 그들 (kutul)
npd	[demonstrative]	거기 (keki), 그때 (kuttae), 이때 (ittae)
nnn	[number]	1, 3, 1994
nn	[numeral]	하나 (hana), 둘 (tul), 세 (sey), 열둘 (yeltul)
Predicates		
pv	[predicate, verb]	자르(다) (caru(ta)), 밀(다) (mil(ta))
pa	[pred., adjectives]	노랗(다) (norah(ta)), 깨끗하(다) (kkaekkusha(ta))
pad	[demonstrative adjective]	그렇(다) (kureh(ta)), 아무렇(다) (amwureh(ta))
px	[auxiliary]	(-게) 되(다) ((key)toi(ta)), (-지) 말(다) ((ci)mal(ta))
Modifiers		
md	[demonstrative adnoun]	그 (사람) (ku(saram)), 이 (나무) (i (namwu))
mn	[numeral adnoun]	한 (사람) (han (saram)), 두 (명) (twu(myeng))
m	[adnoun]	새 (물건) (sae (mwulken)), 옛 (서적) (yeys(secek))
Adverbs		
at	[time adverb]	내일 (naeil), 일찍 (ilccik), 이미 (imi)
ad	[demonstrative]	이리 (iri), 어디 (eti), 요리 (yori), 여기 (yeki), 어찌 (ecci)
ajw	[word-conjunctive]	또는 (ttonun), 및 (mic)
ajs	[sentence-conjunctive]	그러나 (kurena), 이른바 (irunpa), 특히 (thukhi)
a	[adverb]	매우 (maewu), 과연 (koayen), 반드시 (pantusi)
Independents		
i	[interjection]	예 (yey), 아이구 (aikwu), 여보세요 (yeposyeyo)
Particles		
jc	[case particle]	-이 (i), -가 (ka), -를 (rul), -에서 (eyse)
jcm	[adnominal]	-의 (ui), -까지의 (kkaciui), -께로의 (kkey)
jcv	[vocative]	-야 (ya), -여 (ye), -시여 (siye), -이야 (iya)
jca	[adverbial]	-에게 (eykey), -한테 (hanthey), -대로 (taero)
jcp	[predicative]	-이(다) (i(ta)), -이(로구나) (i(rokwuna))
jx	[auxiliary]	-도 (to), -는 (nun), -까지 (kkaci), -ㄴ랑 (nrang)
jj	[conjunctive]	-와 (oa), -과 (koa), -며 (mye), -에다 (eyta), -이랑 (irang)

Table 2: Tag set list with examples.

Endings		
ecq	[equal conjunctive ending]	-거나 (kena), -느니 (nuni),-(하)며 (mye)
ecs	[subordinate]	-도록 (torok), -면 (myen), -으려 (urye)
ecx	[auxiliary]	-게 (되다) (key (toita)), -지 (않다) (ci (anhta)
exm	[adnominal transformation]	-던 (ten), -은 (un), -을 (ul), -는 (nun)
exn	[nominal transformation]	-기 (ki), -음 (um), -ㅁ (m)
exa	[adverbial transformation]	-(하)게 ((ha) key)
efp	[prefinal]	-시- (si), -었- (ess), -옵- (op), -던- (tun)
ef	[final]	-는구나 (nunkwuna), -어라 (era), -습니다 (sumnita)
Suffixes		
xn	[noun suffix]	(사람)들 ((saram)tul), (김)씨 ((kim)ssi)
xpv	[verb derived suffix]	(위반)하(다) ((wuipan)ha(ta)),
		(높)이(다) ((noph)i(ta)), (밝)히(다) ((park)hi(ta))
xpa	[adjective derived suffix]	(가난)하(다) ((kanan)ha(ta))
		(정성)스럽(다) ((cengseng)surep(ta))
xa	[adverb derived suffix]	(조용하)게 ((coyongha)key), (간단)히 ((kantan)hi)
Symbols		
s,	[comma]	
s.	[sentence closer]	
s‘	[left quotation or left parenthesis]	
s’	[right quotation or right parenthesis]	
s-	[connection]	
su	[unit symbol]	
sw	[currency unit symbol]	
sy	[other symbols]	

Table 3: Tag set list with examples continued.

grammar of dependent CFG rules can be composed to be used in testing a couple of probabilistic parsing methods we have developed. Figure 2 shows the bracketed output of the example sentences.

4. Conclusion and Future Plan

The project is still many years away till the reasonable completion. Completed annotated corpus will contain more than 10 million word units from balanced Korean texts, and 5 million bracketed units. In fact, the project is a part of bigger initiative in which full scale lexicographic and linguistic studies are also included. The expected products from the studies are, for example, various statistics, linguistic patterns, data models, and language models. Thus far, corpus analysis tools including concordance package, morphological analyser, and automatic tagger are developed, and bracketing system is

Input sentences:

("컴퓨터 화면 한 구석에 나타난 숫자가 빠르게 올라갔다.")
("khemphywuthe hoamyen han kwuseke nathanan swuscaka pparukey olakassta.")
("화면 중앙에는 럭비공 모양의 영상이 천천히 돌고 있었다.")
("hoamyen cwungangeynun lekbikong moyangi chenchenhi tolko issessta.")

Tagged sentences:

(컴퓨터/nc 화면/nc 한/mn 구석/nc 에/jca 나타나/pv ㄴ/exm 숫자/nc 가/jc 빠르/pa 게/exa 올라가/pv 았/efp 다/ef ./s.)
(화면/nc 중앙/nc 에는/jca 럭비공/nc 모양/nc 의/jcm 영상/nc 이/jc 천천히/a 돌/pv 고/ecx 있/px 었/efp 다/ef ./s.)

Fig 1: Tagged example sentences.

under way.

The public release of the output from the project may include analysis tools as well as corpora, but its formal setting is not set up yet. The KAIST trees are expected to be done at the end of 1995, and the complete version will be released at early 1996 through Korean Linguistic Data Consortium that is operated by the ministry of culture and ministry of science and technology.

Acknowledgement

In fact, the work reported in this paper is the result of the inspiration of too many people to be listed in the title. In particular, we thank Hiongun Kim for his effort to build bracketed corpus.

References

Lee, W. J.; Choi, K. S.; Lim, Y. J.; and Lee, Y. J. (1994). An automatic tagging system and environment for the construction of Korean text database, to appear in *WESTPRAC-V* , August 94.

```
(  (  (  (  (  ( 컴퓨터/nc
                 화면/nc )
               {  한/mn
                  구석/nc   에/jca } )
            나타나/pv   ㄴ/exm )
       숫자/nc   가/jc )
       빠르/pa  게/exa
    올라가/pv   았/efp   다/ef )
./s. )
```

```
(   (   ( 화면/nc
        중앙/nc  에는/jca )
        ( ( 럭비공/nc
          모양/nc   의/jcm )
        영상/nc   이/jc )
        천천히/a
        { 돌/pv   고/ecx
    있/px   었/efp   다/ef } )
./s. )
```

Dependency CFG rules

Nc	$\longrightarrow$	Nc Nc	; 컴퓨터 화면 (khemphywuthe hoamyen)
NcJca	$\longrightarrow$	Mn NcJca	; 한 구석에 (han kusekey)
NcJca	$\longrightarrow$	Nc NcJca	; 화면 구석에 (hoamyen kusekey)
PvExm	$\longrightarrow$	NcJca PvExm	; 구석에 나타난 (kusekey nathanan)
NcJc	$\longrightarrow$	PvExm NcJc	; 나타난 숫자가 (nathanan swuscaka)

Fig 2: Bracketed example sentences.