Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3826–3833
Marseille, 11–16 May 2020

An Enhanced Mapping Scheme of the Universal Part-Of-Speech for Korean

Maria Myung-Hee Kim and Nathalie Colineau
Defence Science and Technology (DST)
DST Edinburgh, SA 5111
{maria.kim, nathalie.colineau}@dst.defence.gov.au

Abstract

When mapping a language specific Part-Of-Speech (POS) tag set to the Universal POS tag set (UPOS), it is critical to consider the individual language's linguistic features and the UPOS definitions. In this paper, we present an enhanced Sejong POS mapping to the UPOS in accordance with the Korean linguistic typology and the substantive definitions of the UPOS categories. This work updated one third of the Sejong POS mapping to the UPOS. We also introduced a new mapping for the KAIST POS tag set, another widely used Korean POS tag set, to the UPOS.

Keywords: Universal Part-of-Speech, Sejong POS, KAIST POS, Korean NER

1. Introduction

The Universal Part-Of-Speech tag set (UPOS) (Petrov et al., 2012) has been widely used as a standard for mapping various POS tag sets across multiple languages to facilitate building and evaluating multi-lingual taggers and parsers (McDonald et al., 2013; Rosa et al., 2014; Park, 2017). In 2014, the extended UPOS (UD annotation guidelines v1, 2014) was released as part of the Universal Dependencies (UD) project, which aims to develop cross-language treebank annotation. The recent UD version 2.5 has released 157 treebanks in 90 languages in Nov 2019[1] .

When mapping a language specific POS tag set to the UPOS, it is important to map it correctly based on: (1) the definitions of the UPOS categories; and (2) the individual language's linguistic topology. It should not be just matching to the equivalent categories' name[2]. When the original UPOS (UPOS12) was proposed by Petrov, each category was defined operationally (Petrov et al., 2012). However, there were some ambiguities to the interpretation of these categories.

This problem was addressed in the extended UPOS (UPOS14) in 2014 (UD annotation guidelines v1, 2014) [3]. Clear definitions and explanatory examples were provided for the 17 categories of the UPOS including five new categories[4], which were identified as essential categories to many languages. Following the detailed guidelines, one can build an appropriate mapping scheme and justify their mapping results.

The first morpheme-based mapping for Korean POS tag set to the UPOS12 was introduced by Park and colleagues (2016). They proposed a mapping for the Sejong POS tag set, one of the most widely used Korean POS tag set. Recently, Park and Tyers (2019) updated this mapping to the UPOS14. In this work, they also discussed how the mapping can be applied to build a training corpus for Named Entity Recognition (NER) and Semantic Role Labelling (SRL) tasks.

While providing an updated mapping, Park and Tyers (2019) did not make substantial changes to the initial UPOS mapping (Park et al., 2016)[5]. More importantly, the updated mapping was not discussed nor justified.

In this paper, we review the Sejong POS mapping to the UPOS14 proposed by Park and Tyers (2019), and suggest and discuss linguistically motivated changes. We also present two new mapping schemes for Korean POS tag sets (Sejong POS and KAIST POS) to the UPOS14. Finally, we discuss the potential impact of the new Sejong POS mapping to the UPOS on Korean NER task a through qualitative analysis.

2. Korean Segmentation

English is defined as a fusional language where *a single inflection* [6] *can convey multiple grammar roles*. For example, the English sentence "*John draw**s** a picture.*" consists of four words ('*John*', '*draws*', '*a*' and '*picture*') and five morphemes ('*John*', '*draw*', 's', '*a*' and '*picture*'); the one bound morpheme **–s** carries three grammatical roles including person (third person), number (singular) and tense (present). English has one inflectional morpheme per word and is a weakly inflected language with only eight inflectional morphemes[7] (Brinton, 2000; Yule, 2010). As a result, English text is usually segmented at the word level without compromising further downstream processing in Natural Language Processing (NLP).

Korean, on the other hand, is defined as an agglutinative language where *each inflection conveys only a single grammatical role* (Eifring and Theil, 2004). For example,

[1] https://universaldependencies.org/

[2] Note that words called as numerals in some languages (e.g. Czech) should be tagged with the UPOS ADJ (adjective) as recommended by UD annotation guidelines v1, 2014.

[3] UD annotation guidelines v2 was released in 2016 (UD annotation guidelines v2). The UPOS from v1 was kept the same in v2 except for two minor revisions (https://universaldependencies.org/v2/summary.html); it includes updates on definitions of AUX and PART tags.

[4] PROPN (proper noun), AUX (auxiliary), SCONJ (subordinate conjunction), SW (symbol) and INTJ (interjection)

[5] Only two mappings were updated to the newly introduced UPOS14 categories: PROPN and INTJ.

[6] The modification of a word to express different grammatical categories such as tense, case, voice, aspect, person, number, gender and mood.

[7] Eight inflectional morphemes: two for nouns (**-s** for plural and **'s** for possessive), two for adjectives (**-er** for comparative and **–est** for superlative) and four for verbs (**-s** for 3rd singular present tense, **-ed** for past tense, **-ed** for past participle and **–ing** for present participle).

as shown in Table 1, the sentence "나는 친구들을 만났다." consists of three words and eight morphemes. Moreover, Korean is a highly inflected language with many inflectional morphemes including about 150 postpositions and about 500 verb endings (Kang, 2006). Unlike English, Korean usually has multiple morphemes per word.

Words	Morphemes with POS
나는 "I"	나/NP (pronoun) 는/JKS (nominative postposition)
친구들을 "friends"	친구/NNG (general noun) 들/XSN (noun derived suffix) 을/JKO (accusative postposition)
만났다 "met"	만나/VV (verb) 었/EP (prefinal ending) 다/EF (final ending)

Table 1: A Korean sentence split into three words and eight morphemes.

In Korean, words split by a space are called *eojeols*, which are formed by joining a content morpheme (e.g., a noun or a verb stem) and functional morphemes (e.g., postpositions, suffixes and endings). Therefore, to process text and segment sentences into linguistic units some have relied on space segmentation and others have relied on morphological analysis. We discuss the pros and cons of both approaches.

2.1 Eojeol Segmentation

Initially, Petrov and colleagues (2012) provided an eojeol based mapping to the UPOS12 for the 187 Sejong POS patterns derived from the Sejong treebank. McDonald and colleagues (2013) then provided the Google UD Treebanks (GUD) representing 6K sentences [8] applying the UPOS12 for Korean and five other European languages (German, English, Swedish, Spanish and French) in eojeol (token) segmentation. They identified Korean as an outlier in the cross-lingual transfer parsing evaluation. The Google UD Korean Treebank (GKT) [9]was then *automatically* converted to the UPOS14 following the UDv2 guidelines (UD annotation guidelines v2, 2016). To alleviate the data sparsity issue caused by a coarse-grained eojeol segmentation, Chun and colleagues (2018) added morpheme segmented Sejong POS tags obtained from the KOMA tagger[10] in the GKT. Although both the UPOS14 tags (based on the eojeol segmentation) and the Sejong POS tags (based on the morpheme segmentation) are included in the GKT, they remain largely separate tags because they cannot be automatically linked to each other due to their differences on the segmentation granularity.

To enhance the eojeol based UPOS14 mapping to Korean predicates, Noh and colleagues (2018) utilised morphological analysis results to clarify the roles of Korean predicates.

One limitation observed with eojeol segmentation is that the number of POS patterns (i.e. combination of morphemes within an eojeol) increases exponentially as the number of eojeols increases. As mentioned by Park and Tyers (2019), with about 450K eojeols in the Sejong treebank, almost 5K POS patterns were found and with about 9.2M eojeols in the Sejong corpus, over 50K POS patterns were found.

Furthermore, the treatment of nouns with postpositions has proven to be difficult with eojeol segmentation. As shown in Figure 1, without morphological analysis, it is very hard to distinguish whether '-**가**', '-**에게**', and '-**을**' are postpositions or a part of (proper) nouns. Korean postpositions can be used in a flexible manner without affecting the meaning of the sentence. For example, as shown in Figure 2, the meaning of the four sentences does not change regardless of the postpositions' usage. This can be problematic for NER task which aims to extract named entities, in particular, proper nouns.

민우+**가**	소라+**에게**	선물+**을**	주었다.
minu+ga	sola+ege	seonmul+eul	ju-eoss-da.
"Minwoo"	"to Sora"	"a present"	"gave."

Figure 1: An example of *noun* + **postposition** eojeols

S1: 나+**는**	너+**를**	사랑한다.
S2: 나	너+**를**	사랑한다.
S3: 나+**는**	너	사랑한다.
S4: 나	너	사랑한다.
"I"	"you"	"love"

Figure 2: An example of **postpositions** flexible usage

2.2 Morpheme Segmentation

Many Korean NLP applications are based on morpheme segmentation including phrase-structure parsing (Choi et al., 2012; Park et al., 2016) and statistical machine translation (Park et al., 2016; Park et al., 2017). For Korean word embedding models, morpheme segmentation is usually used to avoid expensive computational cost caused by the exponential increase in new vocabularies with eojeol segmentation (Lee, 2019). It is more efficient to segment words into morphemes to cover all the possible inflected verb forms.

Compared to eojeol segmentation, morpheme segmentation involves a lot more work as it requires a morphological analysis. But it presents a number of advantages, like the ability to conduct fine-grained segmentation to facilitate a numbers of NLP tasks. Fortuitously, there are a handful of open-source Korean morphological analysers available. For example, KoNLPy, a python-based wrapper, provides an easy access to five widely used Korean taggers including KKMA, KOMORAN, MECAB-KO, HANNANUM and OKT.

It is essential to segment Korean text into morphemes to understand the function of each morpheme, especially for the Part-of-Speech (POS) task in NLP. Morpheme segmentation also alleviates the issues of data sparsity and postposition analysis. Therefore, we chose this segmentation approach for the POS mapping proposed.

[8] derived from news articles and web blogs

[9] GKT is a subset of the GUD for Korean.

[10] A general-purpose morphological analyser for Korean (Lee and Rim, 2009) that produces the Sejong POS tag set.

3. Mapping Sejong POS to UPOS

To develop a mapping scheme for the Sejong POS tag set to the UPOS14, we started by reviewing the most recent mapping by Park and Tyers (2019) with the up-to-date UPOS definitions (UD annotation guidelines v2, 2016). Out of 45 Sejong POS tags, we revised 12 Sejong POS tag mapping while agreeing on the remaining 33.

As Park and Tyers (2019) did not discuss or provide explanations about their mapping, it was hard to compare the two mapping schemes. In the following subsections, we present and discuss the proposed changes. We also present relevant examples derived from the Google Korean UD Treebank (GKT) corpus[11] for the proposed mapping on the 12 Sejong POS tags.

Herewith UPOS denotes the UPOS14 for simplicity.

3.1 Sejong SL and SH tags

The Sejong SL tag denotes words written in foreign languages and the Sejong SH tag denotes words written in Chinese characters.

In the initial Sejong POS mapping to the UPOS12 (Park et al., 2016), the SL and SH tags were mapped to the UPOS X tag, which was defined as '*a catch-all for other categories such as abbreviations or foreign words* (Petrov et al., 2012)'. Then, when the UPOS14 was proposed, the definition of the X tag changed completely and was restricted to '*words that are not possible to analyse'*[12] (UD annotation guidelines v1, 2014). Despite the changed definition of the X tag, in Park and Tyers' work (2019), the SL and SH tags were still mapped to the X tag. Therefore, we proposed a new mapping for these two tags based on their actual usage and the new X tag definition.

We analysed the SL and SH tagged words found in the Google UD Korean Treebank (GKT) corpus. We manually examined all the instances of the SL and SH tagged words from a subset of the GKT corpus (development corpus), which consists of 950 sentences. We found 185 SL tags across 99 sentences; among them 136 were assigned to words used as proper nouns (e.g. 'KIA', 'SK' and 'LG전자'), 29 were assigned to words used as general nouns (e.g., 'TV' and 'CEO'), and 20 were incorrectly[13] assigned to words expressing units (e.g. 'km' and 'kg'). We also found 27 SH tags across 17 sentences; among them 22 SH tags were assigned to words used as proper nouns (e.g. '李小龍' – 'Bruce Lee') and 5 SH tags were assigned to words used as general nouns (e.g.'역사(力士)' – 'a man of great strength').

As both tags were mainly used as proper nouns (73% of SL tag and 78% of SH tag) and in some cases as general nouns (27% of SL tag and 22% of SH tag), we propose to map them to the UPOS PROPN tag as shown in Table 2.

Sejong POS	UPOS14	
	Previous	New
SL (Foreign characters)	X	PROPN
SH (Chinese characters)	X	PROPN

Table 2: Proposed mapping for the SL and SH tags

While somewhat imperfect, this new mapping allows us to preserve some information. We will discuss the impact of this new mapping on the Korean NER task in the Discussion section.

3.2 Sejong JC and EC tags

In Park and Tyers' work (2019), all the Sejong POS tags for postpositions including the JC tag (conjunctive particle) were mapped to the UPOS ADP tag (adposition). This tag represents prepositions and postpositions and is defined as '*a complement composed of a noun phrase, noun, pronoun or clause that functions as a noun phrase to express its grammatical and semantic relation to another unit within a clause'* (UD annotation guidelines v1, 2014).

However, as discussed by Ahn and Song (2011), in Korean, coordinate conjunction for noun is expressed by conjunctive particles. For example, in (1)

'메리+**와** 탐' (1)
(Mary **and** Tom)

the conjunctive particle '-**와**' is used to link the two proper nouns 'Mary' and 'Tom. Following Ahn and Song's analysis, we propose to remap the JC tag to the UPOS CCONJ tag as shown in Table 3.

The Sejong EC tag represents all three types of conjunctive endings: *equal conjunctive ending*, *subordinate conjunctive ending* and *auxiliary conjunctive ending*. Each of them has a corresponding UPOS tag: equal conjunctive ending could be mapped to the CCONJ tag (coordinate conjunction), subordinate conjunctive ending to the SCONJ tag (subordinate conjunction) and auxiliary conjunctive ending to the PART tag (particle). Since we have only one Sejong tag (EC) to represent them all, we need to choose the most appropriate one.

Sejong POS	UPOS14	
	Previous	New
JC (Conjunctive particle)	ADP	CCONJ
EC (Conjunctive ending)	PART	CCONJ/ SCONJ

Table 3: Proposed mapping for the JC and EC tags

Park and Tyers (2019) mapped all the Sejong POS tags for verb endings, including the EC tag (conjunctive ending), to the UPOS PART tag. The UPOS PART tag is defined as '*function words that must be associated with another word or phrase to impart meaning and that do not satisfy definitions of other universal parts of speech such as adpositions, coordinate conjunctions, subordinate conjunctions or auxiliary verbs. In general, the PART tag should be used restrictively and only when no other tag is possible'* (UD annotation guidelines v2, 2016, we underlined the text in this quote). Based on the UPOS

[11] The GKT corpus consists of three parts for development, training and testing. It contains a total of 6339 sentences, 80,322 eojeols and 151,737 morphemes.

[12] '*those ordinary loan words should be assigned a normal POS tag and X tag should be used very restrictively for*' (UD annotation guidelines v1, 2014)

[13] In the GKT corpus, units are tagged inconsistently with the Sejong SW (symbol) and SL tags even within a sentence. Thus, we considered them as incorrectly tagged words.

PART definition, we argue that the use of the UPOS CCONJ and SCONJ tags would be more appropriate for the EC tag.

As mentioned in Ahn and Song's work (2011), a coordinate conjunction for verb is expressed by equal conjunctive endings. For example, in (2)

'메리+는 자+**고** 탐+은 노래+한다.' (2)
(Mary sleeps **and** Tom sings.)

the equal conjunctive ending '**-고** (and)' is used to link the two verbs 'sleeps' and 'sings'. Similarly, a subordinate conjunction is expressed by subordinate conjunctive endings. For example, in (3)

'메리+가 아파+**서** …' (3)
(**because** Mary is sick…)

the subordinate conjunctive ending '**-서** (because)' is used to form a subordinate clause.

Both the CCONJ and SCONJ tags seem to be valid mapping for the EC tag. Choosing one over another will result in losing some information. Therefore, we included each mapping in our summary table (Table 8). The choice should be made based on the corpus and the downstream NLP task.

It is worth noting that the KKMA Korean POS tagger outputs 56 POS tags based on the 46 Sejong POS tags[14]. The additional POS tags include the conjunctive endings: ECE (equal conjunctive ending), ECD (subordinate conjunctive ending) and ECS (auxiliary conjunctive ending) allowing a 1-to-1 mapping to the UPOS CCONJ, SCONJ and PART tags.

3.3 Sejong NNB and NR tags

The NNB tag represents both unit bound nouns[15] and bound nouns. Unit bound nouns express the unit of quantity and bound nouns are special function words that must be associated with another word or phrase to impart meaning. Thus, the NNB tag denotes *dependent* nouns.

In the initial Sejong POS mapping to the UPOS12 (Park et al., 2016), all noun related Sejong POS tags, including the NNB tag, were mapped to the UPOS NOUN tag. In the UPOS14, the definition of the NOUN tag was clarified from simply noted as '*Nouns*' in the UPOS12 to '*Nouns are a part of speech typically denoting a person, place, thing, animal or idea. The NOUN tag is intended for common nouns only'* (UD annotation guidelines v1, 2014). However, despite this change, the Sejong NNB was still mapped to the UPOS NOUN tag in the recently updated mapping proposed by Park and Tyers (2019).

We, therefore, propose to remap the Sejong NNB tag to the UPOS ADP tag in line with ADP's definition '*a complement composed of a noun phrase, noun, pronoun, or clause that functions as a noun phrase, and that form a single structure with the complement to express its grammatical and semantic relation to another unit within a clause'* (UD annotation guidelines v1, 2014, we underlined the text in this quote).

Similarly, the Sejong NR tag (numeral) represents numbers for quantity and sequence written in the Korean alphabet (*Hangeul*). In the UPOS14, the definition of the NUM tag was updated to '*cardinal numbers are covered by NUM whether they are expressed as words (four), digits (4) or Roman numerals (IV)*' (UD annotation guidelines v1, 2014). Despite the evident mapping between the NR tag and the UPOS NUM tag, Park and colleagues (2016) and Park & Tyers (2019) mapped the NR tag to the UPOS NOUN tag. We believe it is an oversight and changes should be made to map correctly the NR tag to the UPOS NUM tag. This would allow identifying the numeral value correctly as illustrated by the example below. As shown in Table 4, in '2905**만**5000배럴', '**-만**' means 10,000. If the NR tag was mapped to the UPOS NOUN tag, the numeral value '29,050,000' could not be captured correctly.

	Sejong POS	UPOS14	
		Previous	New
2905	SN	NUM	NUM
만	NR	NOUN	NUM
5000	SN	NUM	NUM
배럴	NNB	NOUN	ADP

Table 4: Example of NR and NNB usage ('-배럴' means 'barrel', a unit for petrol)

Remapping the NNB tag to the UPOS ADP tag indicates that dependent nouns can be distinguished from general nouns and treated as functional words rather than content words. Remapping the NR tag to the UPOS NUM tag allows numeral values written in Korean characters to be correctly treated as numeral values rather than as nouns.

Table 5 below summarises the proposed changes.

Sejong POS	UPOS14	
	Previous	New
NNB (Bound noun)	NOUN	ADP
NR (Numeral)	NOUN	NUM

Table 5: Proposed mapping for the NNB and NR tags

3.4 Sejong VX, VCP, VCN and EP tags

In the initial Sejong POS mapping to the UPOS, Park and colleagues (2016) mapped the Sejong VX (auxiliary verb), VCP (positive copular) and VCN (negative copular) tags to the UPOS VERB tag as it was the only UPOS tag available for verb related tags.

In the revised UPOS14, the AUX tag was introduced with the definition of '*an auxiliary is a function word that accompanies the lexical verb of a verb phrase and expresses grammatical distinctions not carried by the lexical verb, such as person, number, tense, mood, aspect, voice or evidentiality. The class AUX includes copulas (in the narrow sense of pure linking words for nonverbal predication)'* (UD annotation guidelines v2, 2016, we underlined the text in this quote). However, in the recent mapping by Park and Tyers (2019), the introduction of the AUX tag was not taken into account and the Sejong VX, VCP and VCN tags were still mapped to the UPOS VERB tag. We suggest taking advantage of this new UPOS tag.

[14] The mappings between the Sejong POS tag set and the KKMA POS tag set – http://kkma.snu.ac.kr/documents/?doc=postag

[15] There is no English equivalent to the unit bound noun (NNB) while it is well developed in Korean, Japanese, Chinese and Thai (Ahn and Song, 2011).

Remapping the Sejong VX, VCP and VCN tags to the AUX tag allow us to distinguish auxiliary verbs from main verbs (see Table 6 below).

Sejong POS	UPOS14	
	Previous	New
VX (Auxiliary verb)	VERB	AUX
VCP (Positive copular)	VERB	AUX
VCN (Negative copular)	VERB	AUX
EP (Prefinal ending)	PART	AUX

Table 6: Proposed mapping for the VX, VCP, VCN and EP tags

The Sejong EP tag represents prefinal ending, which precedes verb ending *to express tense or honorification*. This tag was previously mapped to the UPOS PART tag, which *should be used restrictively only when no other tag is suitable* according to its definition (described in the section 3.2). We propose to remap the Sejong EP tag to the UPOS AUX tag as it better satisfy the definitions of the UPOS AUX and PART tags.

3.5 Sejong MAJ and SW tags

The Sejong MAJ tag represents conjunctive adverbs that join two complete sentences. Conjunctive adverbs are generally located in front of a sentence, which it is modifying. For example, see (4).

'지구는 돈다.'
(The Earth spins.)

(4)

'**그러나/MAJ**, 아무도 그것을 믿지 않았다.'
(**However**, nobody believed it.)

As mentioned earlier, in Korean, coordinate and subordinate conjunctions are expressed by postpositions and verb endings *within a sentence*. Korean conjunctive adverbs such as **그리고** (additionally), **그러나** (however), **그래서** (therefore) *go over the boundary of a sentence*; and therefore should be seen as adverbs like in English (Ahn and Song, 2011). As shown in Table 7, we propose to remap the MAJ tag to the UPOS ADV tag.

The Sejong SW tag denotes symbols such as maths symbols and currency symbols. In the initial mapping scheme by Park and colleagues (2016), the SW tag was mapped to the UPOS X tag as there was no UPOS tag available for symbols in the UPOS12.

The UPOS SYM tag was introduced in the revised UPOS14 to cover symbols. However, when Park and Tyers revised the Sejong POS mapping to the UPOS (2019), they did not make use of this new tag and kept the initial mapping to the UPOS X tag. Therefore, we propose to remap the Sejong SW tag to the UPOS SYM tag as shown in Table 7. This revision allows symbols to be tagged correctly rather than to be treated as unanalysable characters.

Sejong POS	UPOS14	
	Previous	New
MAJ (Conjunctive adverb)	CCONJ	ADV
SW (Symbols)	X	SYM

Table 7: Proposed mapping for the MAJ and SW tags

3.6 Mapping Results

Table 8 summarises all the changes we propose (highlighted in bold font). These changes affected about one third of the existing mappings. We believe that this revision takes full advantages of the up-to-date UPOS definitions, the newly introduced UPOS tags and also takes into account linguistic usage. As discussed earlier, the EC tag has been assigned to both the CCONJ and SCONJ tags. Our updated mapping now includes the UPOS AUX and SYM tags, not previously considered.

Sejong POS	UPOS
VA	ADJ (adjective)
MAG, **MAJ**	ADV (adverb)
IC	INTJ (interjection)
NNG, XR	NOUN (noun)
NNP, **SL**, **SH**	PROPN (proper noun)
VV	VERB (verb)
NNB, JKS, JKC, JKG, JKO, JKB, JKV, JKQ, JX	ADP (adposition)
VX, VCP, VCN, EP	AUX (auxiliary)
JC, EC	CCONJ (coordinate conjunction)
MM	DET (determiner)
NR, SN	NUM (numeral)
EF, ETN, ETM, XPN, XSN, XSA, XSV	PART (particle)
NP	PRON (pronoun)
EC	SCONJ (subordinate conjunction)
SF, SP, SE, SO, SS	PUNCT (punctuation)
SW	SYM (symbol)
NA, NF, NV	X (other)

Table 8: Mapping the 45 Sejong POS tags to the UPOS

4. KAIST POS to UPOS

We also developed a mapping for the KAIST POS tag set, another widely used Korean POS tag set, to the UPOS. Unlike the Sejong POS tag set, the KAIST POS tag set is organised into a hierarchy comprised of four levels of tag granularity: 9, 22, 26 and 69 tags. We chose to work with the most fine-grained 69 tags as this set matches the most closely to the Sejong 45 POS tags.

To map the KAIST POS to the UPOS, we first reviewed the Korean POS tags comparison chart[16] between the Sejong POS and the KAIST POS to identify differences. Where 1-to-1 mapping between the Sejong POS and the KAIST POS was identified, the same UPOS tag was mapped to both tag set. This represented 22 KAIST POS tags out of the 69 tags.

Where 1-to-many mapping was identified (i.e., one Sejong POS tag corresponds to several KAIST POS tags), we reviewed the mapping to the UPOS. In most cases (39 out of 42), the same UPOS was mapped to both tag sets. For example, the Sejong NNP (proper noun) tag corresponds to four KAIST tags: NQPA (family name, e.g., '문'), NQPB (given name, e.g., '재인'), NQPC (family+given name, e.g., '문재인'), and NQQ (general

[16] https://konlpy-ko.readthedocs.io/ko/v0.4.3/morph/#comparison-between-pos-tagging-classes

proper noun, e.g., 대한민국). These four KAIST tags were mapped to the UPOS PROPN tag as was the Sejong NNP tag.

We examined more closely the three remaining cases. As mentioned in the section 3.2, the Sejong EC tag (conjunctive ending) covers three types of conjunctive endings (i.e., equal conjunctive ending, subordinative conjunctive ending and auxiliary conjunctive ending). In contrast, the KAIST POS tag set offers a specific tag for each of these conjunctive endings: ECQ (equal conjunctive ending), ECS (subordinative conjunctive ending) and ECX (auxiliary conjunctive ending). Therefore, we propose to map these KAIST POS tags with their corresponding tags in the UPOS: CCONJ, SCONJ and PART respectively. It is worth noting here that the mapping for the different types of conjunctive endings differs between Sejong and KAIST. It seems that the KAIST POS tags better aligns with the UPOS enabling the mapping to preserve as much information as possible.

Finally, we identified five cases (MAD, JGT, JP, XSAM and XSAS) where there were no correspondences between the Sejong and the KAIST POS. The KAIST POS tag set offers more detailed categories that are not represented in the Sejong POS tag set.

Taking into account the UPOS definitions and the KAIST POS guideline (Choi et al., 1994; "KAIST POS Tag Set", 2003), we mapped the MAD tag (demonstrative adverbs) to the UPOS ADV tag; the JGT tag (joint case postposition) to the UPOS ADP tag; the JP tag (predicative marker) to the UPOS ADP tag; the XSAM and XSAS tags (both adverb derived suffixes) to the UPOS PART tag.

Table 9 shows the new mapping of the KAIST POS to the UPOS. Note that none of the KAIST POS tags were mapped to the UPOS X tag.

KAIST POS	UPOS
PAD, PAA	ADJ
MAG, MAJ, MAD	ADV
II	INTJ
NCPA, NCPS, NCN, NCR	NOUN
NQPA, NQPB, NQPC, NQQ, F	PROPN
PVD, PVG	VERB
NBN, NBS, NBU, JCS, JCC, JCM, JCO, JCA, JCV, JCR, JXC, JXF, JGT, JP	ADP
PX, EP	AUX
JCJ, ECC	CCONJ
MMD, MMA	DET
NNC, NNO	NUM
EF, ECX, ETN, ETM, XP, XSNU, XSNCA, XSNCC, XSNA, XSNS, XSNX, XSVV, XSVN, XSVA, XSMS, XSMN, XSAM, XSAS	PART
NPP, NPD	PRON
ECS	SCONJ
SF, SE, SP, SD, SL, SR	PUNCT
SY, SU	SYM
N/A	*X*

Table 9: Mapping the 69 KAIST POS tags to the UPOS

5. Discussion

Now that we have updated the Sejong POS mapping to the UPOS and proposed one for the KAIST POS, we discuss the impact this have on Korean NER task. We did not compare parser performances but instead we propose to illustrate the impact of this new UPOS mapping on NER task through a few examples. Our aim is to demonstrate the gain this new mapping enables by better identifying complex named entities, irrespective of individual NER system's performances.

The most important change we made is to reduce the unnecessary use of the UPOS X tag, which is reserved for unanalysable words. We discuss below the impact of this change for the Sejong POS SL and SW tags.

The examination of the SL tag usage showed that people can refer to entities both in Korean or their foreign correspondences. For example, the organisation named entity *Samsung* could be referred to by its English name ('Samsung') or by its Korean name ('삼성'). Under the previous mapping, all foreign words were tagged with the UPOS X tag and would not have been recognised as proper nouns, and therefore, as named entities. Remapping the SL tag to the UPOS PROPN tag allows us to recover named entities that would have been ignored.

Similarly, for more complex named entities (e.g., 'Samsung SDI'), we want to be able to capture the entity as a whole, irrespective of the language within which it is expressed (e.g., Korean, foreign or a mix of the two). As shown in Table 10, under the previous mapping, only entities expressed in Korean would have been identified. Again, remapping the SL tag to the UPOS PROPN tag allows us to recover named entities expressed in foreign languages (partly or in full).

Similarly, for more complex named entities (e.g., 'Samsung SDI'), we want to be able to capture the entity as a whole, irrespective of the language within which it is expressed (e.g., Korean, foreign or a mix of the two). As shown in Figure 4, under the previous mapping, only entities expressed in Korean would have been identified. Again, remapping the SL tag to the UPOS PROPN tag allows us to recover named entities expressed in foreign languages (partly or in full).

Currency symbol plays an important role in the identification of references to monetary amount in text. For example, in '$500' ($/SW+500/SN), the numeral value '500' is recognised as a currency value as it is preceded by a currency symbol '$'. Under the previous mapping, symbols were tagged with the UPOS X tag, thus it would have been difficult to recognise currency symbols. This would have resulted in missing money named entities or capturing it without currency symbols. Remapping the SW tag to the UPOS SYM tag allows us to recognise '500' as a monetary value, capture '$500' as a whole, and identify it as a MONEY named entity.

As suggested by Park and Tyers (2019), the mapping for the Sejong POS tag set to the UPOS tag set can be applied directly to the Sejong corpus to build UPOS tagged training corpus for downstream NLP tasks such as NER. The same can be done to the KAIST corpus by using our new mapping for the KAIST POS tag set. Moreover, to develop new corpus for Korean, one could also use one of the Korean POS taggers tailored to either Sejong or KAIST POS tags and then convert the output to the UPOS using the new mapping scheme we propose.

	Sejong POS	Previous UPOS	NE	New UPOS	NE
Samsung	SL	**X**	**O**	PROPN	ORG
삼성	NNP	PROPN	ORG	PROPN	ORG
Samsung SDI	SL SL	**X X**	**O O**	PROPN PROPN	ORG ORG
삼성 에스디아이	NNP NNP	PROPN PROPN	ORG ORG	PROPN PROPN	ORG ORG
삼성 SDI	NNP SL	PROPN **X**	ORG **O**	PROPN PROPN	ORG ORG

Table 10: Mapping of the SL tag to the UPOS and NE extracted (for the entity 'Samsung' and 'Samsung SDI' written in English, Korean and both)

The mapping of the Korean POS to the UPOS can also be used to compare Korean parser performances. Although the KoNLPy[17] provides an easy access to various Korean parsers, it is not easy to directly compare their performance as they output different Korean POS tags (i.e., Sejong POS or KAIST POS).

6. Conclusion

When mapping a fine-grained tag set to a coarse-grained tag set, information loss is inevitable. While the UPOS does not capture as much information as specific Korean POS tag sets, it is a useful representation when working with a multi-lingual system. This paper proposes a new mapping for Korean POS tag set to the Universal POS tag set (UPOS), for the two most widely used Korean POS tag sets (the Sejong POS tag set and the KAIST POS tag set). It also discusses segmentation issues, provides explanation about the proposed mapping, and the impact this has on Korean NER task through qualitative analysis.

7. Bibliographical References

Ahn J. and Song K. (2011). A Contrastive Study of Korean-English Word Classes in a Typological Perspective. The Linguistic Association Korea Journal, 19 (6): 213-232.

Brinton, L. J. (2000). The Structure of modern English: A linguistic introduction. John Benjamin Publishing Company. Amsterdam/Philadelphia.

Choi D., Park J., and Choi K. (2012). Korean Treebank Transformation for Parser Training. In Proceedings of the ACL2012 Joint Workshop on Statistical Parsing and Semantic Processing of Morphologically Rich Languages, pages 78-88, Jeju, Republic of Korea. Association for Computational Linguistics.

Choi K., Han Y.S., Han Y.G. and Kwon O. (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. Nara, Japan.

Chun J., Han N., Hwang J. D., Choi J. D. (2018). Building Universal Dependency Treebanks in Korean. In Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC'18), Miyazaki, Japan, 2018.

Eifring H. and Theil R. (2004). Linguistics for Students of Asian and African Languages. Institutt for østeuropeiske og orientalske studier.

Kang S. (2006). 형태소 분석의 이해 [Understanding morphological analysis] [PowerPoint slides] Kookmin University. Retrieved from http://cafe.daum.net/nlpk/7dLs/5.

Lee G. (2019). Sentence Embedding Using Korean Corpora, Acorn Publisher, Seoul, 1st edition.

McDonald R., Nivre J., bach-Brundage Y. Q., Goldberg Y., Das D., Ganchev K., Hall K., Petrov S., Zhang H., Tackstrom O., Bedini C., Castello N. B., Lee J. (2013). Universal Dependency Annotation for Multilingual Parsing. In the Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 92-97.

Noh Y., Han J., Oh T. and Kim H. (2018). Enhancing Universal Dependencies for Korean. In Proceedings of the second Workshop on Universal Dependencies (UDW 2018), pages 108-116.

Park J. (2017). Segmentation Granularity in Dependency Representations for Korean. In Proceedings of the Fourth International Conference on Dependency Linguistics (Depling2017), pages 187-196.

Park J., Dugast L., Hong J., Shin C. and Cha J. (2017). Building a Better Bitext for Structurally Different Languages through Self-training. In Proceedings of the First Workshop on Curation and Applications of Parallel and Comparable Corpora, pages 1-10, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Park J., Hong J. and Cha J. (2016) Korean Language Resources for Everyone. In the Proceedings of PACLIC 2016, the 30th Pacific Asia Conference on Language, Information and Computation, pages 49-58.

Park J. and Tyers F. (2019). A New Annotation Scheme for the Sejong Part-of-Speech Tagged Corpus, In Proceedings of the 13th Linguistic Annotation Workshop, pages 195-202.

Petrov S., Das D. and McDonald R. (2012). A Universal Part-of-Speech Tagset. In Proceedings of LREC 2012, the 8th International Conference on Language Resources and Evaluation.

Rosa R., Mašek J., Mareček D., Popel M., Zeman D., Žabokrtský Z. (2014). HamleDT 2.0: Thirty Dependency Treebanks Stanfordized. In Proceedings of LREC. pages 2334-2341.

Yule, G. (2010). The Study of Language. Cambridge University Press, Cambridge, Fourth edition, pages 69-70

"KAIST POS Tag Set", (10 Jul 2003) Retrieved from http://semanticweb.kaist.ac.kr/research/morph/ (Accessed: 5 Nov 2019).

"UD annotation guidelines v1", (2014-10-01), Universal Dependencies, Retrieved from

[17] https://konlpy-ko.readthedocs.io/ko/v0.4.3/

http://universaldependencies.org/docsv1/index.html (Accessed: 10 Oct 2019).

"UD annotation guidelines v2", (01 Dec 2016), Universal Dependencies, Retrieved from http://universaldependencies.org/guidelines.html (Accessed: 24 Oct 2019).