Enhancing Universal Dependencies for Korean

Youngbin Noh[1] **Jiyoon Han**[2] **Taehwan Oh**[3] **Hansaem Kim**[2]†

[1]Department of Cognitive Science, Yonsei University, Seoul, South Korea
[2] Institution of Language and Information Studies, Yonsei University, Seoul, South Korea
[3] Department of Korean Language and Literature, Yonsei University, Seoul, South Korea
{vincenoh, clinamen35, ghksl0604, khss}@yonsei.ac.kr

Abstract

In this paper, for the purpose of enhancing Universal Dependencies for the Korean language, we propose a modified method for mapping Korean Part-of-Speech(POS) tagset in relation to Universal Part-of-Speech (UPOS) tagset in order to enhance the Universal Dependencies for the Korean Language. Previous studies suggest that UPOS reflects several issues that influence dependency annotation by using the POS of Korean predicates, particularly the distinctiveness in using verb, adjective, and copula.

1 Introduction

The Universal Dependencies (UD) approach aims to find morphological and syntactic characteristics that can be applied to several languages for cross-lingual language processing. This approach converts the language resources of each language has into one unified format (CoNLL U-Format) in order to simplify general language processing.

The number of language resources varies among all languages. Designed with a focus on the characteristics of resource-rich languages, the CoNLL U-Format does not completely reflect the distinctiveness of Korean annotation. The CoNLL U-Format sets up minimum processing unit based on eojeol (white space), However, in Korean, the basic unit for language processing is not only eojeol but the also morphemes with eojeol. Therefore, an analysis of the Korean language with white space as the minimum unit may lead to the omission of some important information, or yield inaccurate results.

In Korean, the unit divided by white space is called eojeol, this is composed of a noun or verb stem combined with a postposition (josa) or ending (eomi) that function as inflectional and derivational particles. Significantly, based on the type of ending with which the stem of the predicate is combined to form an eojeol, the eojeol takes on different functions. Therefore, this paper suggests that Universal Part-of-speech (UPOS) be analyzed by taking this characteristic into consideration.

The CoNLL U-Format assigns a UPOS to each eojeol, and marks it with a language-specific part-of-speech tag (XPOS) beside it. This paper suggests a methodology that is able to clarify the UPOS and Dependency annotation by using XPOS after processing the Korean language.

For the purpose of enhancing Korean UD, the focus should be on the processing of predicates. In the Korean, most predicates consist of the combination of a stem and an ending, which is similar to Japanese predicate construction but differs from that of English and European languages. However, unlike Japanese text which can be segmented into stem and ending separately to give UPOS, as there is no white space in the language itself, Korean text includes a white space unit which has a construction that is different from English, and thus, we should consider this specific property.

This paper suggests a scheme for the mapping of part-of-speech that forms Korean predicates on UPOS, as well as suggests a method of annotating the dependency relations in case of verb sequences. In section 2, we examine the attempts to convert and build Korean language resources into UD. Furthermore, based on an analysis of the contents of previous Part-of-Speech (POS) annotations and dependency annotations, we search for the necessary areas of improvement for Korean annotation. In section 3, we suggest a UPOS for several issues that can have influences the dependency annotation by using XPOS of Korean predicates. In section 4, we will suggest a Korean

† corresponding author

Proceedings of the Second Workshop on Universal Dependencies (UDW 2018), pages 108–116
Brussels, Belgium, November 1, 2018. ©2018 Association for Computational Linguistics

dependency annotation modified through the suggested UPOS.

2 Previous works

Since the late 1990s, there have been attempts to build a syntactic parsing corpus of the Korean language based on the dependency structure. Korean National Corpus in the 21st Century Sejong Project (KNC), which has been established and is the most well-known syntactic parsing corpus, is based on the binary phrase structure and can easily to be converted to have a dependency structure. In fact, the dependency structure analysis corpus has been used widely in Korean language processing, and with a recent attempt to unify its form with UD. The efforts to apply current tag system of UD to the Korean language began with the Google Universal Dependency Treebank (UDT) Project (McDonald et al., 2013), which attempted to combine the Stanford Tag System and the Google System.

This paper discusses a total of three corpora: the Google Korean Universal Dependency Treebank (GKT), the Parallel Universal Korean Dependency Treebank (PUD), and the KAIST Korean Universal Dependency Treebank (KTB). All of the three corpora were tagged in CoNLL U-Format. The Google Korean Universal Dependency Treebank was first converted from the Universal Dependency Treebank v2.0 (legacy), and then enhanced by Chun et al. (2018). The KAIST Korean Universal Dependency Treebank was generated by Chun et al. (2018) from the constituency-based trees in the KAIST Tree-Tagging Corpus. The Parallel Universal Dependencies (PUD) treebanks are created from Raw Text to Universal Dependencies for the CoNLL 2017 shared task on Multilingual Parsing.

As stated in the Introduction, it is necessary to focus on predicate processing for enhanced Korean UD. Therefore, in section 2, this paper, examines GKT, PUD, and KTB centered on predicates; In 2.1, we will compare and analyze POS annotation methods of eojeols containing predicates in three corpora; and in 2.2, we will examine the Dependency annotation methods of these three corpora for eojeols containing predicates.

2.1 Part-of-speech annotation

For the POS annotation of eojeols containing predicates, all GKT, PUD, and KTB currently published on the UD Website follow the types of predicates contained in the eojeols without considering the function of the eojeols. It designates itself as VERB if the predicate in the eojeol is a verb and designates itself as ADJ if the predicate in the eojeol is an adjective, similar to the process of lemmatization.

(1) GKT Example - UPOS annotation as VERB for verb (VV) contained eojeol

# text = 조화가 잘 되어 아담한 감을 준다.

	조화+가	잘	되+어	아담하+ㄴ	감+을	주+ㄴ다	.
	jo-hwa-ga	jal	doe-eo	a-dam-han	gam-eul	jun-da	
	“Harmony + tpc”	“well”	“become”	“neat”	“Impression + obj”	make	
UPOS	NOUN	ADV	**VERB**	ADJ	NOUN	**VERB**	PUNCT
XPOS			**VV+EC**			**VV+EF**	

For copula, GKT and KTB adopt a similar methods of POS annotation while PUD takes different one. GKT and KTB label all eojeols that contain copula as VERB; however, PUD segments the copula from eojeols and assigns an AUX.

(2) GKT Example - UPOS annotation as VERB for copula (VCP) contained eojeol

# text = 바로 이곳입니다.

	바로	이곳+이+ㅂ니다	.
	ba-lo	i-gos-ib-ni-da	
	“exactly”	“here”	
UPOS	ADV	**VERB**	PUNCT
XPOS		**NP+VCP+EF**	

# text = 설립 이사장인 청암 박태준

	설립	이사장+이+ㄴ	청암	박태준
	seol-lib	i-sa-jang-in	cheong-am	park-tae-jun
	“Establishing”	“chairman”	“Cheongahm”	“Park Tae jun”
UPOS	NOUN	**VERB**	NOUN	NOUN
XPOS		**NNG+VCP+ETM**		

(3) PUD Example - UPOS annotation as AUX for copula(VCP) contained eojeol

# text = 비협조적인 사람이며

	비협조적	이+ㄴ	사람	이+며
	bi-hyeob-jo-jeog	in	sa-lam	i-myeo
	“Uncooperative”	“is”	“person”	“is”
UPOS	NOUN	**AUX**	NOUN	**AUX**
XPOS		**VCP+ETM**		**VCP+EC**

In addition, GKT and PUD tag both the main predicate as well as the auxiliary predicate as VERB. On the other hand, KTB gives an AUX tag to auxiliary predicates to differentiate between main predicates and auxiliary predicates.

(4) GKT Example - UPOS annotation as VERB for eojeol containing auxiliary verb(VX)

# text = 사용이 두드러지고 있다.

	사용+이	두드러지+고	있+다	.
	sa-yong-i	du-deu-leo-ji-go	iss-da	
	“Use+tpc”	“prominent”	“is”	
UPOS	NOUN	**VERB**	**VERB**	PUNCT
XPOS		**VV+EC**	**VX+EF**	

(5) KTB Example - UPOS annotation as AUX for eojeol containing auxiliary verb(VX)

# text = 현판이 달려 있었습니다.

	현판+이	달리+어	있+었+습니다	.
	hyeon-pan-i	dal-lyeo	iss-eoss-seub-ni-da	
	“Signboard+tpc”	“hang”	“being”	
UPOS	NOUN	**VERB**	**AUX**	PUNCT
XPOS		**VV+EC**	**VX+EF**	

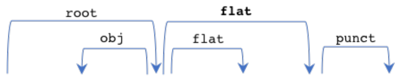
2.2 Dependency annotation

Of the 37 universal syntactic relations labels that represent Universal Dependency Relations, GKT, PUD, and KTB show the biggest difference in the `aux` (Auxiliary) and labels related to MEW (multi-word-expression), `compound` (Compound), `fixed` (Fixed Multiword Expression), and `flat` (Flat Multiword Expression). GKT demonstrates quite a low frequency of `aux,` which is because auxiliary predicates are not classified separately in the POS annotation process but are processed with VERB. Since an auxiliary predicate is not used alone but appears next to the main predicate, AUX which is a POS tag, and `aux,` which is a label and syntactic tag, should be proportional to each other. As such, it is natural for `aux` to show high frequency, as is the case with KTB. Instead, a `flat` label appears in GKT with high frequency whereas it appears with low frequency in other corpus as the relationship between the auxiliary predicate and the main predicate is processed as `flat`.

In this way, the Korean language has a predicate formed with continual eojeols in several cases; therefore it becomes critical to establish dependency relations between eojeols that form a predicate in the Korean language.

(6) GKT Example – tags VX as `flat`

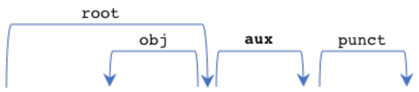
# text = 이야기를 하려 하지 않았다.



	Root	이야기+를	하+려	하+지	않+았+다	.
		ha-lyeo	ha-lyeo	ha-ji	anh-ass-da	
		“speak+obj”	“try”	“did”	“not”	
UPOS	ROOT	NOUN	VERB	VERB	**VERB**	PUNCT
XPOS					**VX+EC**	

(7) KTB Example – tags VX as(with) `aux`

# text = 말을 하지 않았다.



	Root	말+을	하+지	않+았+다	.
		mal-eul	ha-ji	anh-ass-da	
		“Talk+obj”	“did”	“not”	
UPOS	ROOT	NOUN	VERB	**AUX**	PUNCT
XPOS				**VX+EP+EF**	

3 Part-of-speech annotation

The target of this paper is limited only to the POS that can be used as a predicate in a sentence. In the Korean language, the POS applicable to the predicates are verb, adjective, copula, and auxiliary. The predicates of the Korean language can be composed of either a single eojeol or multiple ones. In the former case, one eojeol composed of a stem and ending with a verb or adjective functions as a predicate; alternatively, noun and copula combined with ending form an eojeol. In the latter case, stems of the main verb and auxiliary verbs are combined with an ending forming two or more eojeols in sequence, or many types of main verb stems are combined with an ending forming two or more eojeols in a sequence.

In this manner, the Korean language is very diverse in terms of predicate configurations; as the morphological and syntactic functions of predicate-related POS contained in these configurations are different, the POS annotation method in the prior process of parsing becomes an important issue that influences clear dependency annotation thereafter.

3.1 Verb

A Korean verb stem cannot be used alone in a sentence; it must be combined with an ending to form an eojeol. The types of ending are broadly categorized into final endings, connective endings, and conversion endings, and the function of an eojeol in a sentence is dependent on the ending with which the stem is combined. When combined with a final or connective ending, it has the function of a predicate, whereas it has the function of a modifier or substantive on being combined with a conversion ending.

As described in section 2, in GKT, KTB, and PUD, eojeols that contain Korean verbs are mostly tagged as VERB regardless of the type of ending that is combined with the verb stem. This simple labeling method does not consider the actual function of an eojeol in which the verb is contained. When this approach is used, the POS annotation is very likely to become redundant, which does not improve the accuracy of the dependency annotation.

This paper suggests that POS annotation should be performed according to the function of each eojeol within a sentence. In an eojeol where the verb stem functions as a predicate by being combined with a final or connective ending, the POS tag of VERB can be given. On the other hand, in an eojeol where the verb stem functions as modifying the following noun by being combined with an adnominal ending, the POS tag of ADJ can be given. When a verb stem is combined with a nominal ending, the function of the eojeol can be changed according to the type of combined inflectional particle.

Therefore, the POS tag of the eojeol follows the type of the combined inflectional particle. For helpful information in deciding the POS annotation of each eojeol, we can refer to the KNC POS tagset, which analyzes eojeol in morpheme units. Specifically, if the morpheme unit annotation of an eojeol is “VV (verb) + EF (final ending)” or “VV + EC (connective ending)”, the UPOS of VERB is allotted; if the morpheme unit annotation is “VV + ETM (adnominal ending)”, the tag ADJ is allotted.

However, among other connective endings combined with verb stem, “-게 (*-ge*)” provides an adverbial function to the eojeol, and it can often be tagged as ADV rather than VERB. It is the same in the case of the adjective and copula below. An eojeol made with a verb stem combined with a nominal ending and inflectional particle can be tagged with NOUN if it is “VV + ETN (nominal ending) + {JKS (subjective case particle), JKC (compliment case particle), JKO (objective case particle), JC (conjunctive particle)},” with ADJ if “VV + ETN + JKG (adjective case particle)”, and with ADV if “VV + ETN + JKB (adverbial case particle)”. The POS assignment method described above is applied only when the eojeol is not a part of a paragraph that presupposes a subject-predicate relation.

By following this method, more accurate UPOS annotation can be obtained by using morphological annotation information of the existing KNC. The UPOS established in this way would also be helpful in determining the dependencies relations.

(8) Verb contained eojeol UPOS annotation obtained from XPOS

# text = 자전거 타고 대전역 가는 길

	자전거	타+고	대전역	가+는	길
	ja-jeon-geo	ta-go	dae-jeon-yeog	ga-neun	gil
	“Bicycle”	“ride”	“Daejeon Station”	“go”	“way”
UPOS	NOUN	**VERB**	NOUN	**ADJ**	NOUN
XPOS		**VV+EC**		**VV+ETM**	

3.2 Adjective

Unlike English adjectives, Korean adjectives are sometimes classified as stative verbs, because a Korean adjective can function as predicate without a support verb. Therefore, it is hard to apply ADJ of UPOS to intact Korean adjectives. An English adjective does not change its form depending on whether it takes the role of predicate or modifier, and it can even form an eojeol on its own. Hence, it does not result in problem if it is tagged as ADJ. However, a Korean adjective stem, just like a verb, can complete an eojeol and be used in a sentence only when it is combined with an ending, and its function within the sentence can be changed based on the type of ending it is combined with.

In GKT, PUD, and KTB, eojeols that contain Korean adjectives are mostly tagged as ADJ regardless of the type of ending that is combined with the adjective stem in the sentence, or the function of the eojeol, in which the adjective is contained. Sometimes, an eojeol that functions as a predicate rather than a modifier is tagged as ADJ; and if this eojeol containing an adjective which functions as a predicate by being combined with a final ending or connective ending, is tagged as POS of ADJ, the sentence becomes a non-sentence, as it does not have a predicate.

Focusing on this characteristic of Korean adjectives, this paper suggests performing POS annotation for an eojeol that contains an adjective based on the function of the eojeol within the sentence. The approach is to annotate an eojeol in which an adjective stem functions as a predicate within a sentence when combined with a final or connective ending as VERB, and to annotate an eojeol in which an adjective stem functions as a modifier within sentence by being combined with adnominal ending as ADJ.

As stated above, Korean adjectives are analogous with stative verbs; hence, if one is combined with a final or connective ending, it functions as a predicate within a sentence. Therefore, a UPOS annotation as VERB is not irrational in the least, and it is suitable to give an ADJ tag to an eojeol that modifies a following eojeol by being combined with an adnominal ending. In addition, if an adjective stem is combined with a nominal ending, in the manner described in section 3.1, its function is changes according to the type of postposition with which it is combined; therefore, the POS of the eojeol can be determined according to the type of postposition.

(9) Adjective contained eojeol UPOS annotation obtained from XPOS

# text = 제일 가까운 스타벅스가 어디 있지

	제일	가깝+ㄴ	스타벅스+가	어디	있+지
	je-il	ga-kka-un	Starbucks+ga	eo-di	iss-ji
	“Most”	“close”	“Starbucks +tpc”	“where”	“be”
UPOS	NOUN	**ADJ**	NOUN	NOUN	**VERB**
XPOS		**VA+ETM**			**VA+EF**

An eojeol containing an adjective can also be annotated by referring to the KNC POS tagset that performs analysis in morpheme units. This information will be helpful in clarifying the dependency relations of a sentence by being used in dependency annotation.

3.3 Copula

The Korean copula “-이-(-*i*-)” is a unique POS that gives a predicate function to a noun. It appears with a noun and is similar to the English verb “be,” as it has the function of a predicate. But unlike the verb “be” which functions as a predicate alone by forming an eojeol without a noun, it can form an eojeol only by forming a “noun+copula+ending” structure. In addition, Korean copula, just like verb stem or adjective stem, can function as a predicate by combining it with a final or connective ending, or as a modifier or substantive by combining with a conversion ending.

In GKT and KTB, eojeols that contain copula are mostly tagged as VERB regardless of the type of ending that is combined with copula, similar to the eojeols that contain a Korean verb; in PUD, a copula is segmented from the eojeol and tagged as AUX. The former does not consider the actual function of the eojeol containing the copula, and the latter takes a method out of UD’s POS annotation guideline, which considers the eojeol as a basic unit.

This paper suggests differentiating the POS annotation of eojeols that contain copula according to the function of the eojeol in a sentence, just like the Korean verb or adjective stated above. To be specific, if an ending that completes an eojeol located next to a “noun+copula” structure is a final or connective ending when the eojeol functions as a predicate within the sentence, it is tagged as VERB; if the adnominal ending functions as a modifier, it is tagged as ADJ; and if combined with a nominal ending, POS annotation is done according to the type of inflectional particle. Here, we can also refer to the KNC POS tagset that annotates Korean language in morpheme units. Unlike Korean verbs or adjectives, the KNC POS tagset analysis on the eojeol containing copula that forms an eojeol by combining it with a noun would be like “NN* (noun) + VCP (copula) + E* (ending).”

(10) Copula contained eojeol UPOS annotation ontained from XPOS

# text = 설립 이사장인 청암 박태준

	설립	이사장+이+ㄴ	청암	박태준
	seol-lib	i-sa-jang-in	cheong-am	park-tae-jun
	“Establishing”	“chairman”	“Cheongahm”	“Park Tae jun”
UPOS	NOUN	**ADJ**	NOUN	NOUN
XPOS		**NNG+VCP+ETM**		

3.4 Auxiliary

The Korean auxiliary verb is different from the English auxiliary verb in several respects. Firstly, most English auxiliary verbs take forms that are different from main verbs; however in several cases, Korean auxiliary verbs are homonyms that take the same forms as the main verbs. Additionally, unlike English, which has completely different figures of “main verb||auxiliary verb” combinations and “main verb||main verb” combinations, in the Korean language the “main verb||auxiliary verb” combinations and “main verb||main verb” combinations, these combinations have the same syntactic structure in the Korean Language. Finally, in principle the Korean auxiliary verb is written with a space separating it from the main verb in order to form a separate eojeol, but sometimes it forms one eojeol with main verb in order to take the function of one predicate.

(11) ‘Main verb||main verb’ combination

# text = 김치를 맛있게 먹고 나오다

	김치를	맛있+게	먹+고	나오+다
	gim-chi-leul	mas-iss-ge	meog-go	na-o-da
	“Kimchi+obj”	“delicious”	“eat”	“out”
XPOS			**VV+EC**	**VV+EF**

(12) ‘Main verb||auxiliary verb’ combination

# text = 사용이 두드러지고 있다.

	사용+이	두드러지+고	있+다
	sa-yong-i	du-deu-leo-ji-go	iss-da
	“Use+tpc”	“prominent”	“is”
XPOS		**VV+EC**	**VX+EF**

In GKT and PUD, eojeols that contain the main verb and auxiliary verb are not divided; however, both of them are tagged as VERB. In this case, “main verb||main verb” combinations and “main verb||auxiliary verb” combinations are not distinguished, and in the case of the “main verb||auxiliary verb” combinations, it is hard to understand which one of two eojeols takes the role of the main predicate and which one takes the auxiliary function. In addition, “main verb||main verb” combinations and “main verb||auxiliary verb” combinations form different dependency relations. If POS annotation is unable to give this information properly, the whole sentence has to be analyzed again in the dependency annotation process.

This paper suggests applying different tags to the “main verb||main verb” combinations as wel as the “main verb||auxiliary verb” combinations by strictly classifying both of them. Sometimes the form of a Korean auxiliary verb is difficult to be distinguished from the main verb, but it is a closed set and small in number. In the KNC POS tagset that analyzes the Korean language in morpheme units, the main verb stem is tagged as VV or VA while the auxiliary verb is tagged as VX; Using this information, we can clearly and simply classify main verbs and auxiliary verbs in UD POS annotation. Therefore, the main verb can be tagged by VERB and the auxiliary verb by AUX, and when a main verb and an auxiliary verb form an eojeol, it can be tagged as VERB without segmenting the eojeol.

(13) ‘Main verb||main verb’ combination UPOS annotation obtained from XPOS

# text = 김치를 맛있게 먹고 나오다

	김치를	맛있+게	먹+고	나오+다
	gim-chi-leul	mas-iss-ge	meog-go	na-o-da
	“Kimchi+obj”	“delicious”	“eat”	“out”
UPOS			**VERB**	**VERB**
XPOS			**VV+EC**	**VV+EF**

(14) ‘Main verb||auxiliary verb’ combination UPOS annotation obtained from XPOS

# text = 사용이 두드러지고 있다.

	사용+이	두드러지+고	있+다
	sa-yong-i	du-deu-leo-ji-go	iss-da
	“Use+tpc”	“prominent”	“is”
UPOS		**VERB**	**AUX**
XPOS		**VV+EC**	**VX+EF**

3.5 Application result

When applying our proposal to GSD, the results are the same as in Table 1.

XPOS	UPOS	correct	total	revise	%
VV+EC	VERB	3,443	3,602	159	4%
VV+EF	VERB	414	560	146	26%
VV+ETM	ADJ	8	2,385	2,377	100 %
VV+-게	ADV	14	161	147	91%
VV+ETN+JKB	ADV	23	26	3	12%
XPOS	**UPOS**	**correct**	**total**	**revise**	**%**
VA+EC	VERB	609	1152	543	47%
VA+EF	VERB	7	278	271	97%
VA+ETM	ADJ	497	839	342	41%
VA+-게	ADV	233	250	17	7%
VA+ETN+JKB	ADV	2	5	3	60%
XPOS	**UPOS**	**correct**	**total**	**revise**	**%**
NN+VCP+EC	VERB	355	403	48	12%
NN+VCP+EF	VERB	575	575	0	0%
NN+VCP+ETM	ADJ	4	268	264	99%
NN+VCP+ETN +JKB	ADV	1	1	0	0%
XPOS	**UPOS**	**correct**	**total**	**revise**	**%**
VX+*	AUX	55	1,730	1,675	97%

Table 1 : application result on GSD

Because the UPOS is based on information of XPOS inside CoNLL U-Format, it can be automatically converted. The eojeol contaning VV showed the highest conversion rates to ADJ and ADV. And for eojeols contaning VA, the conversion rates to VERB and ADV was highest. In the case of eojeols contaning VCP, the conversion rates to VERB and ADJ was highest. Most eojeols starting with VX were converted to AUX.

4 Dependency annotation

4.1 Head final

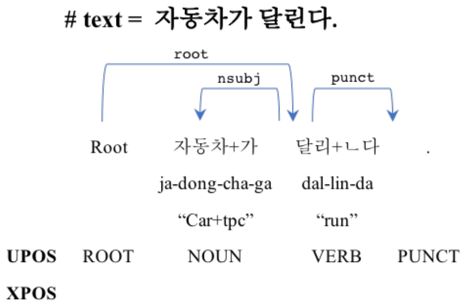
Unlike English, Korean language is a head-final language in which complement comes first followed by a head of verb phrase. This head is marked as `root` in the tag system of Universal Dependency Relations. This `root` is the core of dependency annotation as other sentence components are subordinated to this label. There are two kinds of Korean sentences; simple sentences and compound-complex. Compound-complex sentences can be divided into compound sentences that contain two consecutive simple sentences, and complex sentences that contain clauses with various kinds of sentence functions. It is easy to set up the root for a simple sentence, as there is only one predicate in the sentence.

However, it is difficult to determine the head of the sentence in compound-complex sentences. In the case of complex sentences, the last predicate is likely to be the head, but in the case of compound sentences, it is hard to assign a head as the ranks of two predicates within the sentence are equal. Accordingly, we would like to apply the head-final principle to compound sentences based on the cases of simple sentence and complex sentences.

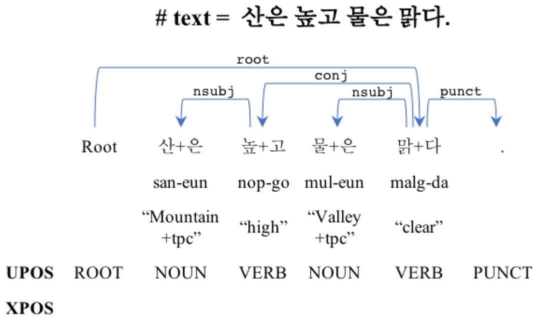
In the Japanese language, there is a tendency of setting up the right-most predicate stem in a sentence as the root while examining GSD, PUD, Modern, BCCWJ, and KTC corpora revealed by UD. In Korean, based on this criterion, it is necessary to set up the right-most eojeol containing a predicate as the root. This can be used to minimize confusion in the Korean language, which has a complex sentence structure.

Korean predicates play various role by being combined with endings; therefore, it is essential to first check whether the eojeol actually plays the role of predicate, upon setting up a predicate located in places other than the sentence final as the root. The error rate can increase if this process is omitted. The error rate can increase if this process is omitted. Therefore, following the principle of head final will reduce analysis errors and increase the efficiency of the processing.

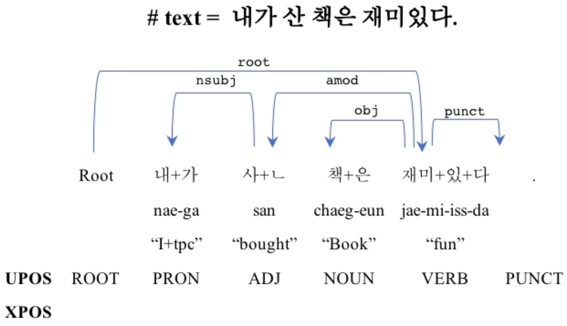
(15) Simple sentence



(16) Compound sentence



(17) Complex sentence



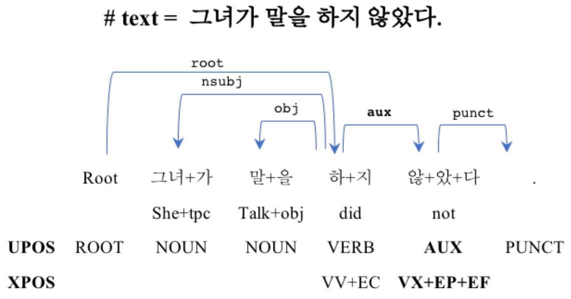
4.2 Verb sequence separated by white space

If two or more consecutive eojeols take on the same role of predicate within a sentence, the relationship between these eojeols should be revealed. Predicate eojeols are combined in the following cases: combination of main predicate with auxiliary predicate; and combination of two main predicate and main predicate.

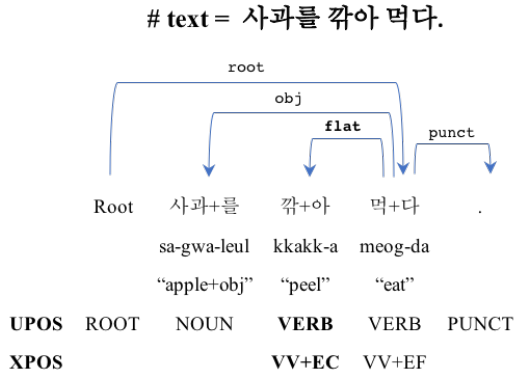
In the first case, `root` is assigned to the main predicate and `aux` is assigned to the auxiliary predicate. In the existing dependency syntactic parsing or structure analysis, if the head final is provided, head position is allotted to the auxiliary predicate; however, following the UD system, `aux` position is given to the auxiliary predicate and the head position is not. This shows that the auxiliary predicate does not describe actual contents of a sentence, but takes on an auxiliary function. Through this processing, we can resolve the controversy over whether to accept the Korean language should be accepted as a head final language.

In the case of a combination of two main predicates, the relationship is determined by the ending of the preceding main predicate. The following main predicate is labeled as `root`. However, if the preceding main predicate is combined with a connective ending, it can be assigned as `flat`; if the preceding main predicate is combined with an adverbial ending, it can be designated as `advcl` or `advmod`, depending on the combination relationship of the preceding Eojeol. Based on the KNC tagset, the connective ending is tagged by EC, and if the morpheme is “-게,” it can be considered as an adverbial ending. In the case of an adnominal ending, there is an extra label of ETM in the KNC tagset.

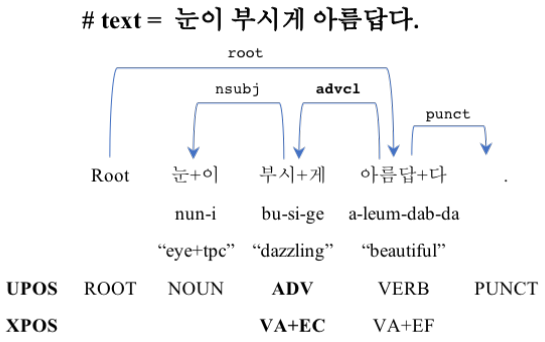
(18) Aux



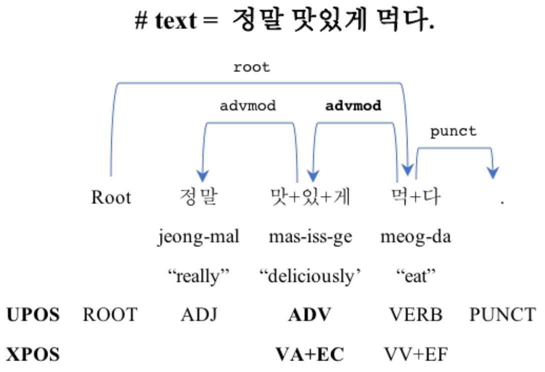
(19) flat



(20) advcl



(21) advmod



5 Conclusion and Future works

This paper aimed to suggest a plan for improving UD Tagging by focusing on the predicate. A Korean eojeol consists of nouns and verbs combined with propositions and endings that function as inflectional and derivational particles. The function of the stem of predicate depends on which ending is combined with the eojeol. Therefore, we proposed modified UPOS tagging and dependency annotation to reflect the syntactic characteristics of the Korean language using language-specific XPOS.

This paper developed the discussion by focusing only on predicates-related contents. eojeols with other functions that were not considered in this paper will be examined in future studies.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF), funded by the Korean Government (MSIT) (NRF-2017 M3C4A7068186).
This work was supported by the National Research Foundation of Korea (NRF), funded by the Korean Government (MSIT) (NRF-2009-361-A00027).

References

De Marneffe, M. C., & Manning, C. D. 2008. Stanford typed dependencies manual (pp. 338-345). Technical report, Stanford University.

Hansaem Kim, Korean National Corpus in the 21st Century Sejong Project(2006), Proceedings of the 13th National Institute of Japanese Literature (NIJL) International Symposium(pp 49–54)

Jayeol Chun1 , Na-Rae Han2 , Jena D. Hwang3 , Jinho D. Choi1, 2018. Building Universal Dependency Treebanks in Korean, In LREC.

Nivre, J., De Marneffe, M. C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., ... & Tsarfaty, R. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In LREC.

McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., ... & Bedini, C. 2013. Universal dependency annotation for multilingual parsing. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (Vol. 2, pp. 92-97).

Park, J., Hong, J. P., & Cha, J. W. 2016. Korean Language Resources for Everyone. In Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation: Oral Papers pp. 49-58.

Sag, Ivan A., Thomas Wasow, and Emily M. Bender. 2003. Syntactic Theory: A Formal Introduction, second edition. CSLI Publications.

Tanaka, T., Miyao, Y., Asahara, M., Uematsu, S., Kanayama, H., Mori, S., & Matsumoto, Y. 2016. Universal Dependencies for Japanese. In LREC.

金山博, 宮尾祐介, 田中貴秋, 森信介, 浅原正幸, & 植松すみれ. 2015. 日本語 Universal Dependencies の試案. 言語処理学会第 21 回年次大会, 505-508.