October 2005

Guidelines for Penn Korean Treebank Version 2.0

Na-Rae Han
University of Pennsylvania, nrh@ling.upenn.edu

Shijong Ryu
University of Pennsylvania

University of Pennsylvania Institute for Research in Cognitive Science Technical Report No. IRCS-5-03.

Guidelines for Penn Korean Treebank Version 2.0

Abstract

The Korean Treebank Annotations Version 2.0 is a second volume of The Korean Treebank Annotations (Palmer et al., 2002; Han et al., 2002). It contains new texts that are from the news domain: the original corpus for the Korean Treebank 2.0 was extracted from The Korean Newswire corpus published by LDC, catalog number LDC2000T45. The Korean Treebank Annotations Version 2.0 consists of 647 news articles in 112 files which contain 132,040 words and 5,010 sentences. There are 40,252 unique words and 13,844 unique morphemes (12,681 unique morphemes excluding foreign characters and arabic numbers). The annotated text measures about 8.5MB in size.

While annotating the new texts, many new linguistic constructions and phenomena were encountered which called for setting additional guidelines. Furthermore, a few guidelines used for the first volume of the Korean Treebank were re-examined and modified in the second volume. This document outlines the guidelines that were newly introduced for the second volume of the Penn Korean Treebank, as well as the ones that have been revised since the publication of volume 1.0. Therefore, this is not a self-contained document, but is rather an addendum to the two previously published guidelines for the Penn Korean Treebank (Han and Han, 2001; Han et al., 2001).

Comments

University of Pennsylvania Institute for Research in Cognitive Science Technical Report No. IRCS-5-03.

Guidelines for Penn Korean Treebank Version 2.0

Na-Rae Han, Shijong Ryu

October 20, 2005

Contents

1 Introduction

The Korean English Treebank Annotations (Palmer et al., 2002; Han et al., 2002) is an electronic corpus of Korean and English parallel texts annotated with morphological and syntactic information. Annotation of the Korean part of the corpus was done in accordance with two published guidelines: "Part of Speech Tagging Guidelines for Penn Korean Treebank" (Han and Han, 2001)"Bracketing Guidelines for Penn Korean Treebank" (Han et al., 2001). The corpus consists of Korean and English bilingual texts extracted from military training manuals.

The Korean Treebank Annotations Version 2.0 (Han et al., to be published) [1] is a second volume of the corpus, and it contains new texts that are from the news domain. The original corpus for the Korean Treebank 2.0 was extracted from The Korean Newswire corpus published by LDC, catalog number LDC2000T45. The Korean Newswire corpus is a collection of Korean Press Agency news articles from June 2, 1994 to March 20, 2000. The texts included in the Korean Treebank 2.0 was selected from the March 2000 portion of the news articles (files `20000302.SGM` – `20000320.SGM`). [2]

The corpus consists of 647 news articles in 112 files which contain 132,040 words and 5,010 sentences. There are 40,252 unique words and 13,844 unique morphemes (12,681 unique morphemes excluding foreign characters and arabic numbers). The annotated text measures about 8.5MB in size.

While annotating the new texts, many new linguistic constructions and phenomena were encountered which called for setting additional guidelines. Furthermore, a few guidelines used for the first volume of the Korean Treebank were re-examined and modified in the second volume. This document outlines the guidelines that were newly introduced for the second volume of the Penn Korean Treebank, as well as the ones that have been revised since the publication of volume 1.0. Therefore, this is not a self-contained document, but is rather an addendum to the two previously published guidelines for the Penn Korean Treebank (Han and Han, 2001; Han et al., 2001).

The Penn Korean Treebank 2.0 corpus is currently in negotiation to be released by the Linguistic Data Consortium. In this release, a new edition of the Korean Treebank corpus will be included alongside the new volume of Korean Treebank 2.0, dubbed Korean Treebank Annotations Version 1.1, which has been edited to conform to the newly revised guidelines illustrated in this document.

[1] We are extremely grateful to Martha Palmer for her continued support and guidance in this project. We also thank Beatrice Santorini for insightful discussions in setting some of the guidelines. We also would like to thank Seung-yun Yang, Sook-Hee Chae and Seunghun Lee who participated in the earlier stage of the project, as well as Kyuchul Yoon, Hyunsook Shin, and Eunjong Kong from Ohio State University who lent us valuable help with the part of the annotation process. The work reported in the document was supported by contract DAAD 19-03-2-0028, awarded by the Army Research Lab.

[2] The sentence ID field of the Korean Treebank 2.0 matches the file name and the document ID (`<DOCID>` field) of the Korean Newswire Corpus. For example, the Korean Treebank sentence `3200090:3` found in file `320009.fid` corresponds to the third sentence of the Korean Newswire document `KPA20000320.0090` found in file `20000320.SGM`. Note that the sentence number field (`:3` of `3200090:3`) increments throughout a Korean Treebank file, which consists of multiple articles, and is not reset between articles.

2 Tokenization in Korean Treebank 2.0

2.1 Tokenization Marker '~' Introduced

In the Korean Treebank, raw sentences appearing above each bracketed tree have been already tokenized. Periods, commas, quotation marks and other symbols appear separated out in the sentence field. In Korean Treebank 1.0, sentence fields contain such tokenized words, which look like the following:

(1) `;;B;01:23: 4 중 대는 " 정산 15 " 이고 6 중 대는 " 정산 17 " 입니다 .`

In Korean Treebank 2.0, tokenization marker ~ is introduced, which is prefixed onto a token to indicate that it had been separated from the preceding element during the tokenization process. That is, "AB" in the original text is tokenized to "A ~B". Introduction of this tokenization marker ensures that the original sentence before the tokenization process is easily recoverable. Without such marking, the quotation markers in the above sentence from Treebank 1.0 cannot be locally determined whether they were originally attached to the preceding word or the following word, or both. Example sentences in Korean Treebank 2.0 with the tokenization markers are shown below:

(2) `;;3200011:1: 한국 전수일 감독의 ' ~허공에 멈추는 새 ~' ~가 19 일 스위스 프리부르 국제영화제에서 대상을 받았다 ~.`
`;;3020013:10: 국제통화기금 ~( ~IMF ~) ~은 ...`

2.2 Conservative Tokenization Strategies

In Korean Treebank 1.0, tokenization was applied generously in favor of simpler morpho-syntax and transparent syntactic structures. For example, only affixation was allowed in word-formation, and as a result word-phrases that do not conform to the limited set of morpho-syntactic rules were routinely separated apart into a sequence of words. For example, noun compounds such as 호출/NNC+대호/NNC were disallowed and were instead represented as two separate nouns as in 호출/NNC 대호/NNC; the sentence field also reflected this tokenization. In the Korean Treebank 2.0, such liberal use of tokenization was recognized as an undesirable practice that introduces distortion into naturally occurring texts. Hence, more conservative tokenization strategies were employed, where only symbols are subject to tokenization in principle. As a result, original "띄어쓰기 (word-spacing)" of the text is preserved for the most part. Forced tokenization (where no symbols are involved) is used sparingly only for those cases where insertion of a space would result in a grammatical word-spacing alternative. Moreover, location of forced tokenization is marked with the tokenization marker '~'. There are two occasions where such forced tokenization is necessary: (a) when the original spacing is clearly an error or (b) when syntactic annotation requires token-separation within a word-phrase. Following are examples of word-spacing errors:

(3) forced tokenization on word-spacing errors

밥을먹었다 -> 밥을 ~먹었다
많은곳을 봤다 -> 많은 ~곳을 봤다

The school grammar of Korean prescribes separation of a verb and the following auxiliary verb, which is often disregarded in practice. In Korean Treebank, auxiliary verbs take the VP headed by the matrix verb to project up to another VP. This syntactic configuration is impossible when `verb+auxiliary-verb` string is not separated. Hence:

(4) forced tokenization on `verb+auxiliary-verb` construction

```
임무를 맡아온 -> 임무를 맡아 ~온
(VP (VP (NP-OBJ 임무/NNC+을/PCA)
        맡/VV+어)
    오/VX+은/EAN)
```

Noun-verb sequences that form a close semantic unit are often found written as one word, a tendency resulting from the process of incorporation. In the Korean Treebank, the noun must be identified and labeled as an argument of the verb, which is made possible by separating it out from the verb:

(5) forced tokenization on `noun+predicate` construction

```
신경쓰고  -> 신경 ~쓰고
(VP (NP-OBJ 신경/NNC)
    쓰/VV+고/ECS))

강도높게  ->  강도 ~높게
(S (NP-SBJ 강도/NNC)
   (ADJP 높/VJ+게/ECS))

의미있는  -> 의미 ~있는
(S (NP-SBJ 의미/NNC)
   (ADJP 있/VJ+는/EAN))
```

Starting from Korean Treebank 2.0, '중이다' is recognized as an auxiliary verb (see Section 7.1), which frequently appears attached to a Cino-Korean verbal noun. In order to give it a separate node as a `VX`, it is separated out from the noun:

(6) forced tokenization on 중/NNX+이/CO:

```
... 뉴욕에서 유세 ~중이던 고어 부통령은 ...
(NP-SBJ (S (WHNP-1 *op*)
           (S (NP-SBJ *T*-1)
              (VP (VP (NP-ADV 뉴욕/NPR+에서/PAD)
                      (VP (VV 유세/NNC)))
                  (VX 중/NNX+이/CO+던/EAN))))
        (NP 고어/NPR
            부통령/NNC+은/PAU))
```

3 Revised Part-of-Speech Tagging Guidelines

3.1 Treatment of Allomorphy

A large number of inflectional suffixes and post-position markers in Korean have allomorphs, whose distributions are conditioned by the phonological environment in which they appear. For example, the "topic" postposition marker takes three different forms '은', '는', and 'ㄴ'; the past-tense pre-verbal-ending suffix '었', '았', and 'ㅆ'.

The position taken in Korean Treebank 1.0 was not to posit a single lexical representation for such sets of allomorphs, opting instead to list appropriate allomorphic forms within context. In Korean Treebank 2.0, however, allomorphs are treated as having a single representative form. All allomorphs of a given lexical item therefore show up as a single form in the morphologically analyzed string. For example, the topic markers in '학교-는', '학생-은' and '너-ㄴ' are equally assigned 은/PAU:

(7) '은' as the representative form for the Korean topic postposition:

학교/NNC+은/PAU	학생/NNC+은/PAU	너/NPN+은/PAU
↕	↕	↕
학교는	학생은	넌

The criteria used in determining the representative form among allomorphs are as follows:

(8) Criteria for determining the representative form

a. The representative form should be fully syllabic, i.e. '은' is chosen over 'ㄴ'.

b. The form for the post-consonantal environment is chosen, i.e. '이' instead of '가'.

c. Epenthetic vowels are included, i.e. '으로' and not '로'. [3]

d. For vowel harmony, '어' is chosen over '아', i.e. '어서' and not '아서'.

The following is a list of common allomorphic morphemes and their representative forms:

(9) Common allomorphs and their representative forms

allomorph	usage	representative form
은/는/ㄴ	학생은/고수는/넌	은
이/가	학생이/고수가	이
을/를	학생을/고수를	을
과/와	학생과/고수와	과
으로/로	학생으로/고수로	으로
이라도/라도	학생이라도/고수라도	이라도
이야/야	학생이야/고수야	이야
아/야	복순아/영희야	아
었/았/ㅆ	먹었고/잡았고/샀고	었
어/아/null	먹어/잡아/사	어
어서/아서/서	먹어서/잡아서/사서	어서

[3] This clause is in fact redundant, as epenthetic vowels are used in post-consonantal environments only which is covered by criterion (b).

은/ㄴ	먹은/산	은
음/ㅁ	먹음/삼	음
으시/시	잡으시고/오시고	으시
으니/니	오니/잡으니	으니
을까/ㄹ까	먹을까/살까	을까
습니다/ㅂ니다	먹습니다/삽니다	습니다
이/null	학생이다/고수다	이

3.2 EAU Merged with ECS

In Korean Treebank 1.0, auxiliary endings are recognized as a separate part-of-speech category and are given the tag EAU. There were four verbal endings which belonged to the category: '아', '게', '지', '고'. In Korean Treebank 2.0, they are merged with the category ECS, i.e., 'Coordinate, Subordinate, Adverbial, Complementizer Ending', which is the more general category for all non-sentence-final verbal ending suffixes. With the exception of '지', the EAU verbal endings are permitted in non-auxiliary environments, which resulted in ambiguity in tagging depending on the environment (see Section 4.12 of the POS Tagging Guidelines):

(10) '어' as EAU or ECS in Korean Treebank 1.0:

```
먹어 보았다:      먹/VV+어/EAU 보/VX+았/EPF+다/EFN
갈아 넣었다:      갈/VV+아/ECS 넣/VV+었/EPF+다/EFN
```

As a result of the merge, such ambiguity in POS tags no longer exists, and they are tagged as ECS in all environments.

(11) '어' is always ECS in Korean Treebank 2.0:

```
먹어 보았다:      먹/VV+어/ECS 보/VX+었/EPF+다/EFN
갈아 넣었다:      갈/VV+어/ECS 넣/VV+었/EPF+다/EFN
```

3.3 PAN (Adnominal Postposition) Introduced

A new part-of-speech tag PAN, 'Adnominal Postposition', is created in Korean Treebank 2.0. The PANs are essentially post-position markers, but share with other adnominal POSs such as DAN (Adnominal Determiner) and EAN (Adnominal Ending) the property of modifying the noun elements that follow them. There are two morphemes that are PAN: '의' and '이라는':

(12) 의/PAN and 이라는/PAN

```
철수의 성적:        철수/NPR+의/PAN 성적/NNC
철수라는 학생:      철수/NPR+이라는/PAN 학생/NNC
```

In Korean Treebank 1.0, '의' was classified as PCA, a Case Postposition. While it is true that the postposition marker encodes the Genitive **Case**, its syntactic property of modifying nouns is vastly different from other postposition markers which typically encode the relation between the predicate and the root noun.

Also, '이라는' was treated as a complex morphemic unit 이/CO+라는/EAN, which is made up of a copula followed by an adnominal verbal ending suffix. This inevitably lead to a syntactic analysis involving a full-blown relative clause for the constructions with the expression:

(13) '철수라는 학생' involves a relative clause in Treebank 1.0

```
(NP (S (WHNP-1 *op*)
       (S (NP-SBJ *T*-1)
          (VP (NP 철수/NPR+이/CO+라는/EAN))))
    (NP 학생/NNC))
```

In Korean Treebank 2.0, they are assigned a much simpler syntactic structure of a noun modifying a noun:[4]

(14) '철수라는 학생' is a simple noun phrase in Treebank 2.0

```
(NP (NP 철수/NPR+이라는/PAN)
    (NP 학생/NNC))
```

'이라는', however, retains its complex morphologicial structure and therefore induces a clausal syntactic structure in non-appositive environments:

(15) '그 학생이 철수라는 사실' in Treebank 1.0 and 2.0

```
(NP (S (NP-SBJ 그/DAN
               학생/NNC+이/PCA)
       (VP (NP 철수/NPR+이/CO+라는/EAN)))
    (NP 사실/NNC))
```

Such type of ambiguity can also be found in other items including '이라고', which is similarly ambiguous between 이라고/PAD and 이/CO+라/EFN+고/PAD[5]. When a noun phrase exists which functions as the subject of the 'Noun+이라고' unit, '이라고' is treated as a copula followed by a verbal ending; otherwise, the entire expression is tagged as a postposition marker:

(16) parallel ambiguity in '이라고':

```
나는 (VP (NP-OBJ 철수를) (NP-COMP 형/NNC+이라고/PAD) 부른다)
나는 (VP (S-COMP (NP-SBJ 철수가) (VP (NP 바보/NNC+이/CO+라/EFN+고/PAD))) 생각한다)
```

3.4 Treatment of XSV and XSJ

In Korean Treebank 1.0, four verbalization suffixes (XSV) were recognized: '하', '되', '시키', '받'. Three additional XSVs are introduced in Korean Treebank 2.0: '어지', '어하', '당하'. '어지' attaches to verbal and adjectival roots; '어하' mostly attaches to adjectival roots to turn them into a verb[6]; '당하' attaches to nominal roots:

(17) '-어지/어하' as verbalization suffix

[4]It should be distinguished from another lexical item '이란':

```
책/NNC+이란/PAU 우리/NPN+의/PAN 마음을 살찌우는 영혼의 양식이다.
```

Here, it is the phonological contraction of '이라는 것은', and is tagged PAU .

[5]이/CO+라고/ECS in Korean Treebank 1.0.

[6]Only one exception is found in KTB 2.0 where it attaches to a verbal root: 꺼리/VV+어하/XSV+다/EFN.

예뻐지다: 예쁘/VJ+어지/XSV+다/EFN
알려지다: 알리/VV+어지/XSV+다/EFN
어려워하다: 어렵/VJ+어하/XSV+다/EFN
공격당하다: 공격/NNC+당하/XSV+다/EFN

Note that '어지', when attached to an adjective root, was treated in Korean Treebank 1.0 as a part of the root:

(18) a. '예뻐지다' in Korean Treebank 1.0:

예뻐지/VJ+다/EFN

b. '예뻐지다' in Korean Treebank 2.0:

예쁘/VJ+어지/XSV+다/EFN

As in Korean Treebank 1.0, a word with a derivational suffix projects up to an appropriate part-of-speech node. Therefore, the word in the following example with two derivational suffixes XSJ and XSV projects up to VJ first and then to VV:

(19) syntactic projection of derived words

곤란해했다: (VP (VV (VJ 곤란/NNC+하/XSJ+어하/XSV+었/EPF+다/EFN)))

Even when 어지/XSV is attached to a VV root, which is already a verb on its own, the word projects up to a VV node as seen below. This is due to the fact that the derived verb 밝히/VV+어지/XSV is considered a new verb which has different subcategorization properties from those of the original root 밝히/VV.

(20) VV+XSV projects to VV

밝혀졌다: (VP (VV 밝히/VV+어지/XSV+었/EPF+다/EFN))

Note, however, that the '-어 지-' construction is treated as a verbal ending and an auxiliary verb ('-어/ECS 지/VX') when they are written separated out:

(21) '-어 지-' tagged as -어/ECS 지/VX

먹히어 지다: 먹히/VV+어/ECS 지/VX+다/EFN

Likewise, '당하' is tagged as XSV when it follows a common noun and turns it into verb. Note that '당하' can also function as a main verb, if it is separated from a noun by a space:

(22) 공격당하다: 공격/NNC+당하/XSV+다/EFN
거절당하다: 거절/NNC+당하/XSV+다/EFN
그 범인은 처벌/NNC+당하/XSV+었/EPF+다/EFN

공격 당하다: 공격/NNC 당하/VV+다/EFN
거절을 당했다: 거절/NNC+을/PCA 당하/VV+었/EPF+다/EFN

Also, use of XSV/XSJ suffixes is limited to those cases where the stem indeed is a common noun after separating out the suffixes. Therefore, '조사' in '조사하다' receives tag NNC while '끔찍' in '끔찍하다' does not:

(23) 조사하다: 조사/NNC+하/XSV+다/EFN
수영하다: 수영/NNC+하/XSV+다/EFN
필요하다: 필요/NNC+하/XSJ+다/EFN

반짝하다: 반짝하/VV+다/EFN
출렁거리다: 출렁거리/VV+다/EFN
훌륭하다: 훌륭하/VJ+다/EFN
불구하다: 불구하/VJ+다/EFN
끔찍하다: 끔찍하/VJ+다/EFN
빽빽하다: 빽빽하/VJ+다/EFN

3.5 Treatment of NPR (Proper Noun)

NPR is a morpheme-level tag that represents "proper noun", which is the kind of noun that refers to "names". NPRs in the most obvious cases are illustrated in the examples below, where a single morpheme constitutes a name:

(24) simple cases of NPR:
김대중/NPR
러시아/NPR
혼다/NPR
홍콩/NPR+과/PCJ

Given the apparent connection between the NPR tag and "names", it is easy to get into the mind-set: "This is a name referring to a single entity, therefore the entire thing should receive NPR." Under this extreme approach, long names which in themselves contain multiple morphemes are treated as one single NPR:

(25) "one NPR tag per name" approach:

전국인민대표대회/NPR
한국전력/NPR+의/PAN
대만해협/NPR
유엔안전보장이사회/NPR
한국농촌경제연구원/NPR+의/PAN
중소기업협동조합중앙회/NPR
석유수출국기구/NPR

However, it soon becomes clear on closer inspection that it is impossible to fully represent "names" by NPR tag, because NPR is defined strictly on the morpheme level and names are not. NPR can only apply to morphemic units while "name"s can be represented by larger units, namely word (단어), word-phrase (어절), and phrase (구):

(26) multi-word (i.e. phrasal) names:

한글과 컴퓨터
이라크 항공
워싱턴 타임즈
유엔 안전보장 이사회

This shows that the "name-NPR" equation does not hold beyond the simplest cases presented above in (24). NPR therefore cannot be viewed as a tag that bears any systemic relation to "names". "Name" is in fact a semantic concept that is best annotated on a separate level, such as named entity annotation. From this point of view, the long names in (27) must be broken down into their component morphemes as seen below, which is the position adopted in Korean Treebank 2.0.

(27) names are broken down to component morphemes:

전 국/NNC+인 민/NNC+대표/NNC+대회/NNC
한 국/NPR+전 력/NNC+의/PAN
대만/NPR+해협/NNC
유 엔/NPR+안전/NNC+보 장/NNC+이 사회/NNC
석 유/NNC+수 출 국/NNC+기 구/NNC

3.6 Man-Ha/VX → (VX Man/NNX+Ha/XSJ)

In Korean Treebank 1.0, '만하-' was treated as an auxiliary verb. In Korean Treebank 2.0, however, it is decomposed into a dependent noun (NNX) followed by an adjectivization suffix (XSJ):

(28) a. treatment of '만하-' in Treebank 1.0
쓰/VV+르/EAN 만 하/VX+다/EFN

b. treatment of '만하-' in Treebank 2.0
쓰/VV+르/EAN 만/NNX+하/XSJ+다/EFN

This decision was made in order to ensure consistency with variations of the construction such as below, where a post-position marker intervenes between '만' and '하'. Analyzing '만' as a dependent noun is inevitable in such cases.

(29) 보/VV+을/EAN 만/NNX+도/PAU 하/VJ+다/EFN

3.7 'Pu-Teo', 'Kka-Ji' Invariably PAU

In Korean Treebank 1.0, '부터' and '까지' were treated as ambiguous between PAD (adverbial postpositoin marker) and PAU (auxiliary postposition marker) tags. Specifically, they were tagged PAD when they convey the sense of geographical origin and destination respectively, and PAU in other cases. Starting from Korean Treebank 2.0, the noun phrases that they attach to are no longer considered an argument of a verb but rather an adjunct. It follows from this that the postposition markers are no longer tagged PAD, which generally indicates the argument-status of the head noun; they are invariably treated as PAU.

(30) '부터' and '까지' are invariably tagged as PAU:
집/NNC+부 터/PAU 학 교/NNC+까 지/PAU

3.8 Allow both ADV/NPN for Some Pronouns

In Korean Treebank 1.0, '언제' was always tagged as NPN, so it has the same POS tag as other WH-items such as 어 디/NPN and 누 구/NPN. Starting from Korean Treebank 2.0, '언제' is tagged ADV when it is used adverbially, and NPN when it is used nominally:

(31) '언제' is either ADV or NPN:

'언제 왔니?'	언제/ADV 오/VV+었/EPF+니/EFN ?/SFN
'언제가 좋으니?'	언제/NPN+이/PCA 좋/VJ+으니/EFN ?/SFN

3.9 More on Dependent Noun(NNX) Examples

The following '초' and '말' are tagged as NNX:

(32)
```
20세기 초/NNX
지난달/NNC 초/NNX+부터/PAU 2월 초/NNX+까지/PAU

20세기 말/NNX
이/DAN 달/NNC 말/NNX+까지/PAU 끝내자.
```

The following '내'(inside) and '외'(outside) are tagged as NNX (cf. Tagging Guidelines p.6):

(33)
```
이 구역/NNC 내/NNX+에/PAD 들어오지 마시오
시험이 예상/NNC 외/NNX+로/PAD 까다롭다
```

3.10 Treatment of 'To, Ku, Si, Kun, Myeon, eup, Ri, Tong'

'도, 구, 시, 군, 면, 읍, 리, 동' should be always tagged as NNC, not XSF (cf. Tagging Guidelines p.21).

(34)
```
서울시:  서울/NPR+시/NNC
경기도:  경기/NPR+도/NNC
```

When '도' has the meaning of 'island', words occurring with this morpheme should be tagged as NPR as a whole.

(35)
```
제주도/NPR
울릉도/NPR
```

3.11 Treatment of 'Ko, Ra-Ko'

Complementizer postpositions '고' occurs on the predicate of a complement clause under verbs such as '말하다, 생각하다, 믿다, 요구하다'. It should be tagged as PAD (cf. Tagging Guidelines p.3).

(36)
```
그는 집에 있었다고 말했다:
    (NP-SBJ 그는) (VP (S-COMP 집에 있었다+고/PAD) 말했다)
```

'(이)라고' that follows a direct quotative complement clause or a simple noun should be tagged as PAD as a whole (cf. Tagging Guidelines p.15, p.33-4).

(37)
```
"언제 오겠니"라고 물었다:
    (VP (S-COMP ''언제 오겠니"+라고/PAD) 물었다)

완다라고 부른다:
```

```
    완다/NPR+라고/PAD 부른다

존슨이라고 부른다:
    존슨/NPR+이라고/PAD 부른다
```

'(이)라고' that is used complementizer of the verbs '말하다, 생각하다, 믿다, 요구하다' should be separated and tagged as 라/EFN+고/PAD. In this case, a copula must be recovered before '라', if necessary.

(38)
```
나는 철수가 완다라고 믿는다:
    나는 철수가 완다/NPR+이/CO+라/EFN+고/PAD 믿는다
```

'라고', '(으)라고' that follow an adverbial clause should be tagged as ECS as a whole. Examples are:

(39)
```
몸이 정상이 아니라고 슬퍼하지 마라:
    아니/VJ+라고/ECS 슬퍼하지

어머니께서 생활비에 보태 쓰라고 돈을 보내 주셨다:
    보태 쓰/VV+라고/ECS 돈을 ...

책을 읽으라고 시켰다:
    읽/VV+으라고/ECS
```

3.12 Treatment of Surface Form 'Ta-Ko'

'다고' that is used complementizer of the verbs '말하다, 생각하다, 믿다, 요구하다' should be separated and tagged as 다/EFN+고/PAD (cf. Tagging Guidelines p.18, p.33-4).

(40)
```
그 곳에 전화기가 있다고 생각한다:  있/VV+다/EFN+고/PAD
그는 "편지를 쓴다"고 말했다:       쓰/VV+는다/EFN+고/PAD
```

'다고' that follows an adverbial clause should be tagged as ECS as a whole. Examples are:

(41)
```
날 어리다고 얕보지 마세요:
    어리/VJ+다고/ECS

일류 대학을 가겠다고 열심히 공부하고 있다:
    가/VV+겠/EPF+다고/ECS

허락 없이 제 물건을 썼다고 야단야단이다:
    쓰/VV+었/EPF+다고/ECS
```

3.13 Treatment of 'iss-Ta' and 'Kye-Si-Ta'

If '있다, 계시다' are used as auxiliary predicates, then they should be tagged as VX.

(42)
```
먹고 있다:          먹/VV+고/ECS 있/VX+다/EFN
잡수시고 계신다:    잡수시/VV+고/ECS 계시/VX+는다/EFN
```

3.14 Treatment of 'Teul'

'들' usually attaches to a singular noun and turns it into a plural noun. In that case, it is tagged as XSF (cf. Tagging Guidelines p.20, p.26).

(43) 여기에는 차들이 많다: 차/NNC+들/XSF+이/PCA
저 사람들을 보아라: 사람/NNC+들/XSF+을/PCA
학생들에게만 나누어 주었다: 학생/NNC+들/XSF+에게만/PAD

But '들' can occur on words other than nouns. In some cases, '들', even when attached to a noun, does not convey the plural sense. In these environments, '들' is tagged as PAU.

(44) 빨리들 먹어라: 빨리/ADV+들/PAU
저기 가고들 있구먼: 가/VV+고/ECS+들/PAU
모두 자리에 앉게들: 앉/VV+게/EFN+들/PAU
말씀들 나누세요: 말씀/NNC+들/PAU
어서 밥들 먹어라: 밥/NNC+들/PAU

Note that it can also be tagged as NNX in the particular context shown below:

(45) 소, 개, 닭 들:
소/NNC ,/SCM 개/NNC ,/SCM 닭/NNC 들/NNX

배, 감, 포도 들이 많다:
배/NNC ,/SCM 감/NNC ,/SCM 포도/NNC 들/NNX+이/PCA

4 Case Studies

4.1 VV or VX

Some verbs are ambiguous between VV and VX, when they follows a predicate with a '어' ending. One way of distinguishing between the two cases is to replace '어' with '어서': if it is grammatical and preserves the overall meaning, then the second verb is tagged as VV, otherwise, it is tagged as VX.

(46) 그림을 그려 주었다: 그림/NNC+을/PCA 그리/VV+어/ECS 주/VV+었/EPF+다/EFN
책을 읽어 주었다: 책/NNC+을/PCA 읽/VV+어/ECS 주/VX+었/EPF+다/EFN

4.2 VV or VJ

Some predicates such as '크다, 붉다, 밝다, 설다, 맞다, 늦다, 굳다' are ambiguous between VV and VJ. One way of distinguishing between the two cases is to add present tense marker '는' to the predicates: if it is possible, then the predicate is tagged as VV, otherwise, it is tagged as VJ.

(47) 철수는 마음이 크다: 크/VJ+다/EFN
아이들이 큰다: 크/VV+는다/EFN

달이 매우 밝다: 밝/VJ+다/EFN
날이 밝는다: 밝/VV+는다/EFN

굳고 단단한 물건: 굳/VJ+고/ECS
빵이 돌처럼 굳는다: 굳/VV+는다/EFN

4.3 PAD or NNX

If '대로', '만큼' immediately follow a noun without a space, they are postposition markers and therefore are tagged as PAD.

(48) 명령대로: 명령/NNC+대로/PAD
나도 그 사람만큼 뛸 수 있다: 사람/NNC+만큼/PAD

If '대로', '만큼' are modified by a relative clause, they are dependent nouns (NNX).

(49) 말한 대로: 말한 대로/NNX
나도 참을 만큼 참았다: 참을 만큼/NNX

5 Confusing Examples

- '라'
  '라' is tagged as EFN if it indicates the sentence is a imperative. But '라' is tagged as ECS if it follows copula '이' or adjective '아니', and conjoins two sentences.

  너 자신을 알/VV+라/EFN
  깊이 파/VV+라/EFN.
  뜻밖의 일/NNC+이/CO+라/ECS 어리둥절했다.
  기대했던 대로 가 아니/VJ+라/ECS 크게 실망했다.
  사람이 아니/VJ+라/ECS 짐승이다.

- '말라'
  In most cases, '말' in '말다' is an auxiliary verb and is tagged as VX. But if it is used with the form '말고' and immediately follows a noun, then it is tagged as PAU as a whole.

  가지 마라: 가/VV+지/EAU 말/VX+라/EFN
  너말고 네가 가라: 너/NPN+말고/PAU

- '모르다'
  In general, '모르다' is tagged as VV. But when it follows a predicate with ECS '을지, 는지, 은지' and means 'might', it is tagged as VX:

  나는 그 사람을 모른다: 그/DAN 사람/NNC+을/PCA 모르/VV+는다/EFN
  손이가 노래할지도 모른다: 노래/NNC+하/XSF+을지도/ECS 모르/VX+는다/EFN
  그 해는 어찌나 추웠는지 모른다: 춥/VJ+었/EPF+는지/ECS 모르/VX+는다/EFN

- '요'
  '요' is tagged as PAU if it attaches to a noun, but it is tagged as EFN if it attaches to a verb. '요' is tagged as ECS if it follows copula '이' or adjective '아니', and functions as a coordinate ending:

나는요 : 나/NPN+는/PAU+요/PAU
갔어요 : 가/VV+었/EPF+어요/EFN
이것은 감이요 저것은 사과다 : 감/NNC+이/CO+요/ECS

- '중'
'중' is tagged as NNC if it has the meaning of 'monk', 'middle', or 'among'. but it is tagged as NNX if it follows the deverbal nouns and has the meaning of 'during' of 'throughout'.

그/NPN+는/PAU 중/NNC+이/PCA 되/VV+었/EPF+다/EFN.
이/DAN 중/NNC+에서/PAD
회의/NNC 중/NNX, 방문/NNC 중/NNX, 오전/NNC 중/NNX

- '이니'
In most cases, '이니' is a combination of a copula and a sentence final ending marker (이/CO+니/EFN). But when it attaches to nouns in a list context, it should be tagged as PCJ as a whole.

이것이 책/NNC+이/CO+니/EFN?
해방/NNC+이니/PCJ, 통일/NNC+이니/PCJ, 그게 무슨 말/NNC+이/CO+니/EFN?

- '한편'
'한편' is tagged as NNC when it precedes postpositions or it has modifiers. Otherwise, '한편' is tagged as ADC.

말과 당나귀는 한편/NNC+으로/PAD 비슷한 점이 있으면서,
다른 한편/NNC+으로/PAD 구별되는 점이 있다.

부실 금고의 정리를 조속히 추진/NNC+하/XSV+는/EAN 한편/NNC
우량금고의 모범사례는 발굴해 전파할 예정이다.

한편/ADC 런던/NPR+회의/NNC+이/PCA 끝난 후 유가는 다시 올랐다.

6 Revised Bracketing Guidelines

6.1 'e-Seo' as Subject Case Marker

In Korean Treebank 2.0, '에서' following a noun which refers to an organization or socio-political entity is recognized as a subject case marker. It is tagged as PCA accordingly, and the whole noun phrase is categorized as NP-SBJ:

(50)
```
정부에서 새로운 법령을 공포하였다:
(S (NP-SBJ 정부/NNC+에서/PCA)
   (VP (NP-OBJ 새로운 법령을)
       (VV 공포하였다)))

우리 학교에서 응원상을 받았다:
(S (NP-SBJ 우리 학교/NNC+에서/PCA)
   (VP (NP-OBJ 응원상을)
       받았다))
```

6.2 Ha/VX→ Ha/VV in '-To-Rok/Ki-Ro Ha-' Constructions

In Korean Treebank 1.0, the '-기로 하-' construction was analyzed in such a way that the verb accompanied by '-기로' was treated as the main verb while '하-' was relegated to the role of an auxiliary verb:

(51) treatment of '-기로 하-' in Korean Treebank 1.0:

```
(S (NP-SBJ 철수/NPR+가/PCA)
   (VP (VP 가/VV+기로/ECS)
       하/VX+었/EPF+다/EFN)
   ./SFN)
```

In Korean Treebank 2.0, the annotation guideline is revised so '하-' is now recognized as the main verb of the construction which takes a clausal argument headed by 기/EAN+으로/PAD:

(52) treatment of '-기로 하-' in Korean Treebank 2.0:

```
(S (NP-SBJ *pro*)
   (VP (S-COMP (NP-SBJ 철수/NPR+이/PCA)
               (VP 가/VV+기/ENM+으로/PAD))
       하/VV+었/EPF+다/EFN)
   ./SFN)
```

This decision was due to the observation that the verb '하-' has some degree of agentivity of its own therefore cannot be an auxiliary verb. As a piece of supporting evidence, the subject of '하-' in the example above can be different from '철수', as indicated by the empty pronoun occupying the position: the person who is going is '철수', but the decision may well have been made by some other person or persons.

I much the same way, '-도록 하-' construction in Korean Treebank 1.0 was analyzed in such a way that the verb accompanied by '-도록' was treated as the main verb while '하-' was relegated to the role of an auxiliary verb:

(53) treatment of '-도록 하-' in Korean Treebank 1.0:

```
(S (NP-SBJ 철수/NPR+가/PCA)
   (VP (VP 가/VV+도록/ECS)
       하/VX+었/EPF+다/EFN)
   ./SFN)
```

For reasons analogous to the ones given for the case of '-기로 하-', the treatment for '-도록 하-' is revised as follows:

(54) treatment of '-도록 하-' in Korean Treebank 2.0:

```
(S (NP-SBJ *pro*)
   (VP (S-COMP (NP-SBJ 철수/NPR+이/PCA)
               (VP 가/VV+도록/ECS))
       하/VV+었/EPF+다/EFN)
   ./SFN)
```

Again, as a result of the revision, the subject of the verb '하-' is now allowed to be different from the subject of the embedded clause, which makes possible the more favorable semantic interpretation where some person other than 철수 is the agent of the decision of 철수's going.

6.3 Extension of S-level Tag

When a sentential element is ended with the suffixation of '음' or '기', it is bracketed as S with an appropriate function tag, i.e., S-SBJ, S-OBJ, or S-COMP. We do not further project it to an NP (cf. Bracketing Guidelines p.7-9).

(55) 확력 지원은 보통 철수를 엄호하기 위해서 혹은 후방으로의 대이동을 은폐하기 위해서 가능한 한 최대로 합니다.

```
(S (NP-SBJ *pro*)
   (VP (NP-OBJ 확력/NNC
               지원/NNC+은/PAU)
       (VP (ADVP 보통/ADV)
           (VP (S (S (NP-SBJ *pro*)
                     (VP (S-OBJ (NP-SBJ *pro*)
                                (VP (NP-OBJ 철수/NNC+을/PCA)
                                    (VV 엄호/NNC+하/XSV+기/ENM)))
                         위하/VV+어서/ECS))
                  (ADCP 혹은/ADC)
                  (S (NP-SBJ *pro*)
                     (VP (S-OBJ (NP-SBJ *pro*)
                                (VP (NP-OBJ (NP 후방/NNC+으로/PAD+의/PAN)
                                            (NP 대/XPF+이동/NNC+을/PCA))
                                    (VV 은폐/NNC+하/XSV+기/ENM)))
                         위하/VV+어서/ECS)))
               (VP (NP-ADV (S (NP-SBJ *pro*)
                              (ADJP (VJ 가능/NNC+하/XSJ+은/EAN)))
```

```
                        (NP 한/NNX))
                    (NP-ADV 최대/NNC+으로/PAD)
                    (VP 하/VV+읍니다/EFN)))))
    ./SFN)
```

When a sentence is followed by an final ending (EFN) such as '은지' and '느냐', it can also be a sentential subject, sentential object or sentential complement. It is bracketed as S with appropriate function tag, i.e., S-SBJ, S-OBJ, or S-COMP.

(56) 그들의 군사 칭호가 뭔지는 제가 잘 모르겠습니다.

```
(S (S-OBJ-1 (NP-SBJ (NP 그/NPN+들/XSF+의/PAN)
                    (NP 군사/NNC
                        칭호/NNC+이/PCA))
            (VP (NP 무엇/NPN+이/CO+는지/EFN+은/PAU)))
   (S (NP-SBJ 제/NPN+이/PCA)
      (VP (S-OBJ *T*-1)
          (VP (ADVP 잘/ADV)
              (VP 모르/VV+겠/EPF+읍니다/EFN))))
   ./SFN)
```

(57) 어떤 종류의 기준선이 사용되느냐에 따라 방위에는 세 가지 종류가 있소.

```
(S (S (NP-SBJ *pro*)
      (VP (S-COMP (NP-SBJ (NP 어떤/DAN
                              종류/NNC+의/PAN)
                          (NP 기준선/NNC+이/PCA))
                  (VP (VV 사용/NNC+되/XSV+느냐/EFN+에/PAD)))
          따르/VV+어/ECS))
   (S (NP-ADV 방위/NNC+에/PAD+은/PAU)
      (S (NP-SBJ 세/NNU
                 가지/NNX
                 종류/NNC+이/PCA)
         (ADJP 있/VV+소/EFN)))
   ./SFN)
```

In all other cases where S is an argument, S itself is simply treated as a complement of a verb, i.e., S-COMP. This includes, but is not limited to, the cases where a sentence is followed by EFN+PAD such as '다고' and '라고' as well as an inflectional ending '도록', etc.:

(58) 그는 무전기가 고장났다고 말했다.

```
(S (NP-SBJ 그/NPN+은/PAU)
   (VP (S-COMP (NP-SBJ 무전기/NNC+이/PCA)
               (VP 고장나/VV+었/EPF+다/EFN+고/PAD))
       말하/VV+었/EPF+다/EFN)
   ./SFN)
```

중대 특무장은 중대 성원들이 전투에 필요한 무기와 탄약을 틀림없이 갖도록 합니다.

```
(S (NP-SBJ 중대/NNC 특무장/NNC+은/PAU)
   (VP (S-COMP (NP-SBJ 중대/NNC 성원/NNC+들/XSF+이/PCA)
               (VP (NP-OBJ (S (WHNP-1 *op*)
                              (S (NP-SBJ *T*-1)
                                 (ADJP (NP-COMP 전투/NNC+에/PAD)
                                       (VJ 필요/NNC+하/XSJ+은/EAN))))
                           (NP 무기/NNC+와/PCJ 탄약/NNC+을/PCA))
                   (VP (ADVP 틀림없이/ADV)
                       (VP 갖/VV+도록/ECS))))
       하/VV+읍니다/EFN)
   ./SFN)
```

6.4 Extension of Complex Auxiliary Predicate

'수 있다' and '수 없다' occur at the end of sentences and correspond in meaning to English auxiliary predicates such as 'can' and 'cannot'. Label '수 있다' and '수 없다' as VX and treat them as auxiliary predicates.

(59) 한국탁구가 2000 년 시드니올림픽 본선에 남녀복식 2 개조씩을 파견할 수 있게 됐다.

```
(S (NP-SBJ 한국/NPR+탁구/NNC+이/PCA)
   (VP (VP (VP (NP-COMP 2000/NNU
                        년/NNX
                        시드니/NPR+올림픽/NNC
                        본선/NNC+에/PAD)
               (NP-OBJ 남녀/NNC+복식/NNC
                       2/NNU
                       개/NNX+조/NNC+씩/XSF+을/PCA)
               (VV 파견/NNC+하/XSV+을/EAN))
           (VX 수/NNX
               있/VV+게/ECS))
       되/VX+었/EPF+다/EFN)
   ./SFN)
```

가르켜 드릴 수가 없습니다.

```
(S (NP-SBJ *pro*)
   (VP (VP (VP (NP-OBJ *pro*)
               가르키/VV+어/ECS)
           드리/VX+을/EAN)
       (VX 수/NNX+이/PCA
           없/VJ+읍니다/EFN))
   ./SFN)
```

7 New Issues in Bracketing Guidelines

7.1 'Jung-i-Ta' as VX

In Korean Treebank 2.0, '중이다' is recognized as an auxiliary verb (also noted previously in Section 2.2). The decision was made in order to reflect the auxiliary-verb-like nature of '중이다', which tends to act like an auxiliary verb encoding an aspectual sense.

(60) '중이다' is recognized as VX:

```
나는 숙제를 하는 중이다 ~.
(S (NP-SBJ 나/NPN+은/PAU)
   (VP (VP (NP-OBJ 숙제/NNC+을/PCA)
           하/VV+는/EAN)
       (VX 중/NNX+이/CO+다/EFN))
   ./SFN)
```

'중이다' frequently appears attached to a Cino-Korean verbal noun. In order to give it a separate node as a VX, it is separated out from the noun, and the construction is annotated thusly:

(61) forced tokenization on 중/NNX+이/CO:

```
... 뉴욕에서 유세 ~중이던 고어 부통령은 ...
(NP-SBJ (S (WHNP-1 *op*)
           (S (NP-SBJ *T*-1)
              (VP (VP (NP-ADV 뉴욕/NPR+에서/PAD)
                      (VP (VV 유세/NNC)))
                  (VX 중/NNX+이/CO+던/EAN))))
        (NP 고어/NPR
            부통령/NNC+은/PAU))
```

Note that the recognition as VX only applies when 중/NNX is followed by a copula 이/CO. In all other circumstances, 중/NNX is treated in the same fashion as other dependent nouns:

(62) 중/NNX+에/PAD is not treated as VX:

```
.. 뉴욕에서 유세중에 ...
    (NP-ADV 뉴욕/NPR+에서/PAD)
    (NP-ADV 유세/NNC+중/NNX+에/PAD)
```

7.2 Parallel Treatment of Double Accusative and Double Nominative Constructions

In Korean Treebank 2.0, Korean double accusative and double nominative constructions receive parallel treatments. Therefore (indices 1 and 2 are not present in actual annotation; they are included here for illustration purposes):

(63) a. double accusative construction

```
존이 메리를 팔을 잡았다
(S (NP-SBJ 존/NPR+이/PCA)
   (VP (NP-OBJ1 메리/NPR+을/PCA)
       (VP (NP-OBJ2 팔/NNC+을/PCA)
           잡/VV+었/EPF+다/EFN)))
```

b. double nominative construction

```
존이 키가 작다
(S (NP-SBJ1 존/NPR+이/PCA)
   (S (NP-SBJ2 키/NNC+이/PCA)
      (ADJP 작/VJ+다/EFN)))
```

In NP-OBJ1 NP-OBJ2 VV, the inner object NP-OBJ2 and VV project up to VP; similarly, in NP-SBJ1 NP-SBJ2 VJ, the inner subject NP-SBJ2 and VJ project up to S. The analysis is compatible with some Korean syntactic theories which view the lower S unit a predicative clause ("서술절"). This S is a clausal unit, which dominates a subject and a predicate, yet functions as some sort of predicate relative to its sister NP-SBJ1, which it combines to project another S.

In the above cases, the outer NP element will not form any argument relation with the lexical verb/adjective. In certain other cases, the outer NP element is subcategorized by the lexical verb/adjective, which can be specified in the Korean Propbank:

(64) Propbank representation of "A-이 B-이 있"

```
존이 돈이 있다
(S (NP-SBJ1 존/NPR+이/PCA)
   (S (NP-SBJ2 돈/NNC+이/PCA)
      (ADJP 있/VJ+다/EFN)))

Arg2-nom   : john-nom
Arg1-nom   : money-nom
Rel        : exists
```

This leads to the theoretic implication in Korean grammar that verbs can assign argument roles outside of the lowest S clause that they are contained in. (A similar conclusion is drawn from the treatment of LV in the next section.)

7.3 Treatment of LV Extended to Non-OBJ Arguments

The light-verb construction, currently recognized for an object noun and a light verb pair, is extended similarly to include a subject noun and a light verb pair:

(65) a. light verb construction with NP-OBJ-LV

```
존이 공부를 한다
(S (NP-SBJ 존/NPR+이/PCA)
   (VP (NP-OBJ-LV 공부/NNC+을/PCA)
       (LV 하/VV+는 다/EFN))))
```

b. light verb construction with NP-SBJ-LV

```
흡연이 암과 관계가 있다
(S (NP-SBJ 흡연/NNC+이/PCA)
   (NP-COMP 암/NNC+과/PAD)
   (S (NP-SBJ-LV 관계/NNC+이/PCA)
      (LV 있/VJ+다/EFN)))
```

In NP-OBJ-LV LV pair above, it is the NP-OBJ that assigns argument structure; likewise in NP-SBJ-LV LV pair, it is the NP-SBJ that assigns argument roles to the S-external arguments '흡연이' and '암과'.

One of the theoretic implications introduced then by the treatment of the double-nominative and NP-SBJ-LV constructions is that Korean verbs (or LV constructions) can assign argument roles outside of the lowest S clause that they are contained in. Another related implication is that the phrase structure rule S -> NP-SBJ VP is no longer considered absolute for Korean: S will be viewed as the node NP-SBJ projects up to, while either VP or S can be the sister node to such a NP-SBJ.

7.4 VV Projection of NNC without XSV Suffix

Normally, a verbal noun undergoes verbalization via an attached verbalization suffix (XSV), which then projects up to a VV node (example 66a). In Korean Treebank 2.0, a verbal noun is allowed to undergo verbalization without the presence of a verbalization suffix (66b). Such cases are frequently found in a coordination structure, as illustrated in (67).

(66) a. a noun is verbalized with XSV
```
(VV 이륙/NNC+하/XSV+고/ECS)
```
b. a noun is verbalized without XSV
```
(VV 이륙/NNC)
```

(67)
```
;;3020002:6: 이날 이타크 항공의 일류신 76 기는 117 명의 승객자를 태우고
알 타시드 공군 비행장을 이륙 ~, 사우디의 제다공항에 도착했다 ~.

(S (NP-ADV 이날/NNC)
   (S (NP-SBJ (NP 이타크/NPR
                  항공/NNC+의/PAN)
              (NP 일류신/NPR
                  76/NNU
                  기/NNX+은/PAU))
      (VP (VP (NP-OBJ (NP 117/NNU
                          명/NNX+의/PAN)
                      (NP 승객자/NNC+을/PCA))
              태우/VV+고/ECS)
          (VP (NP-OBJ 알/NNC
                      타시드/NPR
                      공군/NNC
                      비행장/NNC+을/PCA)
              (VV 이륙/NNC))
          ,/SCM
          (VP (NP-COMP (NP 사우디/NPR+의/PAN)
```

```
                              (NP 제다/NPR+공항/NNC+에/PAD))
                    (VV 도착/NNC+하/XSV+었/EPF+다/EFN))))
    ./SFN)
```

7.5 ADV Can Modify Nominal Elements

In Korean Treebank 1.0, there were a few adverbs (ADV) such as '거의, 훨씬, 더, 바로' which were thought to be also capable of being adnominals (DAN) when modifying following nouns, a position taken in order to adhere to the doctrine that adverbials cannot modify nominal elements. In Korean Treebank 2.0, however, they are seen as retaining their original POS of ADV; instead, the view on noun modification is relaxed so that adverbial elements can now modify nouns. Hence:

(68)
```
훨씬 이전에
(NP-ADV (ADVP 훨씬/ADV)
        (NP 이전/NNC+에/PAD))

거의 1달러나 하락했다
(VP (NP-ADV (ADVP 거의/ADV)
            (NP 1/NNU
                달러/NNX+이나/PAU))
    (VP (VV 하락/NNC+하/XSV+었/EPF+다/EFN))))

더 이상
(NP-ADV (ADVP 더/ADV)
        (NP 이상/NNC))

바로 너에게
(NP-COMP (ADVP 바로/ADV)
         (NP 너/NPN+에게/PAD))
```

7.6 ADVP as Arguments

In Treebank 1.0, only noun phrases (NP) and clauses (S) were viewed as capable of functioning as an argument of a verb. Starting from Treebank 2.0, adverb phrases (ADVP) are treated as a COMP argument in some context, mostly involving the verbs 하/VV and 되/VV. Some examples:

(69) a.
```
(VP (NP-OBJ 방침/NNC+을/PCA)
    (ADVP-COMP 분명히/ADV)
    하/VV+었/EPF+다/EFN)
```

b.
```
(VP (NP-OBJ 대/XPF+북/NPR
            투자/NNC+을/PCA)
    (ADVP-COMP 원활히/ADV)
    하/VV+도록/ECS)
```

c.
```
(VP (NP-OBJ 문제/NNC+를/PCA)
    (ADVP-COMP (ADV (VJ 투명/NNC+하/XSJ+게/ECS)))
    하/VV+으며/ECS)
```

d.
```
(VP (ADVP-COMP 이렇게/ADV)
    되/VV+으면/ECS)
```

e. (VP (NP-OBJ 문 제/NNC+를/PCA)
(ADVP-COMP (ADV 가 볍/VJ+ 게/ECS))
보/VV+을/EAN)

ADVP-COMPs differ from ADVPs in that they are not semantic modifiers of the verb. Rather, they ascribe a property to some other element in the argument structure, typically the object (69a, 69b, 69c, 69e) or the subject (69d) in some cases. To further illustrate the point, compare (69b) with the following:

(70) (VP (NP-OBJ-LV 대/XPF+북/NPR
투 자/NNC+을/PCA)
(VP (ADVP 열 심 히/ADV)
(VP (LV 하/VV+도 록/ECS)))

In (69b), 하/VV is a causative verb: it is 투 자 "investment" that is being made 원활 "smooth". In (70), on the other hand, the adverb 열 심 히/ADV describes the mode "enthusiastically" of the act 투 자 를 함 "investing".

As a result of this change, Korean Treebank 2.0 now takes both '분명하게' and '분명히' as an argument in the examples below:

(71) '분명하게/분명히 하다' receive parallel analyses in KTB 2.0

a. (VP (NP-OBJ 요 점 을)
(ADVP-COMP (ADV 분 명 하/VJ+ 게/ECS))
하/VV+으 려 고/ECS)

b. (VP (NP-OBJ 요 점 을)
(ADVP-COMP 분 명 히/ADV)
하/VV+으 려 고/ECS)

In Korean Treebank 1.0, 분 명 하/VJ+ 게/ECS was assigned a clausal structure and was given an argument status on the S node; 분 명 히/ADV, however, could not head a clause and was left as a modifier as a result. Therefore, the syntactic and semantic parallelism of the two constructions was not properly captured in Treebank 1.0 annotations, which was corrected by the revision in Korean Treebank 2.0.

(72) "분명하게/분명히 하다" received disjoint analyses in KTB 1.0

a. (VP (NP-OBJ 요 점 을)
(S-COMP (NP-SBJ *pro*)
(ADJP 분 명 하/VJ+ 게/ECS))
하/VV+으 려 고/ECS)

b. (VP (NP-OBJ 요 점 을)
(VP (ADVP 분 명 히/ADV))
(VP 하/VV+으 려 고/ECS))

7.7 More VX-like Constructions Involving Keos/NNX

Auxiliary predicative noun '것이다' that contributes to modal or aspectual interpretation is labeled as VX. '것' always follows 을/EAN and is followed by copula '이' to be bracketed as VX.

(73) 만날 것이라고 밝혔다:
(VP (S-COMP ... (VP (VP ... 만나/VV+을/EAN)
(VX 것/NNX+이/CO+라/EFN+고/PAD)))
밝히/VV+었/EPF+다/EFN)

먹을 것이다:
(VP (VP ... 먹/VV+을/EAN)
(VX 것/NNX+이/CO+다/EFN))

예쁠 것이다:
(VP (ADJP 예쁘/VJ+을/EAN)
(VX 것/NNX+이/CO+다/EFN))

We do not view the following '것' as VX, although these seem to have the same semantics as their main clause counterpart. They get the usual treatment, as a complementized NP clause with 것/NNX as the head.

(74) (NP-COMP (S ... 초래/NNC+하/XSV+을/EAN) (NP 것/NNX+으로/PAD)) 본다
(NP-OBJ (S ... 초래/NNC+하/XSV+을/EAN) (NP 것/NNX+을/PCA)) 우려하여

Similarly, '뿐' and '터' also are bracketed as VX when occurring between 을/EAN and 이/CO:

(75) 먹을 뿐이다:
(VP (VP ... 먹/VV+을/EAN) (VX 뿐/NNX+이/CO+다/EFN))

집에 있을 터이니 전화해라:
(VP (VP (NP-COMP 집/NNC+에/PAD) 있/VV+을/EAN) (VX 터/NNX+이/CO+니/ECS)

7.8 VV Projection of noun+eu-Ro/PAD

In Korean Treebank 2.0, noun+으로/PAD is allowed to project to VV when it has arguments:

(76) (VP (NP-OBJ 친밀감/NNC+을/PCA)
(VV 바탕/NNC+으로/PAD))

(VP (NP-OBJ 의원/NNC+들/XSF+을/PCA)
(VV 상대/NNC+으로/PAD))

(S (NP-SBJ 최대 수출차는)
(VP (VV 푸조/NPR 시트로앵/NPR+으로/PAD)))

(S (NP-SBJ 반대의사를 밝힌 양당 의원은)
(VP (ADVP 모두/ADV)
(VP (VV 199/NNU+명/NNX+으로/PAD))))

7.9 Treatment of noun+eops-i/ADV

We break apart noun+없이/ADV into noun and 없이/ADV if the noun has a modifier, so that the modifier can modify the noun alone to project the NP argument of 없이/ADV. Note that '없이' is

now prefixed with the tokenization boundary marker '~', as previously explained in Section 2.2. Examples are:

(77) a. 아무런 이유없이:

```
아무런 이유 ~없이
(ADVP (NP-COMP 아무런/DAN
               이유/NNC)
       없이/ADV)
```

b. 유엔의 허가없이:

```
유엔의 허가 ~없이
(ADVP (NP-COMP (NP 유엔/NPR+의/PAN)
               (NP 허가/NNC))
       없이/ADV)
```

If there are no modifiers such as 아무런/DAN, there is no need to force tokenization on '이유없이'; it is tagged for individual morphemes, as in 이유/NNC+없이/ADV.

8 Summary of Tagset in Penn Korean Treebank 2.0

8.1 Content Tags

Category	Tag Description	Tag Label
noun	proper noun	NPR
	common noun	NNC
	dependent noun	NNX
	pronoun, demonstrative	NPN
	ordinal, cardinal, numeral	NNU
	words written in foreign characters	NFW
predicate	verb	VV
	adjective	VJ
	auxiliary predicate	VX
adverb	constituent adverb, clausal adverb	ADV
	conjunctive adverb	ADC
adnominal	configurative, demonstrative	DAN
interjection	exclamation	IJ
list	list marker	LST

8.2 Function Tags

Category	Tag Description	Tag Label	Note
postposition	case	PCA	
	adverbial	PAD	
	adnominal	PAN	new in KTB 2.0
	conjunctive	PCJ	
	auxiliary	PAU	
copula		CO	
ending	final	EFN	
	coordinate, subordinate, adverbial	ECS	
	~~auxiliary~~	~~EAU~~	merged with ECS in KTB 2.0
	adnominal	EAN	
	nominal	ENM	
	pre-final ending (tense, honorific)	EPF	
affix	suffix	XSF	
	prefix	XPF	
	verbalization suffix	XSV	
	adjectivization suffix	XSJ	

8.3 Symbols

Category	Tag Description	Tag Label
comma		SCM
termination	sentence ending markers	SFN
left quotation mark		SLQ
right quotation mark		SRQ
symbol	other symbols	SSY

References

[1] Martha Palmer, Chung-Hye Han, Na-Rae Han, Eon-Suk Ko, Hee-Jong Yi, Alan Lee, Chris Walker, John Duda and Nianwen Xue (2002). "Korean English Treebank Annotations" Linguistic Data Consortium (LDC) catalog number LDC2003L02 and ISBN 1-58563-265-1

[2] Chung-hye Han, Na-Rae Han, Eon-Suk Ko and Martha Palmer (2002). "Development and Evaluation of a Korean Treebank and its Application to NLP" *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002).*

[3] Chung-hye Han, Na-Rae Han, Eon-Suk Ko, Martha Palmer and Heejong Yi (2002). "Penn Korean Treebank: Development and Evaluation" *Proceedings of PACLIC (Pacific Asia Conference on Language, Information and Computation) 16.*

[4] Chung-hye Han, Na-Rae Han, Eon-Suk Ko and Martha Palmer (2002). "Development and evaluation of a Korean Treebank and its application to NLP" *Language and Information*, vol 6.1, pp 123-138.

[5] Chung-hye Han, Na-Rae Han and Eon-Suk Ko (2001). "Bracketing Guidelines for Penn Korean Treebank" IRCS, University of Pennsylvania

[6] Chung-hye Han and Na-Rae Han (2001). "Part of Speech Tagging Guidelines for Penn Korean Treebank" IRCS, University of Pennsylvania

[7] Andy Cole and Kevin Walker (2000). "Korean Newswire" Linguistic Data Consortium (LDC) catalog number LDC2000T45 and ISBN 1-58563-168-X