# Coordinate Structures in Universal Dependencies for Head-final Languages

**Hiroshi Kanayama**
IBM Research
Tokyo, Japan
hkana@jp.ibm.com

**Na-Rae Han**
University of Pittsburgh
Pittsburgh, PA
naraehan@pitt.edu

**Masayuki Asahara**
NINJAL, Japan
Tokyo, Japan
masayu-a@ninjal.ac.jp

**Jena D. Hwang**
IHMC
Ocala, FL
jhwang@ihmc.us

**Yusuke Miyao**
University of Tokyo
Tokyo, Japan
yusuke@is.s.u-tokyo.ac.jp

**Jinho Choi**
Emory University
Atlanta, GA
jinho.choi@emory.edu

**Yuji Matsumoto**
Nara Institute of Science and Technology
Nara, Japan
matsu@naist.jp

## Abstract

This paper discusses the representation of coordinate structures in the Universal Dependencies framework for two head-final languages, Japanese and Korean. UD applies a strict principle that makes the head of coordination the left-most conjunct. However, the guideline may produce syntactic trees which are difficult to accept in head-final languages. This paper describes the status in the current Japanese and Korean corpora and proposes alternative designs suitable for these languages.

## 1 Introduction

The Universal Dependencies (UD) (Nivre et al., 2016, 2017) is a worldwide project to provide multilingual syntactic resources of dependency structures with a uniformed tag set for all languages. The dependency structure in UD was originally designed based on the Universal Stanford Dependencies (De Marneffe et al., 2014), in which the left-most conjunct was selected as the head node in coordinate structures. After some modifications, the current UD (version 2) uses the definition as shown in Figure 1.

The UD principles include a simple mandate: the left word is always the head in parallel and sequential structures, including coordination, apposition and multi-word expressions. The rationale behind this uniformity is that these structures do not involve true dependency, and having a single direction for conj relations on the assumption that coordinate structures are completely paratactic, both within and across languages, is advantageous. However, as discussed in several proposal for extended representation of coordination structures (Gerdes and Kahane, 2015; Schuster and Manning, 2016), they cannot be straightforwardly represented as dependencies. Especially in head-final languages such as Japanese and Korean, the left-headed structure poses some fundamental issues due to hypotactic attributes in terms of syntax in coordinate structures.

This paper points out the issues in the treatment of coordinate structures with evidence of linguistic plausibility and the trainability of parsers, reports on the current status of the corpora in those languages, and proposes alternative representations.

Section 2 describes the linguistic features of head-final languages, and Section 3 points out the problems in the left-headed coordinate structures in head-final languages. Section 4 summarizes the current status of UD Japanese (Tanaka et al., 2016; Asahara et al., 2018) and UD Korean (Chun et al., 2018) corpora released as version 2.2. Section 5 shows the experimental results on multiple corpora in Japanese and Korean to attest the difficulty in training with left-headed coordination. Section 6 proposes a revision to the UD guidelines more suited to head-final languages.

## 2 Head-final languages

Both Japanese and Korean are strictly head-final agglutinative languages in which most dependencies between content words have the head in the

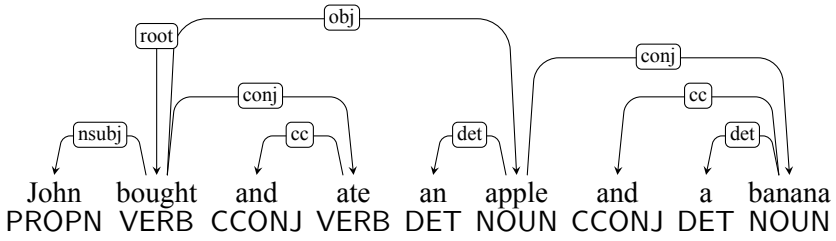



Figure 1: English coordinate structures ("bought and ate" and "an apple and a banana") in UD v2.

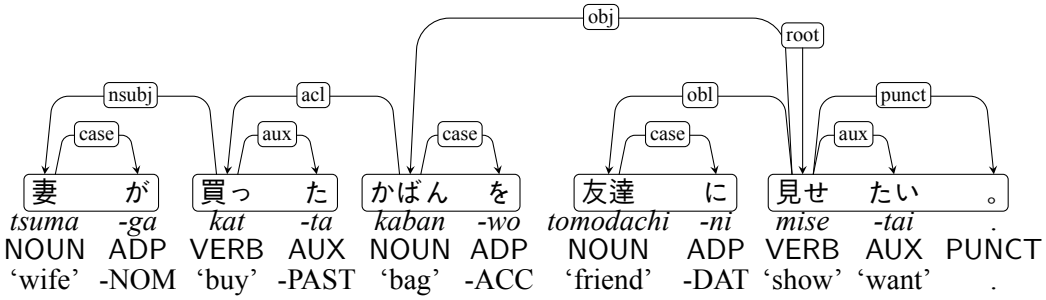



Figure 2: A head-final dependency structure of a Japanese sentence "妻が買ったかばんを友達に見せたい" ('(I) want to show the bag which (my) wife bought to (my) friend').

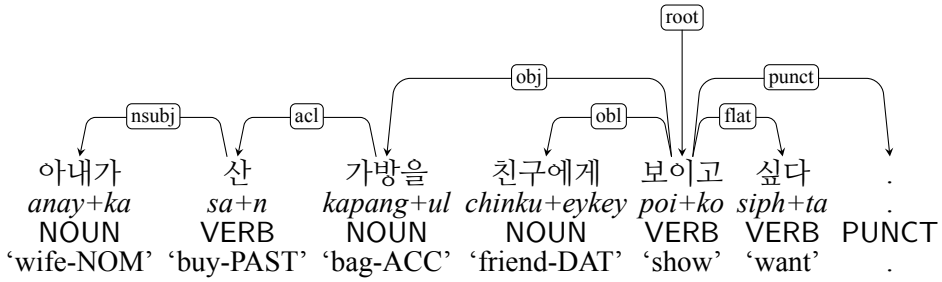



Figure 3: A head-final dependency structure of a Korean sentence "아내가 산 가방을 친구에게 보이고 싶다.", which is parallel to that in Figure 2.

right. Figures 2 and 3 depict the dependency structures in Universal Dependencies for Japanese and Korean sentences, respectively. Both have right-headed dependencies except for functional words and punctuations.

Japanese has a well-known phrasal unit, called *bunsetsu*—each unit is marked with a rounded rectangle in Figure 2. A *bunsetsu* consists of a content word (or multiple words in the case of a compound) and zero or more functional words such as postpositional case markers (ADP), particles (PART) and auxiliary verbs (AUX).

Korean has a similar unit called *eojeol*. It typically consists of a content word optionally followed by highly productive verbal or nominal suffixation, and, unlike Japanese *bunsetsu*, it is marked by white space in orthography. Figure 3 shows a Korean counterpart to Figure 2, where the syntax and the main dependency relations mirror those of the Japanese example. The main departure here is that the Korean UD's treatment of postposition suffixes and verbal endings are dependent morphemes in the *eojeol*-based Korean orthography, and thus, are neither tokenized nor assigned separate dependency relations.

UD corpora from both languages are converted from dependency or constituency corpora based on *bunsetsu* or *eojeol* units. In Japanese, functional words in each *bunsetsu* (ADP, AUX and PUNCT in Figure 2) must depend on the head word in the *bunsetsu* (NOUN and VERB). In the Korean example of Figure 3, the last verb "싶다" ('want') behaves as a function word though it is tagged as VERB, thus it is attached to the main verb with

flat label. As for the dependencies between content words, the right-hand unit is always the head. The exceptions are limited to special cases such as annotations using parentheses, but when the UD's left-headedness principle is adopted, multi-word expressions and coordination are added to exceptional cases.

In addition to these two languages, Tamil is categorized as a rigid head-final language (Polinsky, 2012). According to the typological classification using statistics of UD corpora (Chen and Gerdes, 2017), Japanese and Korean fall into a similar class in terms of distance of dependencies. The same goes for Urdu and Hindi, but they have more flexibility in word order including predicates.

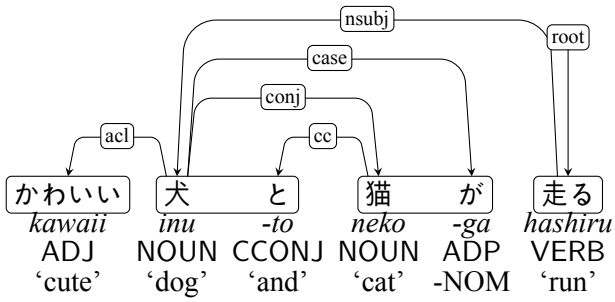



Figure 4: Left-headed representation of a nominal coordination in Japanese "犬と猫" ('dog and cat'), in a sentence "かわいい犬と猫が走る" ('A cute dog and cat run').

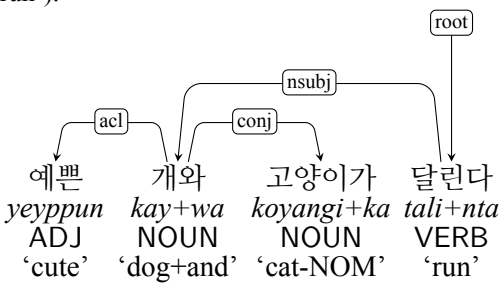



Figure 5: Left-headed representation of a nominal coordination in Korean "개와 고양이" ('dog and cat'), in a sentence "예쁜 개와 고양이가 달린다" ('A cute dog and cat run').

## 3 Issues with left-headed coordination

This section points out several issues regarding Japanese and Korean coordinate structures in Universal Dependencies when the left-headed rules are strictly applied.

### 3.1 Nominal coordination

If a Japanese noun phrase "犬と猫" ('dog and cat') is regarded as a coordination and represented in a left-headed manner under UD, the structure is as Figure 4 in a sentence "犬と猫が走る" ('A cute dog and cat run'). When the particle "と" (*to*) is regarded as a conjunction CCONJ to connect two conjuncts, instead of a case marker attached to the preceding noun "犬" ('dog'), it is made a dependent of the right conjunct, breaking the *bunsetsu* unit in the dependency structure.

Also the nominative case marker "が" (*ga*) following "猫" ('cat') should specify the nominative case of the noun phrase ('dog and cat'), then the case marker is a child of "犬" ('dog') as the left conjunct, which produces a long distance dependency for a case marker which is usually attached to the preceding word.

The Korean counterpart in Figure 5 mirrors the Japanese example, except that again due to the different tokenization scheme the conjunctive particle "와" (*wa*) is kept suffixized in the left nominal conjunct *eojeol*, thus the conjunction relation cc is not overtly marked.

A common problem with adjectival modification in UD shown in Figures 4 and 5 is that there is no way to distinguish between modification of the full coordination vs. of the first conjunct (Przepiórkowski and Patejuk, 2018) . For example, there is no way to specify the scope of the adjective 'cute': the two readings (1) only a dog is cute and (2) both animals are cute.

### 3.2 Verbal coordination

Further critical issues are attested in the verbal coordinate structures. Figure 6 shows the left-headed verbal coordination "食べて走る" ('eat and run') in a noun phrase "食べて走る人" ('a person who eats and runs'), where verb "食べ" ('eat') is the child of "人" ('person'). Despite this dependency relationship, morphological markings tells us a different story: "食べ+て" is an adverbial form that modifies another verb, *i.e.*, "走る" ('run'), and the verb "走る" ('run') is an adnominal form that modifies another noun, *i.e.*, "人" ('person'). Therefore, the dependency between 'eat' and 'person' does not properly reflect the syntactic relationship of the modification of a verb by an adnominal verb, without seeing the whole coordinate structure 'eat and run'. The same set of issues are observed with the corresponding Korean example in Figure 7.

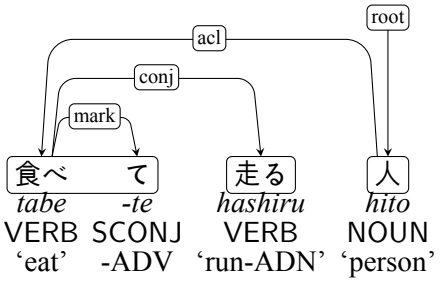



Figure 6: Left-headed representation of a verbal coordination in a Japanese phrase “食べて走る人” (‘A person who eats and runs’).

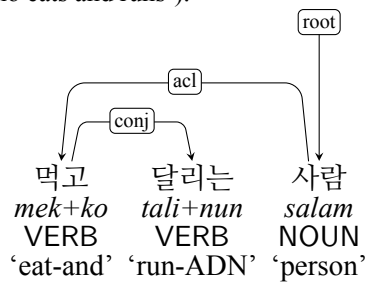



Figure 7: Left-headed representation of a verbal coordination in a Korean phrase “먹고 달리는 사람” (‘A person who eats and runs’).

### 3.3 Ellipsis

It is widely acknowledged that the phenomenon of ellipsis in non-constituent coordination is difficult to represent in UD, which does not allow introduction of covert gap words. Such structures can be even trickier to capture in head-final languages.

Figure 8 shows Japanese examples of non-constituent coordination. (a) is the coordination of “父は山に行き” (‘he goes to a mountain’) and “私は川に行った” (‘I went to a river’). The root node is the rightmost word in the left conjunct chunk. The second example (b) (‘My father went to the mountain, and I, to the river.’) shows the ellipsis of the first verb “行き” (‘go’), which is the root node in (a). The dependency relations of the omitted node that include the root are reduced and attached to the daughter node “父” (‘father’). The label orphan should be assigned between “私” (‘I’) and “山” (‘mountain’), and then, the first word, “父” (‘father’), becomes the root of the sentence. These peculiar tree constructions are caused by the left-headed principle of coordinate structures for a strictly head-final language, where the left conjunct tends to be omitted in this type of ellipsis. Korean likewise exhibits an exact parallel with its predicate ellipsis construction; examples are not shown in the interest of conserving space.

### 3.4 Coordination in Japanese and Korean: grammar vs. meaning

Conjunction is typically schematized as ‘X and Y’, where ‘X’ and ‘Y’ are interchangeable: *commutativity* is a defining characteristic of coordination which forms a basis for its headlessness. The Japanese and Korean examples presented so far, however, depart from this in a fundamental way: coordination in the two languages is asymmetric on the levels of syntax and morphology. Their ‘and’-counterpart is a dependent morpheme attached to the left conjunct,[1] and it is the right conjunct that bears all inflections and syntactic markings. In ellipsis, it’s the left conjunct that is reduced, while the right conjunct, along with requisite inflectional markings, is left standing.

This, then, points strongly towards the right conjunct being the head. Hoeksema (1992) cites four criteria of the ‘head’, which are: semantic, distributional, morphosyntactic, and technical (*i.e.*, phrasal projection); his morphosyntactic criterion states that the head is the locus of inflection, which applies to the right conjunct in the two languages.

On the other hand, there is one source of commutativity for Japanese and Korean coordination, which is *meaning*: namely, the fact that the lexical portions of left and right conjuncts can be swapped with no impact on truth conditions. In nominal coordination (4, 5) this semantic commutativity is robust; in verbal coordination (6, 7, 8), it is more restricted as temporal-sequential or causal interpretation often slips in (*e.g.*, 6, 7 could be understood as ‘eats *and then* runs’), but where it is available it tends to be just as robust (*e.g.*, 8). This would mean that the semantic commutativity is the primary basis for identifying and acknowledging coordination as a phenomenon in these languages, as this property does not extend to grammar.

Back to the grammatical aspect, a natural corollary is that Japanese and Korean coordinate structures are very close to those of nominal modification and subordination. In Korean, “존-의 고양이-가” (*John-uy koyangi-ka*, ‘John’s cat-NOM’) with the genitive marker “-의” (*-uy*) therefore appears to share the same configuration as ‘cat-and dog-NOM’; “먹-고서 달리-는 사람” (*mek-kose tali-nun salam*, ‘person who eats *and then* runs’)

[1] Exceptions exist: Korean and Japanese conjunction and disjunction markers “그리고”, “及び”, “및”, “또는”, “ないし” are whole words.

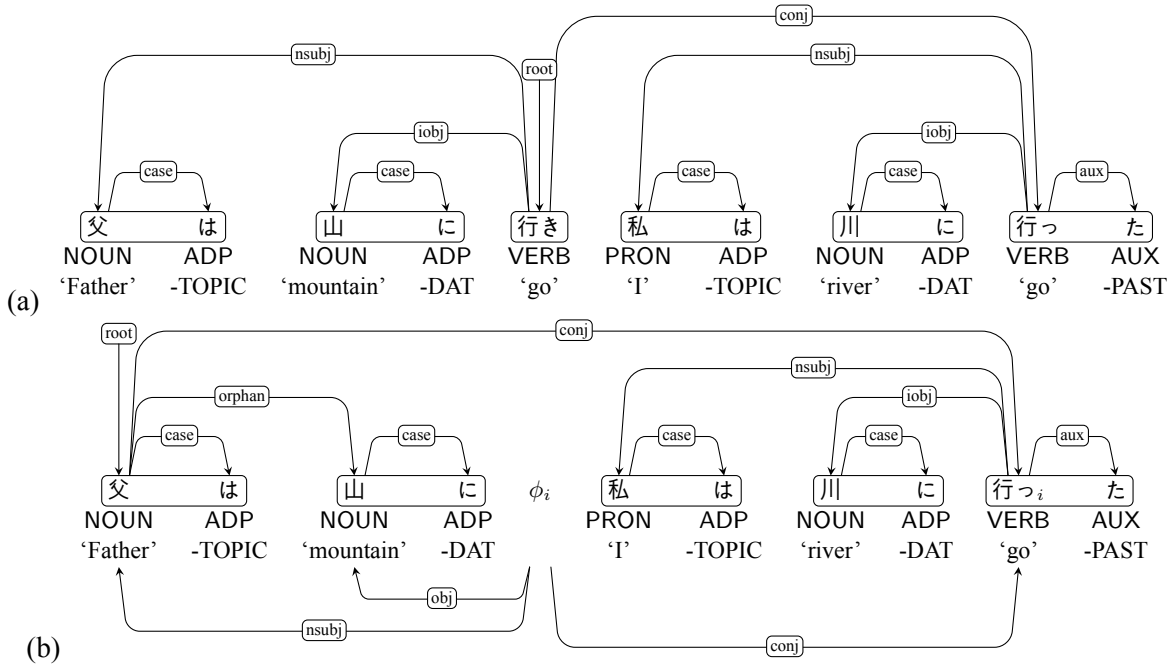



Figure 8: Predicate ellipsis in the non-constituent conjunct coordination.

with the sequential verbal ending “-고서” (*-kose*) likewise is indistinguishable on the surface from the coordination counterpart which uses “-고” (*-ko*, ‘and’) instead. In both cases, the righthand-side elements are unquestionably the head, syntactically and semantically, and they are treated as such in Japanese and Korean UD. Then, the only criteria for distinguishing the coordinate structures from their headed cousins are (1) choice of the suffix, and (2) semantic commutativity. One unfortunate consequence of the current UD principles is that these seemingly parallel pairs of structures in Korean and Japanese must receive vastly different syntactic treatments – one right-headed and the other left-headed – based on these two, non-syntactic, attributes. This creates a point of incongruence in terms of language-internal grammar; additionally, it becomes an engineering-side liability, as we will see shortly in Section 5.

## 4 Current status

Despite the complexities outlined in the previous section, the UD Japanese and UD Korean teams had to work within the bounds of the principles laid out by the Universal Dependencies version 2. Therefore, in the official version 2.2 release used for the CoNLL 2018 shared task (Zeman et al., 2018), UD Japanese and UD Korean adopted two separate strategies in order to ensure compliance, as we will see below.

### 4.1 UD Japanese

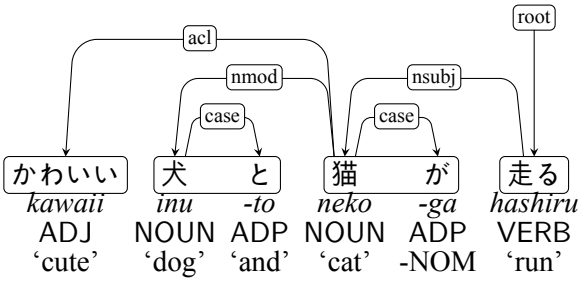



Figure 9: The representation in UD Japanese v2.2 for a sentence “かわいい犬と猫が走る” (‘A cute dog and cat run’).

To sidestep the issues described in Section 3, UD_Japanese-GSD and -BCCWJ opted against using coordinate structures altogether, that is, no conj label appears in the two corpora. Instead, nominal coordination is represented as a type of nominal modification (nmod) as shown in Figure 9, with “と” (*to*) between ‘dog’ and ‘cat’ categorized as ADP along with other case markers. This treatment simplifies the structure: the head of ‘cute’ is now ‘cat’, which clearly signals that the adjective modifies both ‘dog’ and ‘cat’. Moreover, ‘cat’, which is associated with the nominative case marker “が” (*ga*), is seen directly connected with

the verb ‘run’ with the (nsubj) label.

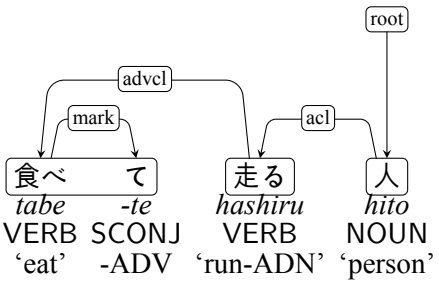



Figure 10: The representation in UD Japanese v2.2 for a phrase “食べて走る人” (‘A person who eats and runs’).

Additionally, the relationship between verbs are not handled as coordination, as shown in Figure 10. A verb connected with “て” (*te*) is regarded as subordination rather than coordination, because the phrase can be read as ‘a person who runs after eating’. It is consistent with the strategy of PoS tagging in UD Japanese to assign SCONJ for conjunctive particles.

Besides the coordination, UD Japanese does not use flat label for sequential nouns, including proper nouns, to avoid the left-headed structures. Instead, compound is used as shown in Figure 11.

UD Japanese_GSD uses fixed for a limited numbers of multi-word functional words, while UD Japanese_BCCWJ does not use it at all. Table 1 shows the distribution of some labels.

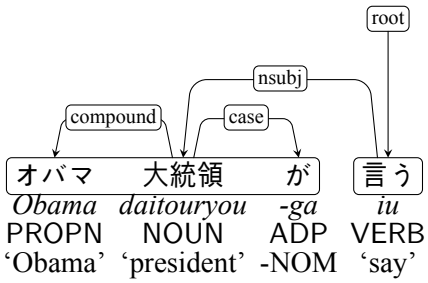



Figure 11: The use of compound in UD Japanese v2.2 for “オバマ大統領が言う” (‘President Obama says’).

| Corpus | root | conj | flat | fixed |
|---|---|---|---|---|
| GSD | 8,232 | 0 | 0 | 338 |
| BCCWJ | 57,256 | 0 | 0 | 0 |

Table 1: Distribution of labels in UD Japanese corpora. root shows the number of sentences.

### 4.2 UD Korean

Unlike the Japanese UD, the Korean UD effort has made a conscious decision to use right-headedness for conjunction following the coordination guidelines proposed by Choi and Palmer (2011). Thus, the coordinate structures in all three of the Korean UD corpora (Chun et al., 2018) were developed with the rightmost conjunct as the head of the phrase, with each conjunct pointing to its right sibling as its head.

For the latest available UD_Korean-GSD, however, the dependencies were converted to left-headed structures post-development in an effort to fully comply with the UD guidelines despite the problems left-headed structures pose for the language as described in Section 3. The other two Korean UD corpora, namely the Kaist and the Korean Penn Treebank, reflect right-headed coordinate structures (Chun et al., 2018). In addition to coordination, UD Korean extends the right-headed dependency structures to noun-noun structures. Unlike the Japanese that has opted to represent sequential nouns as cases of compound (Figure 11), Korean uses right-headed flat and fixed dependencies (Figure 12(a)), assigning the rightmost nominal with the morphological case marking as the phrasal head. Just as with the coordinate structure, these flat dependencies were converted into left-headed structures for the UD_Korean-GSD (Figure 12(b)). Table 2 shows the distributions of conj, flat and fixed labels.

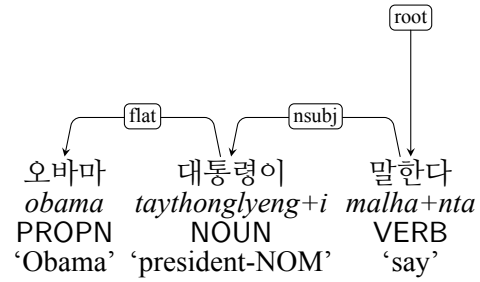
| Corpus | root | conj | flat | fixed |
|---|---|---|---|---|
| GSD | 6,339 | 3,864 | 12,247 | 13 |
| Kaist | 27,363 | 20,774 | 803 | 3,186 |
| Penn | 5,010 | 9,960 | 528 | 18 |

Table 2: Distribution of dependency labels in UD Korean corpora.

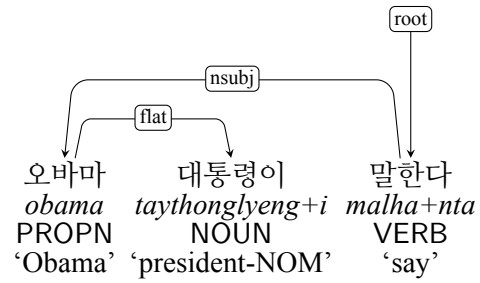
The differing strategies employed in the Japanese and Korean UD produce very different dependencies over structures that should otherwise receive similar analyses. Effectively, despite the syntactic similarities apparent in the two languages, the differences in the UD structures pose a challenge to the cross-lingual transfer learning (Kanayama et al., 2014).

## 5 Parsing Experiments

How well will parsers learn the syntactic structures of left-headed coordination in head-final lan-




(a) Korean right-headed flat structure.




(b) (a) converted to left-headed structure as reflected in the UD_Korean-GSD.

Figure 12: The use of flat in Korean UD v2.2.

guages? To answer this question, we trained and tested UDPipe (Straka et al., 2016) on multiple versions of UD Japanese and Korean corpora.

### 5.1 Japanese

As described in Section 4.1, the current UD Japanese-GSD corpus does not use conj tags. The corpus was converted into another version with coordinations without changing the dependency structures (right-headed coordination), that is, some of nmod and advcl labels are converted into conj label when the original manual annotation used conj regarding them as nominal or verbal coordinations. Also CCONJ tag and cc label are assigned to the coordinative case markers. The corpus was further converted into left-headed coordination, by changing the dependency structures following the UD guidelines.

For each corpora, two models were trained using `train` and `dev` portions; with (1) default UDPipe settings without changing any parameters, and (2) Japanese specific parameters for each phase[2] and precomputed word embeddings.

Given the model trained with each corpus and the raw input text of the `test` portion of corresponding corpus, UDPipe processed tokenization, PoS tagging and parsing. Table 3 shows the F1 values of tokenization (word), PoS tagging (UPOS) and UAS and LAS, for three models and two configurations. Tokenization is not straightforward because there is no whitespace between words, and it lowers scores of downstream processes; PoS tagging and parsing. Japanese specific configuration consistently showed better parsing scores by around 2 points.

Compared to the current UD Japanese ('no coordination'), 'right-head coordination' showed similar UAS values because the dependency relations were almost the same. In both configurations, LAS values dropped by 1.4 points because coordination (conj) cannot be deterministically distinguished from other dependencies (nmod or advcl). 'left-head coordination' further confused the model. UAS scores decreased by more than 3 points due to the difficulty to distinguish coordinate structures which completely change the dependency orientation, and the inconsistent syntactic relationship between the left conjunct and the head word. Also, it is known that shorter length of dependencies are preferred (Futrell et al., 2015) and the right-headed coordination strictly reduces the dependency distance in head-final languages. These results support the advantages of the right-headed strategy in Japanese coordinate structures.

### 5.2 Korean

All three UD corpora in Section 4.2, GSD, Kaist, and Penn Treebanks, are used to conduct similar experiments in Korean. First, raw text from those corpora are combined and fed into the original implementation of Word2Vec (Mikolov et al., 2013)

[2] `--tokenizer=dimension=64;epochs=100;initialization_range=0.1;batch_size=50;learning_rate=0.005;dropout=0.3;early_stopping=1 --tagger=models=2;templates_1=tagger;guesser_suffix_rules_1=12;guesser_enrich_dictionary_1=4;guesser_prefixes_max_1=0;use_lemma_1=0;use_xpostag_1=1;use_feats_1=1;provide_lemma_1=0;provide_xpostag_1=1;provide_feats_1=1;prune_features_1=0;templates_2=lemmatizer;guesser_suffix_rules_2=6;guesser_enrich_dictionary_2=6;guesser_prefixes_max_2=4;use_lemma_2=1;use_xpostag_2=1;use_feats_2=1;provide_lemma_2=1;provide_xpostag_2=0;provide_feats_2=0;prune_features_2=0 --parser=iterations=30;embedding_upostag=20;embedding_feats=20;embedding_xpostag=0;embedding_form=50;embedding_form_file=ud-2.0-embeddings/ja.skip.forms.50.vectors;embedding_lemma=0;embedding_deprel=20;learning_rate=0.02;learning_rate_final=0.001;l2=0.3;hidden_layer=200;batch_size=10;transition_system=projective;transition_oracle=static;structured_interval=8`

| | default parameter | | | | Japanese configuration | | | |
|---|---|---|---|---|---|---|---|---|
| | word | UPOS | UAS | LAS | word | UPOS | UAS | LAS |
| no coordination [UD v2.2] | 91.0 | 88.4 | 75.5 | 74.0 | 91.8 | 89.1 | 77.0 | 75.4 |
| Left-head coordination | 91.0 | 88.2 | 71.7 | 69.9 | 91.6 | 88.6 | 73.6 | 71.8 |
| Right-head coordination | 91.0 | 88.2 | **75.4** | **72.6** | 91.6 | 88.6 | **76.7** | **74.0** |

Table 3: Parsing performance on Japanese UD corpora. F1 values of tokenization, the Universal POS tagging Score (UPOS), the Unlabeled Attachment Score (UAS), and the Labeled Attachment Score (LAS) are shown here.

| | UPOS | | | UAS | | | LAS | | |
|---|---|---|---|---|---|---|---|---|---|
| | GSD | Kaist | Penn | GSD | Kaist | Penn | GSD | Kaist | Penn |
| Left-head coordination | 89.37 | 90.12 | 92.17 | 69.49 | 77.54 | 73.54 | 61.98 | 70.37 | 65.94 |
| Right-head coordination | 89.39 | 90.10 | 92.41 | **77.22** | **83.00** | **78.34** | **65.03** | **75.02** | **69.18** |

Table 4: Parsing performance on the three Korean UD corpora, GSD, Kaist, and Penn. The gold-tokenization is used, and F1 values of UPOS tagging, UAS and LAS are reported.

to train word embeddings, where skip-gram with negative sample is used for language modeling and the vector size of 50 and the minimum count of 3 are used for configuration (the default values are used for all the other parameters).

The GSD and Kaist Treebanks are experimented with the configuration recommended by the UDPipe team, which was optimized on the CoNLL'17 shared task dataset.[3] The Penn Treebank is experimented with mostly the same configuration except that the transition-based parsing algorithm using the SWAP transition with the static lazy oracle is applied because this corpus allows multiple roots as well as non-projective dependencies, which is not assumed for the recommended configuration.

Following the annotation guidelines, the conj, flat, and fixed relations in the version 2.2 of the GSD and Kaist Treebanks are all left-headed. However, the authors of these Korean UD corpora also provide the right-headed version of those corpora from their open-source project. This project provides both left- and right-headed versions of the Penn Treebank as well, which makes it easy for us to make head-to-head comparisons.[4]

Table 4 shows parsing performance of UDPipe on the Korean UD corpora. Significant improvements are found in all three corpora for both the unlabeled and labeled attachment scores when the right-headed version is used. Moreover, our qualitative analysis indicates that the improvements are not just from those three relations, conj, flat, and fixed, but other relations associated with them because the right-headed version makes them more coherent with the other relations.

## 6 Proposal

The strict left-headed constraint for the coordinate structures in the current Universal Dependencies has tied the hands of the two individual language UD projects, driving them to adopt suboptimal solutions: dropping the conjunction category entirely in the case of Japanese, and maintaining two forks of the same data sets in the case of Korean (Section 4). The former approach incurs the loss of a real and essential cross-linguistic parallelism involving conj which undermines the UD framework's premise of universality; the latter risks splintering of the UD as a linguistically diverse yet unified project.

[3] `--tokenizer='dimension=24;epochs=100;initialization_range=0.1;batch_size=50;learning_rate=0.01;dropout=0.2;early_stopping=1'--tagger='models=2;templates_1=tagger;guesser_suffix_rules_1=8;guesser_enrich_dictionary_1=6;guesser_prefixes_max_1=0;use_lemma_1=1;use_xpostag_1=1;use_feats_1=1;provide_lemma_1=0;provide_xpostag_1=1;provide_feats_1=1;prune_features_1=0;templates_2=lemmatizer;guesser_suffix_rules_2=6;guesser_enrich_dictionary_2=5;guesser_prefixes_max_2=4;use_lemma_2=1;use_xpostag_2=0;use_feats_2=0;provide_lemma_2=1;provide_xpostag_2=0;provide_feats_2=0;prune_features_2=1'--parser='iterations=30;embedding_upostag=20;embedding_feats=20;embedding_xpostag=0;embedding_form=50;embedding_form_file=ko-all.vec;embedding_lemma=0;embedding_deprel=20;learning_rate=0.01;learning_rate_final=0.001;l2=0.5;hidden_layer=200;batch_size=10;transition_system=projective;transition_oracle=dynamic;structured_interval=10'`

[4] The official release of the UD Penn Korean Treebank can be obtained only through the Linguistic Data Consortium (LDC) such that the corpus in this open-source project does not include the form field.

Even if one was inclined to regard these drawbacks as merely abstract, hopefully we have sufficiently demonstrated that the adherence to the left-headed principle leads to numerous language-internal inconsistencies (Section 3) and, moreover, has an engineering-side consequence, as parser trainability is negatively impacted (Section 5).

Given these considerations, we propose that the UD guidelines be modified so as to allow flexibility in head orientation for coordinate structures. This move will leave our two UD teams free to apply right-headedness in coordinate structures and hence represent them in a way that is linguistically sound and with engineering-side advantages, all without making a compromise.

Additionally, general UD issues like the scope problem triggered by adjectival modification of coordinate structures (Section 3.1) can be resolved through right-headed attachment (*i.e.*, making the right conjunct ('cat') the head of the coordination). While admittedly right-headed attachment is not a complete solution for UD's general issue of adjectival modification of coordination, for the right-headed languages, at least, would allow the syntax to supply appropriate syntactic structures for the ambiguities present in the text[5].

Furthermore, it is our belief that the change will ultimately prove beneficial to all head-final languages. Rather than viewing this modification as a concession, we invite the UD leadership to consider the fact that coordination manifests differently across languages, and sometimes in a manner that strongly indicates headedness, as it does in Japanese and Korean; extending the head parameter to coordination will therefore strengthen the UD's position of universality. This flexibility may arise another issue in drawing a line between left- or right-headed, but any languages can keep the current strategy without any drawbacks, and apparently, it is beneficial for the rigid head-final languages.

## 7 Conclusion

In this paper, we presented issues that Japanese and Korean face in the representation of coordinate structures within the current design of Universal Dependencies, followed by a proposal for the UD principles to allow right-headedness in coordination. We hope this proposal will lead to more flexibility in the annotation scheme for the two languages, which will be essential in creating corpora that are useful not only for academic research but also for real-world use cases.

[5]Note that in "犬とかわいい猫" ('dog and cute cat'), where 'cute' modifies 'cat' (the head of coordination), ambiguity is resolved through word order (*i.e.*, cannot be read as both of them are cute).

## References


Masayuki Asahara, Hiroshi Kanayama, Takaaki Tanaka, Yusuke Miyao, Sumire Uematsu, Shinsuke Mori, Yuji Matsumoto, Mai Omura, and Yugo Murawaki. 2018. Universal Dependencies Version 2 for Japanese. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan.

Xinying Chen and Kim Gerdes. 2017. Classifying languages by dependency structure. typologies of delexicalized Universal Dependency treebanks. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017), September 18-20, 2017, Università di Pisa, Italy*, 139, pages 54–63. Linköping University Electronic Press.

Jinho D. Choi and Martha Palmer. 2011. Statistical Dependency Parsing in Korean: From Corpus Generation To Automatic Parsing. In *Proceedings of the IWPT Workshop on Statistical Parsing of Morphologically Rich Languages*, SPMRL'11, pages 1–11, Dublin, Ireland.

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)*, Miyazaki, Japan.

Marie-Catherine De Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D Manning. 2014. Universal Stanford dependencies: A cross-linguistic typology. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, volume 14, pages 4585–4592.

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. In *Proceedings of the National Academy of Sciences of the United States of America*.

Kim Gerdes and Sylvain Kahane. 2015. Non-constituent coordination and other coordinative constructions as dependency graphs. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 101–110.

Jack Hoeksema. 1992. The head parameter in morphology and syntax. In Dicky G. Gilbers and S. Looyenga, editors, *Language and Cognition 2*, pages 119–132. University of Groningen.

Hiroshi Kanayama, Youngja Park, Yuta Tsuboi, and Dongmook Yi. 2014. Learning from a neighbor: Adapting a Japanese parser for Korean through feature transfer learning. In *Proceedings of the EMNLP'2014 Workshop on Language Technology for Closely Related Languages and Language Variants*, pages 2–12.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Proceedings of Advances in Neural Information Processing Systems 26*, NIPS'13, pages 3111–3119.

Joakim Nivre, Željko Agić, Lars Ahrenberg, et al. 2017. Universal Dependencies 2.0. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University, Prague.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia.

Maria Polinsky. 2012. Headness, again. *UCLA Working Papers in Linguistics, Theories of Everything*, 17:348–359.

Adam Przepiórkowski and Agnieszka Patejuk. 2018. Arguments and adjuncts in universal dependencies. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3837–3852.

Sebastian Schuster and Christopher D Manning. 2016. Enhanced english universal dependencies: An improved representation for natural language understanding tasks. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 23–28. Portorož, Slovenia.

Milan Straka, Jan Hajič, and Jana Straková. 2016. UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia. European Language Resources Association.

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 2016)*.

Daniel Zeman, Filip Ginter, Jan Hajič, Joakim Nivre, Martin Popel, and Milan Straka. 2018. CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, Brussels, Belgium. Association for Computational Linguistics.