Annotation Issues in Universal Dependencies for Korean and Japanese

Ji Yoon Han♠, Tae Hwan Oh♢, Jin Lee♠, Hansaem Kim♠
♠Interdisciplinary Graduate Program of Linguistics and Informatics,
Yonsei University, Seoul, South Korea
♢Department of Korean language and literature, Yonsei University, Seoul, South Korea
{clinamen35, ghksl0604, sumomo, khss}@yonsei.ac.kr

Abstract

To investigate issues that arise in the process of developing a Universal Dependency (UD) treebank for Korean and Japanese, we begin by addressing the typological characteristics of Korean and Japanese. Both Korean and Japanese are agglutinative and head-final languages. And the principle of word segmentation for both languages is different from English, which makes it difficult to apply UD guidelines. Following the typological characteristics of the two languages and the issue of UD application, we review the application of UPOS and DEPREL schemes to the two languages. The annotation principles for `AUX`, `ADJ`, `DET`, `ADP` and `PART` are discussed for the UPOS scheme, and the annotation principles for `case`, `aux`, `iobj`, and `obl` are discussed for the DEPREL scheme.

1 Introduction

This article investigates issues arising in the process of building a Universal Dependency (UD) treebank for Korean and Japanese. The two languages possess similarities in typology. Korean and Japanese as a language have in common: SOV (subject-object-verb) word order and agglutination. In the design principles presented by the UD project, the principles do not conform to the characteristics of agglutinative languages. The following are part of design principles presented in the UD project:

1. *UD needs to be satisfactory on linguistic analysis grounds for individual languages.*
2. *UD needs to be good for linguistic typology, i.e., providing a suitable basis for bringing out cross-linguistic parallelism across languages and language families.*

For the Korean language, it is challenging to satisfy design principles 1 and 2. Berdicevskis et al. (2018) conducted a study on measuring cross-linguistic complexity using UD language resources. For a total of 37 languages, the study analyzes linguistic complexity utilizing a given measurement framework for morphological and syntactic complexity. In the provided measurement framework, morphological complexity has eight complexity variations that include TTR (Type-Token Ratio) and syntactic complexity has seven complexity variations that include CR_POSP (Diversity of POS bi-gram). The results indicate that measures of syntactic complexity might be on average less robust than those of morphological complexity. Unfortunately, the language analysis unit for Korean and Japanese were deemed too difficult and were excluded from this study.

In UD annotation scheme, the Part-of-Speech (POS) analysis results created a language specific part-of-speech layer (XPOS) field in addition to UPOS that exposes the characteristics within the individual languages. However, it is not easy nor ideal to apply the UD scheme created for an inflected language where the content and functional words are clearly separated by word dividing whitespaces, directly onto an agglutinative language in which functional words are integrated with the content word to form a single unit such as a word or an eojeol. In this study,

Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020), pages 99–108
Barcelona, Spain (Online), December 13, 2020

we examine the problem of how UD is applied to agglutinative languages and morphologically rich languages.

2 Typological characteristics of Korean and Japanese

Since UD was designed primarily for inflected languages, such as English, it is difficult to apply it directly onto the Korean language which is an agglutinative language. Characteristic to agglutinative languages, Korean is highly developed in postpositions and verbal endings. This is a major stumbling block to the application of UD to the Korean language. These problems are present not only in Korean but also in Japanese. Therefore, it is necessary to compare and analyze the application patterns of UD in Korean and Japanese. The following characteristics are mentioned as typological features of the Korean language, but also applies to the Japanese language:

> *1. Subject-Objective-Verb word order by default, but it is a relatively free word order language.*
> *2. As an agglutinative language, Korean is abundant in postpositions and verbal endings. Functional morphemes determine grammatical relations - not by word order*
> *3. Language in which the embedded clause precedes the main clause.*

The Korean and Japanese languages are both agglutinative languages and also commonly follow a SOV word order.(Sohn, 2001) Furthermore, as the most distinctive feature of languages with a SOV word order, such as Korean and Japanese, the two languages are rich in postpositions and verbal endings. This preceding nature is one of the key differences compared to English and other languages where functional words with grammatical functions are placed in front of the content word.

In terms of the location of head directionality, Korean and Japanese place the head on the right-hand side from the existing parsing resources, which is characteristic to head-final languages (Kanayama et al., 2018). In English, which is the standard for Stanford Dependencies, places the head on the left-hand side. In other words, the location of the head in coordinate structures for English is different from languages such as Korean or Japanese. In Korean and Japanese, the head usually appears on the right-hand side, so if the coordinate or parallel elements are annotated, the root is assigned to the right-hand element according to the head-final principle. To prevent this confusion, establishment of principles addressing this issue is in dire need. Figure 1 is an example of the annotation of conj, the relation label between two elements connected by a coordinating conjunction.

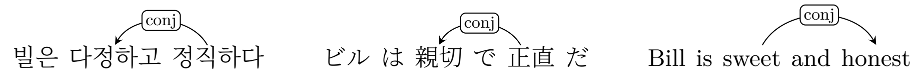

Figure 1: annotation of `conj`; Sentences that translate to ”Bill is sweet and honest.”

In both Korean and Japanese, there is a form of postposition(called particle in Japanese) and verbal endings. These represent grammatical relations among the content words. In the correlating Japanese component, the categories “particle(case):” and “particle(phrase_final)” are included in the part-of-speech, while only “case postpositions” is included in the Korean part-of-speech. In the Korean component, case postpositions are generally accepted as words, but verbal endings are not considered within the category of a “word.” This is because words are achieved only when combined with a verbal stem. Some of the endings in Korean correspond to the particle(case) in Japanese and some mapping to the particle(phrase_final). Compared to the other part-of-speech with relatively clear boundaries and clear lexical meanings, postposition and verbal ending are one cause of difficulties in assigning UD annotation labels to both languages. The boundaries are not only vague, but also carry stronger grammatical meanings as opposed to lexical meanings.

3 The issue of UD application

3.1 Tokenization and word segmentation

Setting the basic unit of the annotation is the most rudimentary step in dependency relations analysis. UD guidelines define dependencies as occurring between syntactic words. However, the criteria for defining syntactic words are vaguely presented, making it difficult to clearly define syntactic words in each of the languages. For English, whitespace boundaries define the unit and usually also the word for POS annotation. Accordingly, the formal boundaries are consistent with the basic units tagged in UPOS and DEPREL. However, the basic unit for morphological analysis in Korean is not defined by whitespace boundaries. In fact, the whitespace boundaries in Korean defines a unit known as "eojeol," which is a combined form of content words and functional words. This is significant for both Korean and Japanese as the function of sequence within a sentence is represented using postpositions and verbal endings included within an eojeol. For example, in table 1, the following phrase "학교 생활을(to school life)" consists of three morphemes: "학교(school)," "생활(life)," and "을(to)." In this case, "학교 (school)" modifies "생활(life)." And "을(to)" is an objective case marker that indicates "school life" functions as an object in a sentence. As shown in this example, there exists a relationship between morphemes that delivers important grammatical information. If, for example, eojeol was the basic annotation unit, only the relationship between "학교(school)" and "생활(life)" would be represented through the annotations. If "을(to)" is annotated independently, a case tag will be assigned during the DEPREL process and indicate that "을(to)" refers to the case in the sentence. But this case information will be missing if annotation is conducted with eojeol as the basic unit. In the case of Japanese, this problem does not occur because all three levels separately annotate "を(to)". The most debated topic on the study of UD application for the Japanese language is on defining the basic unit of annotation. Unlike Korean, Japanese does not use whitespaces as a unit divider. Existing Japanese corpora are annotated with dependency structures, including the Kyoto University Text Corpus (Kurohashi and Nagao, 2003) and the Japanese Dependency Corpus (Mori and Sasada, 2014). These corpora use bunsetsu as the syntactic dependency annotation units for Japanese.

Korean Sentence 학교 생활을 즐겁게 할 수 있을지도 모르는 방법

XPOS (Sejong)	학교 *hakgyo* NNG	생활 *saenghwal* NNG	을 *-ul* JKO	즐겁 *jeulgeop-* VA	게 *-key* EC	하 *ha-* VV	ㄹ *-l* ETM	수 *su* NNB	있 *iss-* VX	을지 *-eulji-* EC	도 *-do* JX	모르 *more-* VV	는 *-nun* ETM	방법 *bangbeop* NNG
eojeol	학교	생활을		즐겁게		할		수	있을지도			모르는		방법

Japanese Sentence 学校生活を楽しくするかもしれない方法

SUW	学校 *gakkō*	生活 *seikatsu*	を *-o*	楽しく *tanoshiku*	する *-suru*	か *-ka*	も *-mo*	しれ *-shire*	ない *-nai*	方法 *hōhō*
LUW	学校生活		を	楽しく	する	かもしれない				方法
bunsetsu	学校生活を			楽しく	するかもしれない					方法

Table 1: Comparison of Korean and Japanese annotation units for the sentence, which translates to "the way you might be able to enjoy school life."

3.2 UPOS annotation

UD guideline uses UPOS tagset, a common morphological analysis scheme, for multilingual processing. Korean uses Sejong Scheme for morphological annotations and Japanese uses Unidic. Table 2 exhibits the mapping of each annotation scheme. This chapter focuses on some of its label: `AUX`, `ADJ`, `DET`, `ADP`, `PART`, `CCONJ`, and `SCONJ`. Balanced Corpus of Contemporary Written Japanese (BCCWJ) has been automatically tokenized and PoS-tagged by NLP analysers in a three-layered tokenization of Short Unit Word (SUW), Long Unit Word (LUW), and bunsetsu. SUWs are defined by the morphological properties and are minimal atomic units that can be combined in ways specific to particular classes of Japanese words. LUWs are defined by the syntactic properties and bunsetsu are word grouping units defined by the dependency structure

(Omura and Asahara, 2018). BCCWJ_DepPara released in 2016 is the bunsetsu-level dependency structure annotations that relies on LUWs (Asahara and Matsumoto, 2016). In 2018, UD Japanese-BCCWJ adopted the SUW word unit. Unlike Japanese, Korean has the unit "eojeol" that is defined by the whitespace dividers. But lexical morphemes and functional morphemes make up one unit of eojeol. Therefore, Korean is also not excused from the conundrum of how to define the basic unit for annotation. Park et al. (2018) and Noh et al. (2018) defined eojeol as the basic unit of UD scheme. Park (2017) defines four different levels of segmentation granularity for Korean. The four levels are eojeol, punctuation, case markers, and verbal endings. The Sejong Treebank adopted eojeol as its basic unit (Hong, 2009). The Exobrain Corpus uses the same annotation system as the Sejong Treebank to express dependency and therefore also use eojeol as the basic annotation unit (Lim et al., 2015). however, there is still a need for discussion on determining the basic unit for syntax annotations and on how to best reflect linguistic characteristics.

UPOS	**Sejong(kor)**	**Unidic(Jap)**
`VERB`	VV+E ([NNG, NNP, MAG, XR])+XSV+E	verb noun(common.verbal suru)
`ADJ`	MM(attributive prenouns) VA+E VCN+E ([NNG, NNP, MAG, XR])+XSA+E ([N, MAG, SN])+VCP+E	adjective_i adnominal noun(adjectival)
`DET`	MM(except numeral & attributive prenouns)	adnominal

`ADP`	(JK, JX)	particle(case)
		particle(binding)

`AUX`	VX+E	verb(bound) adjective_i

`PART`	(EP, EC, EF, ET, XP, XS)	particle(phrase final)
		suffix(adjectival noun)

`CCONJ`	MAJ{및(mich), 또는(tto-neun)}
	JC
		particle(case)
		particle(adverbial)
		conjunction

`SCONJ`	MAJ{All access adverbs except '및(mich), 또는(tto-neun)'}	particle(conjunctive)
		particle(nominal)

Table 2: part of the mapping table between UPOS, Sejong POS and UniDic POS

1) `AUX`

`AUX` seems somewhat applicable to Korean and Japanese, but the specific morphological categories are actually quite different. In Japanese, the Unidic annotation scheme corresponds to bound verb(auxiliary verbs, non-independent verbs) and adjective_i(non-independent adjectives). The Japanese non-independent verbs and non-independent adjectives function similar to that of Korean auxiliary verbs and are categorized into the same morphological categories.

However, the verbal ending in Korean is regarded as an auxiliary verb in Japanese and consequently all annotated as an `AUX`. Therefore, in the case of (a), た*(*`-PAST`*)* is labeled as an auxiliary verb in the past tense and accordingly annotated as `AUX`. In Korean, -았다/-었다*(*`-PAST`*)* is a verb ending that is equivalent to the Japanese auxiliary verb た*(*`-PAST`*)*.In the Korean language, verbal ending is not allowed to construct eojeols independently and is not recognized as a part-of-speech. Likewise, only auxiliary verbs annotated as a VX in the Sejong annotation system are annotated as AUX in UD. In the case of (b), `AUX` is not singled out and assigned in Korean. This is because the Korean verbal ending -하다*(do)* corresponds directly to the Japanese auxiliary verb する*(do)* and does not constitute a separate phrase. However, in (c) and (d), there is a Korean counterpart to the Japanese auxiliary verbs いる*(be)* and ない*(not)*. And for that reason, it is respectively annotated as `AUX` as an independent unit.

(a)	kor	먹었다	
		meogeossda	
		eat+-`PAST`	
		`VERB`	
	jap	食べ	た
		tabe	*-ta*
		eat	-`PAST`
		`VERB`	`AUX`

(b)	kor	공부하다	
		gongbuhada	
		study+do	
		`VERB`	
	jap	勉強	する
		benkyō	*-suru*
		study	do
		`VERB`	`AUX`

(c)	kor	먹고	있다
		meokgo	*itda*
		eat+-`ADP`	-ing
		`VERB`	`AUX`
	jap	食べて	いる
		tabete	*-iru*
		eat+-`ADP`	-ing
		`VERB`	`AUX`

(d)	kor	먹지	않다
		meokji	*anta*
		eat+-`ADP`	not
		`VERB`	`AUX`
	jap	食べ	ない
		tabe	*-nai*
		eat	not
		`VERB`	`AUX`

2) `ADJ` and `DET`

In English, a be verb is used to make an adjective a predicate. In Korean and Japanese, however, an adjective can be used alone as a predicate and can also be used as a modifier. In Japanese, `ADJ` includes adjective_i, adnominal, noun(adjectival). (e) is an example of adjective_i, and (f) is an example of adnominal. noun(adjectival) is noun, but it can function as adjective, such as in (g). In Japanese, the morpheme 健康だ*(healthy)* is converted to "健康な*(healthy)* + noun" when modifying a noun. In this case, 健康 *(health)* has a form of a noun, but annotated as an `ADJ` and な*(affix)* is annotated as an `AUX`.

Lastly, Japanese adnominals are similar to Korean prenoun. In Korean, only attributive prenouns are classified as an `ADJ`. Demonstrative prenouns are classified as a `DET`, and numeral prenouns are classified as a `NUM`. In Japanese, adnominals that convey a meaning of determining something, such as この*(this)*, その*(that)*, あんな*(that)* and どんな*(what)* are classified as a `DET`.

(e)	kor	빨간	사과
		ppalgan	*sagwa*
		red	apple
		`ADJ`	`NOUN`
	jap	あかい	りんご
		akai	*ringo*
		red	apple
		`ADJ`	`NOUN`

(f)	kor	큰	가방
		keun	*gabang*
		large	bag
		`ADJ`	`NOUN`
	jap	大きな	かばん
		ōkina	*kaban*
		lare	bag
		`ADJ`	`NOUN`

(g)	kor	건강한		사람
		geonganghan		*saram*
		healthy		person
		`ADJ`		`NOUN`
	jap	健康	な	人
		kenkō	*-na*	*hito*
		health	-`ADP`	person
		`ADJ`	`AUX`	`NOUN`

3) `ADP` and `PART`

Since Korean and Japanese are agglutinative languages, the postposition or verbal ending is combined with a content word to show various grammatical relationships. Postpositions can also function as case indicators that describe the relationship that the noun it is dependent on has with other words or add meaning to the noun it is dependent on. For example, the Japanese phrase in (i), きれいですね*(pretty)* is broken down in the form of an "adjective_i + auxiliary_verb + particle(phrase_final". But in Korean, the corresponding counterpart is annotated simply as an adjective because the whitespace word boundaries define the basic unit.

Japanese has various types of postpositions(particles) - `ADP` includes particle(case) and particle(binding), `PART` includes particle(phrase_final). `PART` also includes the suffix(adjectival_noun) 的*(affix)*. In Korean, when UPOS tags are allocated by the eojeol unit, postpositions and verbal endings are only annotated as `ADP` and `PART` when they are separated by a punctuation mark or an identifying symbol.

(h)

kor	나는		집에		간다
	naneun		*jibe*		*ganda*
	I+-`NOM`		home+-`DAT`		go
	`NOUN`		`NOUN`		`VERB`
jap	私	は	家	に	行く
	watashi	*-wa*	*ie*	*-ni*	*iku*
	I	-`NOM`	home	-`DAT`	go
	`NOUN`	`ADP`	`NOUN`	`ADP`	`VERB`

(i)

kor	예쁘네요		
	yeppeuneyo		
	pretty		
	`ADJ`		
jap	きれい	です	ね
	kirei-	*-desu-*	*-ne*
	pretty		
	`ADJ`	`AUX`	`PART`

3.3 DEPREL annotation

1) `case` and `aux`

Japanese uses case-marking and predicate-argument structure information to allocate UD DEPREL annotations. When annotating a predicate-argument structure, it often utilizes case-marking information to allocate DEPREL annotations. For example, postposition は*(topic marker)* is a case-marker revealing a dependency relationship (`nsubj`) that also functions as a topic marker at the same time. This has a function similar to the postposition *(은/는(topic marker))* in Korean. In Japanese, a short-unit is laid out as the basic unit for parsing, so the postposition is recognized as an independent unit and assigned the case annotation.

In Korean, on the other hand, since postposition is not a self-dependent component, it is always actualized by merging with another word such as a noun, a pronoun, or a numeral. Therefore, if words are used as the basic unit for parsing, the postposition is not recognized as an independent unit and cannot be processed separately from the preceding element. The only exception to this is if a postposition is separated from the dependent word using a punctuation mark or a symbol, then it is recognized as an independent unit. As such, postposition in Korean and Japanese has similar characteristics and appears at a high frequency.

As previously mentioned, `aux` is a label that corresponds to auxiliary verbs in Korean and modal verbs in Japanese. It supplements the meaning of verbs or adds meaning to the entire sentence. Ultimately, this is deeply correlated to the issue of determining the basic word unit that occurs in the process of language data processing. In Japanese, auxiliary_verbs are assigned the `AUX` label using UPOS and almost always correspond to `aux` in DEPREL. Since the UPOS label is already processed as a separate unit, it is more intuitive to receive a separate annotation in DEPREL. Additionally, the existence of forms that supplement the meaning of a predicate is a common phenomenon in agglutinating words, so we can anticipate high-frequency rates accordingly. Auxiliary verb(VX)s in Korean are also similar in function to auxiliary_verbs in Japanese. In other words, it plays the role of supplementing the lexical meaning or adding the grammatical meaning to the main verb. However, language data shows us that Korean auxiliary verbs are generally less independent than Japanese auxiliary verbs. Auxiliary verbs are often combined with the main verb because the actual meaning on its own is relatively not strong enough. Also important to note, in Japanese, `aux` also includes verb(bound)s that add grammatical meaning to verbs, which in most of them correspond to verbal endings rather than auxiliary verbs in Korean.

2) `iobj` and `obl`

In addition to `case` and `aux`, `iobj` is a label for annotating indirect objects. The Korean and Japanese usage of this label differs significantly. An indirect object is a component that helps distinguish two objects in an argument. For example, in the sentence, “She gave me a book,” the indirect object is “me” and the direct object is “a book.” So in order to represent this

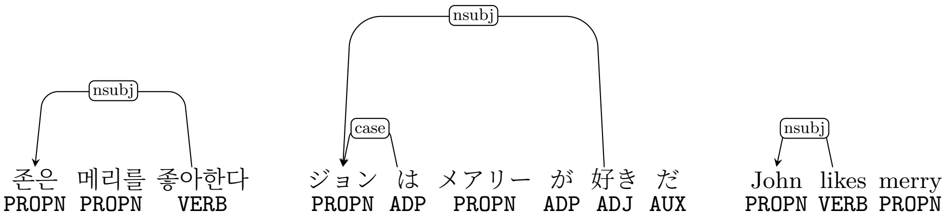

Figure 2: annotation of `case`; Sentences that translate to "John likes merry."

dependency relationship, UD created a label known as `iobj` and suggested guidelines to assign it to indirect objects.

The problem is that in Korean and Japanese, excluding exceptional circumstances, there are very few cases in which two or more essential object arguments appear in a sentence. Instead, various postpositions are combined to show their relationship and role with the verb. For example, in Korean, indirect objects are realized as a noun phrase combined with an adverbial postposition*(에/에게(to))*. These adverbial postpositions function as case-marking indicators that identify indirect objects, but also for some adjuncts like agent, comparison, and destination. Therefore, it is difficult to distinguish between adjunct and indirect objects in Korean. Since the distinction between an indirect object and an adjunct is not clear, the `iobj` label is not used in such arguments. Instead, the label `obl` assigned to adjuncts is used.

By contrast, according to the guidelines released by UD, the Japanese case-marker for indirect objects is presented as に*(to)*. However, there are some problems with this. First of all, in the IPA part-of-speech system, the most common part-of-speech system in Japanese, に*(to)* is classified as an adverbial postposition and it has various meanings similar to the Korean adverbial postposition*(에/에게(to))*.

If the UD annotation scheme does not portray the characteristics of a specific language very well, it is best to selectively apply labels in a way that reflects the characteristics of the language rather than forcefully manipulating the label to use it all up. The recently published study by Omura and Asahara (2018) shows that the BCCWJ-DepPara corpus of the National Institute of the Japanese Language did not fully reflect the DEPREL labels.

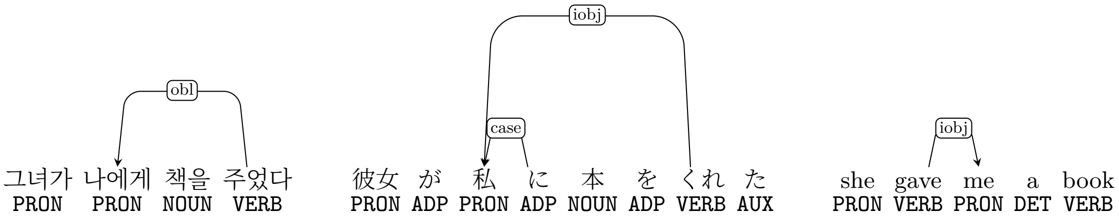
obl
그녀가 나에게 책을 주었다
PRON PRON NOUN VERB
iobj
case
彼女 が 私 に 本 を くれ た
PRON ADP PRON ADP NOUN ADP VERB AUX
iobj
she gave me a book
PRON VERB PRON DET VERB

Figure 3: annotation of `iobj`; Sentences that translate to "She gave me a book."

4 Korean and Japanese UD corpora

There are five Korean treebanks that are registered on the UD project website: the Google Korean UD Treebank (McDonald et al., 2013), the Kaist UD Treebank (Choi et al., 1994), the Parallel Universal Dependencies Treebank (Zeman et al., 2017), the Penn Korean UD Treebank (Chun et al., 2018) and the Sejong UD Treebank (Choi and Palmer, 2011).

Five Japanese treebanks are also registered: the BCCWJ UD treebank (Maekawa et al., 2014), the Kyoto Text UD treebank (Tanaka et al., 2016), the Google Japanese UD Treebank (Zeman et al., 2017), the Parallel Universal Dependencies Treebank (Zeman et al., 2017) and the Modern Japanese UD Treebank (Omura and Asahara, 2017).

In this section, the Penn Korean UD Treebank and the BCCWJ UD Treebank, which were most recently revised, are compared in terms of UPOS and DEPREL label usage. The table 3 shows that the difference in UPOS is significant for the `ADP` label. It accounts for 20.1% of

BCCWJ, but only 1.08% of PKT. Whereas `AUX` accounts for only 3.07% in PKT, but 11.58% in BCCWJ. And `ADJ` in the Korean PKT accounted for 5.33% of PKT and 2.16% in the Japanese BCCWJ. With the DEPREL scheme, the biggest gap between the two languages appeared in the total ratio of `cases`. `case` in PKT was not used, but accounted for 19.87% of the total in BCCWJ. Likewise, `iobj` was not used in PKT, but used in BCCWJ. `aux` also made up 3.08% of PKT and 10.36% in BCCWJ.

UPOS	PKTv2020	PC	BCCWJ	PC
ADJ	7,034	5.33%	27494	2.16%
ADP	1,425	1.08%	255976	20.10%
ADV	2,851	2.16%	18815	1.48%
AUX	4,060	3.07%	147400	11.58%
CCONJ	377	0.29%	16142	1.27%
DET	685	0.52%	5357	0.42%
INTJ	0	0.00%	915	0.07%
NOUN	58,367	44.20%	369172	28.99%
NUM	7,602	5.76%	58321	4.58%
PART	290	0.22%	7456	0.59%
PRON	1,142	0.86%	11557	0.91%
PROPN	12,769	9.67%	35938	2.82%
PUNCT	13,428	10.17%	107005	8.40%
SCONJ	533	0.40%	41512	3.26%
SYM	376	0.28%	60957	4.79%
VERB	21,102	15.98%	108692	8.54%
X	0	0.00%	578	0.05%
total	132,041	100.00%	1273287	100.00%

DEPREL	PKTv2020	PC	BCCWJ	PC
acl	11,210	8.49%	31794	2.50%
advcl	5,086	3.85%	30221	2.37%
advmod	3,125	2.37%	16940	1.33%
amod	1,593	1.21%	22253	1.75%
appos	1,173	0.89%	0	0.00%
aux	**4,061**	3.08%	**131946**	10.36%
case	0	0.00%	253009	19.87%
ccomp	1,989	1.51%	0	0.00%
cc	473	0.36%	16120	1.27%
compound	21,433	16.23%	170525	13.39%
conj	7,155	5.42%	0	0.00%
cop	0	0.00%	5661	0.44%
csubj	8,012	6.07%	0	0.00%
dep	10	0.01%	81623	6.41%
det	685	0.52%	5356	0.42%
fixed	589	0.45%	0	0.00%
flat	739	0.56%	0	0.00%
goeswith	2,199	1.67%	0	0.00%
discourse	0	0.00%	834	0.07%
dislocated	0	0.00%	379	0.03%
iobj	0	0.00%	15689	1.23%
mark	0	0.00%	41369	3.25%
nmod	5,501	4.17%	113787	8.94%
nsubj	4,114	3.12%	55117	4.33%
nummod	7,341	5.56%	53859	4.23%
obj	9,849	7.46%	33059	2.60%
obl	16,891	12.79%	29630	2.33%
orphan	9	0.01%	0	0.00%
punct	13,794	10.45%	106990	8.40%
reparandum	0	0.00%	17	0.00%
root	5,010	3.79%	57109	4.49%
total	132,041	100.00%	1273287	100.00%

Table 3: Universal dependency label comparison between PKT & BCCWJ

5 Conclusion

We reviewed the application of the UD scheme to the Korean and Japanese treebanks as we identified and discussed the areas that require awareness of when constructing a UD treebank for an agglutinative language. We identified issues that arise when determining the basic units and in applying UPOS and DEPREL schemes. For the UPOS scheme, issues related to applying `AUX`, `ADJ`, postposition, and verbal ending were addressed. For the DEPREL scheme, the application and usage of `case`, `aux`, and `iobj` labels are discussed. The above discussions will be essential in establishing standards for building or improving UD treebanks in agglutinative languages in the future.

This review of the current state of UD treebanks for agglutinative languages discloses a need for a UD treebank that better reflects the unique characteristics of the language for construction. PKT was revised to further reflect the unique characteristics of the Korean language by modifying UPOS and DEPREL in reference to the Korean XPOS (Oh et al., 2020). But still, the basic unit for analysis is a word unit, which does not capture all the syntactic functions of postpositions or verbal endings. The Japanese UD makes better use of its unique characteristics than the Korean UD in the sense that it uses its own units. However, as we can see from the application of `iobj` that there is still much room for improvement.

The UD scheme is evolving through continuous research and workshops. As a result, UD treebanks are also becoming more diverse. This study focused on Korean and Japanese to examine the characteristics of agglutinative languages, but the other languages such as Turkish did not explain together, which is left as future work. We hope that this article will contribute to the vitalization of discussions on agglutinative languages.

References

Masayuki Asahara and Yuji Matsumoto. 2016. Bccwj-deppara: A syntactic annotation treebank on the 'balanced corpus of contemporary written japanese'. In *Proceedings of the 12th Workshop on Asian Language Resources (ALR12)*, pages 49–58.

Aleksandrs Berdicevskis, Çağrı Çöltekin, Katharina Ehret, Kilu von Prince, Daniel Ross, Bill Thompson, Chunxiao Yan, Vera Demberg, Gary Lupyan, Taraka Rama, et al. 2018. Using universal dependencies in cross-linguistic complexity research. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 8–17.

Jinho D. Choi and Martha Palmer. 2011. Statistical dependency parsing in korean: From corpus generation to automatic parsing. In *Proceedings of the Second Workshop on Statistical Parsing of Morphologically Rich Languages*, pages 1–11. Association for Computational Linguistics.

Key-Sun Choi, Young S Han, Young G Han, and Oh W Kwon. 1994. Kaist tree bank project for korean: Present and future development. In *Proceedings of the International Workshop on Sharable Natural Language Resources*, pages 7–14. Citeseer.

Jayeol Chun, Na-Rae Han, Jena D Hwang, and Jinho D Choi. 2018. Building universal dependency treebanks in korean. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).*

Y Hong. 2009. 21st century sejong project results and tasks. In *In New Korean Life.*

Hiroshi Kanayama, Na-Rae Han, Masayuki Asahara, Jena D Hwang, Yusuke Miyao, Jinho D Choi, and Yuji Matsumoto. 2018. Coordinate structures in universal dependencies for head-final languages. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 75–84.

Sadao Kurohashi and Makoto Nagao. 2003. Building a japanese parsed corpus. In *Treebanks.*

Joon-Ho Lim, Yongjin Bae, Hyunki Kim, Yunjeong Kim, and Kyu-Chul Lee. 2015. Korean Dependency Guidelines for Dependency Parsing and Exo-Brain Language Analysis Corpus. In *Proceedings of the 27tht Annual Conference on Human and Cognitive Language Technology.*

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, et al. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97.

Hideki Ogura Mori, Shinsuke and Tetsuro Sasada. 2014. A japanese word dependency corpus. In *Proceedings of the seventh International Conference on Language Resources and Evaluation (LREC 2014).*

Younbin Noh, Jiyoon Han, Taehwan Oh, and Hansaem Kim. 2018. Enhancing universal dependencies for korean. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018).*

Taehwan Oh, Jiyoon Han, Hyonsu Choe, Seokwon Park, Han He, Jinho D. Choi, Na-Rae Han, Jena D. Hwang, and Hansaem Kim. 2020. Analysis of the penn korean universal dependency treebank (pkt-ud): Manual revision to build robust parsing model in korean. In *In Proceedings of the 16th International Conference on Parsing Technologies(IWPT 2020).*

Yuta Takahashi Omura, Mai and Masayuki Asahara. 2017. Universal dependency for modern japanese. In *Proceedings of the 7th Conference of Japanese Association for Digital Humanities (JADH2017).*

Mai Omura and Masayuki Asahara. 2018. Ud-japanese bccwj: Universal dependencies annotation for the balanced corpus of contemporary written japanese. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 117–125.

Hyejin Park, Taehwan Oh, and Hansaem Kim. 2018. Universal pos tagset for korean. *The Korean Society for Language and Information*, 22(3):67–89.

Jungyeul Park. 2017. Segmentation granularity in dependency representations for korean. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 187–196.

Ho-Min Sohn. 2001. *The korean language.* Cambridge University Press.

Takaaki Tanaka, Yusuke Miyao, Masayuki Asahara, Sumire Uematsu, Hiroshi Kanayama, Shinsuke Mori, and Yuji Matsumoto. 2016. Universal dependencies for japanese. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1651–1658.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonça, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada, August. Association for Computational Linguistics.