Building Universal Dependency Treebanks in Korean

Jayeol Chun[1], Na-Rae Han[2], Jena D. Hwang[3], Jinho D. Choi[1]
Emory University[1], University of Pittsburgh[2], IHMC[3]
Atlanta GA 30322[1], Pittsburgh, PA 15260[2], Ocala, FL 32502[3]
che.yeol.chun@emory.edu, naraehan@pitt.edu, jhwang@ihmc.us, jinho.choi@emory.edu

Abstract

This paper presents three treebanks in Korean that consist of dependency trees derived from existing treebanks, the Google UD Treebank, the Penn Korean Treebank, and the KAIST Treebank, and pseudo-annotated by the latest guidelines from the Universal Dependencies (UD) project. The Korean portion of the Google UD Treebank is entirely re-tokenized to match the morpheme-level annotation suggested by the other corpora, and systematically assessed for errors. Phrase structure trees in the Penn Korean Treebank and the KAIST Treebank are automatically converted into dependency trees using head-finding rules and linguistic heuristics. Additionally, part-of-speech tags in all treebanks are converted into the UD tagset. A total of 51K+ dependency trees are generated that comprise a coherent set of dependency relations for over a half million tokens. To the best of our knowledge, this is the first time that these Korean corpora are analyzed together and transformed into dependency trees following the latest UD guidelines, version 2.

Keywords: universal, dependency, conversion, korean, treebank

1. Introduction

The Universal Dependencies (UD) project has brought on an increasing momentum to the research community for finding morphological patterns and syntactic relations acceptable to multiple languages (Zeman et al., 2017). The UD project has facilitated collaborative work among several organizations for 70+ languages, and inspired computational linguists to further analyze both resource-rich and -poor languages by suggesting universal guidelines that help them create and augment treebanks in different languages. The UD project has also promoted research on cross-lingual learning that explores the possibility of adapting statistical parsing models from one language to another (McDonald et al., 2013).

Several treebanks had been introduced for Korean, all of which comprised annotation of morphemes and phrase structure trees (Choi et al., 1994; Han et al., 2002; Hong, 2009), each following its own set of guidelines. Phrase structure trees in these treebanks had been converted into dependency trees using head-finding rules and linguistically-motivated heuristics, and used to evaluate Korean dependency parsing performance (Choi and Palmer, 2011; Choi, 2013). The previous efforts did not, however, focus on the compatibility among dependency trees converted from different corpora, resulting in the generation of a distinct set of dependency relations for each treebank.

This paper presents three dependency treebanks in Korean, derived from existing corpora and pseudo-annotated by the latest UD guidelines, version 2. The motivation behind this study is to make a comprehensive analysis between these corpora and convert phrase structure trees across different treebanks into dependency trees with consistent relations, providing a large corpus comprising compatible dependency trees. The contributions of this work are as follows:

- The Google UD Korean Treebank is manually assessed and systematically corrected (Section 3).
- Phrase structure trees in both the Penn Korean Treebank and the KAIST Treebank are converted into dependency trees using the UD guidelines (Sections 4 and 5).
- Corpus analytics are provided that include statistics of the new dependency treebanks, and remaining issues with the current annotation (Section 6).

To the best of our knowledge, this is the first time that these Korean corpora are analyzed together and transformed into dependency trees following the latest UD guidelines.

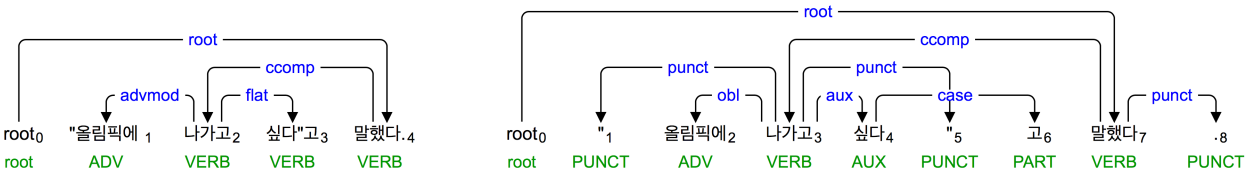
2. Related Work

Petrov et al. (2012) introduced the universal part-of-speech tagset and provided a mapping from 25 different treebank tagsets to this universal set. They showed that parsing performance using the universal POS tagset was comparable to the one using the original tagsets. McDonald et al. (2013) presented the universal dependency annotation and provided pseudo and manually annotated dependency treebanks for 6 languages. They showed promising results for cross-lingual parsing and initiated the effort for developing universally acceptable grammars. The official UD project started with a group of 10 languages (Nivre et al., 2015) and has expanded to over 70 languages. Recently, this project organized the CoNLL'17 shared task on multilingual parsing, involving over 40 languages (Zeman et al., 2017).

3. Google UD Korean Treebank

McDonald et al. (2013) provided the Google UD Treebanks comprising 6K sentences scraped from weblogs and newsire, annotated by their universal dependency guidelines for 6 languages including Korean. These treebanks were annotated before the official UD project started; hence, the guidelines under which the Korean treebank was created differed significantly from that of the version 2 of the UD (UDv2). The Google UD Korean Treebank (GKT) was automatically converted to follow the UDv2 guidelines, and distributed as a part of the CoNLL'17 shared task datasets.

As a means of quality assessment of the newly converted GKT, we perform a manual check over GKT to determine whether or not this automatic conversion generated sound dependency relations.



(a) A dependency tree from the origianl GKT.

(b) The retokenized dependency tree in our corpus.

Figure 1: Examples of the dependency trees in the original Google UD Korean Treebank and our corpus.

3.1. Automatic Morphological Analysis

Korean is an agglutinative language with highly productive verbal and nominal suffixation as well as limited prefixation. Without morphological analysis, then, any system that solely relies on surface forms must contend with the sparsity issue. McDonald et al. (2013) also pointed out that the automatic tokenization carried out for the original GKT was generally too coarse-grained; the suffixes or particles were left in with the tokens, indicating the necessity for future improvements through manual revision and annotation. To help remedy this problem, we augment GKT with automatic morphological analysis obtained by the KOMA tagger, a general-purpose morphological analyzer for Korean (Lee and Rim, 2009) that produces the morpheme tagset defined by the Sejong Treebank (Hong, 2009). The full morphological analysis is included for each token as the last column in our dataset.

3.2. Proper Tokenization

The tokenization in GKT does not split out the inflectional and derivational particle as separate tokens. Furthermore, the improper tokenization of punctuation in GKT can lead to inappropriate dependency relations. In Figure 1(a), the double quotes in the 1st and 3rd tokens are parts of those tokens. The morphological analysis from the KOMA tagger enables to recognize symbols as well as particles so that they are split into separate tokens in our corpus (Figure 1(b)). Dependency labels for these new tokens are inferred from their morpheme tags. Over 9K tokens with embedded punctuation are revised, resulting in 3K additional tokens.

3.3. Dependency Relabeling

There has been more changes made to UDv2 since GKT was converted for the CoNLL'17 shared task. Thus, we apply morpheme-level rules to GKT and relabel all dependency relations to reflect the latest updates in UDv2. In Figure 1(b), the 2nd and 3rd tokens translate to *Olympics+in* and *participate*, respectively. KTB considers *Olympics+in* an adverbial modifier (`advmod`) of *participate*, which is relabeled as an oblique (`obl`) in our corpus, as specified in UDv2.

3.4. Lexical Correction

We manually assess the entire GKT for spelling errors. Social media is one of the main sources for GKT, which include a disproportionately large number of misspellings. Some are common incorrect spellings (e.g., 왠만하ㅌ → 웬만하ㅌ) or deliberate non-standard forms known as 'netspeak' (e.g., Ü른 → ë @), while the rest are simple errors. Additionally, the HTML entity symbols are replaced with corresponding lexical symbols (e.g, `&` → &). The corrected spellings, 146 tokens in total, are provided in the lemma column.

4. Penn Korean Treebank

Han et al. (2002) created the Penn Korean Treebank (PKT) consisting of phrase structure trees for 5K sentences from a military corpus, known as the Virginia corpus. Han et al. (2006) presented the PKT 2.0 comprising manual annotation of morphemes and phrase structure trees for 15K sentences from newswire in Korean. PKT is the only Korean treebank including annotation of empty categories, which enables to generate non-projective dependencies. The Virginia corpus is excluded from our conversion due to the lack of generality in its source, the military domain.

4.1. Part-of-speech Tags

The part-of-speech (`POS`) tags are manually mapped from PKT to UDv2;[1] for the most part, this mapping is categorical. One exception is `DAN`, determiner-adnominal, which encompasses two semantically distinct subgroups: (1)demonstrative prenominals (e.g., ㅌ, 그,) and (2)attribute adjectives that lack predicative counterparts (e.g., È, 헌). The former is mapped to `DET` (determiner); the latter to `ADJ` (adjective). Additionally, we identify nominal and verbal particles whose function are to encode conjunction and assigned them to the appropriate UDv2 `POS` tags. `PCJ` (conjunctive post-position) is singled out and assigned to `CCONJ` (conjunction), while the remaining post-position categories (`PCA`, `PAC`, `PAU`) are mapped to `ADP` (adposition). The `ECS` (coordinate, subordinate, adverbial) verbal endings require additional attention to context: they are categorized as `CCONJ` when they are considered coordinating verbs or verb phrases, and as `SCONJ` when considered coordinating clauses. All remaining verbal endings are categorized as `PART` (particle) along with copula (`CO`) and suffixes (`X*`).

4.2. Dependency Relations

The establishment of dependency relations starts with handling empty categories, which are nominal units that point to the location of their antecedent syntactic elements elsewhere in the sentence. PKT features 4 empty categories that capture long-distance dependencies: trace (`*T*`), ellipsis (`*?*`), empty assignment (`*pro*`) and empty operator (`*op*`). These are redirected to establish correct dependency tree structure. Then each node is assigned its head with head-percolation rules based on Table 1. The dependency

[1]The mappings between the `POS` tagsets from PKT, KTB, and UDv2 are not presented here due to the page limit, but will be provided upon the acceptance.

relationship between the node and its head is inferred by investigating the function tags, phrasal tags as well as morphemes.

Phrase	D	Headrules
S	r	VP;ADJP;S;NP;ADVP;*
VP	r	VP;ADJP;VV|VJ;CV;LV;V*;NP;S;*
NP	r	NP;S;N*;VP;ADJP|ADVP;*
DANP	r	DANP|DAN;VP;*
ADVP	r	ADVP;ADV;-ADV;VP;NP;S;*
ADJP	r	ADJP;VJ;LV;*
ADCP	r	ADC;VP;NP|S;*
ADV	r	VJ;NNC;*
VX	r	NNX;*
VV	r	VV;NNC;VJ;*
VJ	r	VJ;NNC;*
PRN	r	NPR;N*|NP|VP|S|ADJP|ADVP;*
CV	r	VV;*
LV	r	VV;J;*
INTJ	r	INTJ;IJ;VP;*
LST	r	NNU;*
X	r	*

Table 1: Headrules for PKT. **Phrase** lists all phrasal tags in PKT. **D** denotes the search direction, and r denotes searching for rightmost constituent. * denotes any tag headed by what follows. | denotes logical or. ; is a delimiter between tags. Each **Headrule** gives higher precedence to the left tag on the list.

5. Kaist Treebank

Choi et al. (1994) created the KAIST Treebank (KTB) containing phrase structure trees for 31K sentences from various sources including literature, newswire, and academic manuscripts. Trees in this corpus were converted into dependency trees and used as a part of the shared task on parsing morphologically rich languages (Choi, 2013). Unlike PKB, KTB does not include empty categories and function tags, which makes the dependency conversion more challenging.

5.1. Part-of-speech Tags

Similarly to PKT, the KTB POS tag mapping, for the most part, is categorical; exhibiting many-to-one mappings from KTB to UDv2. In some cases, KTB and UDv2 take a different slice through the semantics of what these tags represent. For example, while the KTB's case particles generally map to the UDv2's adpositions (ADP), the conjunctive case particles (jcj) in KTB functionally align with the UDv2's conjunctions (CONJ). Much like PKT, the ending particles (x*) in KTB are analyzed on the basis of semantic context: adveribial derivational suffixes (xsa) signal assignments to the UDv2's adverbs (ADV), while the rest of the ending particles in KTB are considered PART in UDv2.

5.2. Dependency Relations

Except for the empty category handling, KAIST dependency conversion follows the procedure outlined for PKT where the head of nodes is located with head-percolation rules based on Table 2. While the dependency label inference benefits from the rich morphological analysis of KAIST, the small number of phrasal tags and the absence of function words has led to a complication such as the mapping of noun phrases ending with *jxt* to dislocated. Similarly to PKT, where *-SBJ* function tag denotes a subject node, KAIST offers three morpheme tags for the same purpose: *jcs*, *jcc*, and *jxt*. However, while *jcs* and *jcc* roughly correspond to nsubj and csubj, jxt suggests that the phrase is the topic of the phrase or clause, but offers nothing informative in distinguishing whether it is in fact a subject (which it frequently is) and, if so, whether it is a clausal or nominative subject. Although UDv2 offers dislocated for topical elements ubiquitous in languages like Korean and Japanese, KAIST treebank offers no systematic way of distinguishing dislocated from its subject counterparts in nsubj or csubj.

Phrase	D	Headrules
S	r	VP;ADJP;S;NP;ADVP;*
VP	r	pv*|pa*|nc*|VP;ADJP;NP;S;*
NP	r	n*|f;NP;S;pv*;VP;ADJP|ADVP|MODP;*
ADJP	r	ADJP;pa*;n*;ADVP;VP;NP;S;*
ADVP	r	ADVP;VP;ma*;NP;S;*
AUXP	r	AUXP;NP;p*;n*;px;*
MODP	r	mm*;VP;ADJP;NP;*
IP	r	ADVP;ADV;VP;NP;S;*
X	r	*

Table 2: Headrules for KAIST. Refer to Table 1 for tabular details.

6. Corpus Analytics

6.1. Statistics of the New Dependency Treebanks

The three corpora shared NOUN, VERB, ADV and PUNCT as the top parts-of-speech (Figure 2). Beyond these four, no other POS reaches double-digit %, and the relative rankings start to diverge. In both PKT and GKT, PROPN (proper noun) is the fifth-highest ranking POS, while it is seen ranking much lower in KAIST, which instead has ADJ (adjective) taking the spot. NUM (number) is prominent in PKT which is likely a reflection of its news domain. Notably, AUX (auxiliary) is lacking entirely in GKT, where all verbs are uniformly categorized as VERB. Given that auxiliary verb is a well-established category in Korean grammar, we find this a rather puzzling design decision.

The distributions of the dependency labels display intriguing trends across all treebanks (Figure 3). PKT and KAIST appear consistent except in compound, nummod, dislocated and nsubj. As briefly mentioned, compound and nummod are likely domain-specific particularities. As for dislocated and nsubj, the discussion of 5.2. likely explains the discrepancy. GKT's abundant annotation of flat is a remnant of coarse tokenization that led to embedded tokens labeled flat as a whole.

6.2. Discussions

GKT While a number of salient errors has been handled in this work, our analysis show that there are a number of remaining issues with GKT that we strongly recommend be

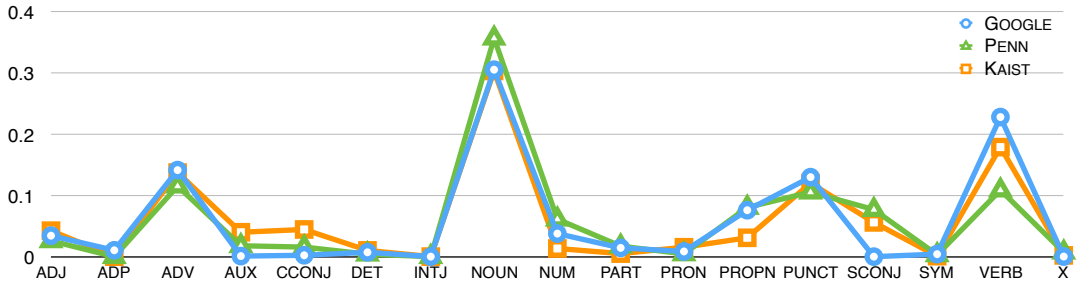


Figure 2: Distributions of part-of-speech tags for all three treebanks.

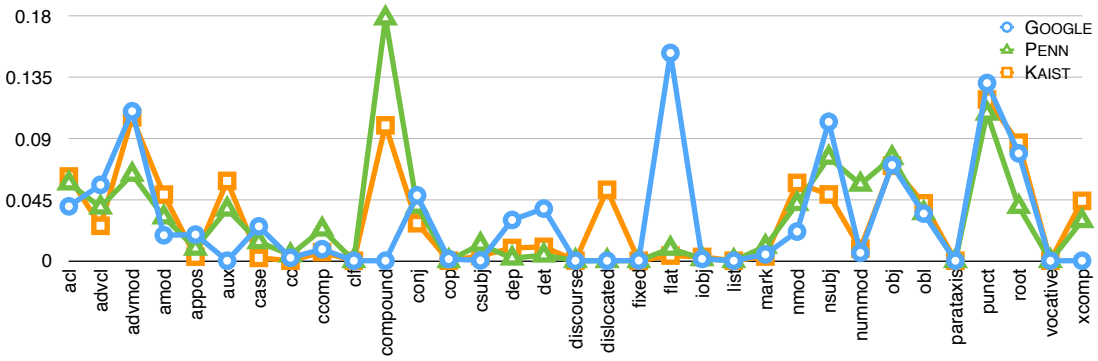


Figure 3: Distributions of the dependency labels for all three treebanks.

addressed in a future release of the data[2]. The errors include structural problems, incorrect argument attachment, and incorrect dependency labelling. Additionally, GKT shows a (mostly) consistent tendency to go with a head-first analysis in cases of conjunction (i.e., *talking* is the direct dependent of *good* for conjunction *talking and reading*) and noun-noun compounds[3], both of which represent inconsistent treatments of a verb-final language.

Additionally, the GKT currently contains duplicates in the dataset, many of which are fairly complex sentences. Out of the 195 duplicates present in the data (out of total 6,339 sentence tokens), 113 duplicates appear verbatim in both the training and test sets (represents over 11% of the test data) and 28 duplicates cross over training and development sets (represents 3% of the development set), which indicates a flawed data sampling process.

PKT and KTB The conversion and error-analysis for PKT has undergone various iterations and the current status of the UDv2 compliant PKT data is near completion. PKT is praised for its strong annotation consistency; that coupled with well-publicized documentation meant that we were able to quickly and reliably implement targeted conversion strategies.

KTB, being our newest converted treebank, has not yet received the in-depth error-analysis that has benefited PKT. One of the major issues with KTB, which we are currently dealing with is that unlike the PKT, KTB lacks function tags. This made it difficult to discern, for instance, clausal subjects from adverbial clauses.

One issue that often came up was the treatment of grammaticalized multi-word expressions such as -ㄹ ƒ t 다 (*-l kesita*) and -ㄹ ˆ 다 (*-l swu issta*). On the face of it, they involve dependent nouns ƒ (*kes*, 'thing') and (*swu*, 'way') respectively to literal translations of '... will be a thing' and 'there is a way to ...'. On the whole, however, they are grammaticalized forms that encode future/irrealis and epistemic modality, respectively: PKT acknowledges this and marks them as multi-word auxiliaries in annotation which facilitated our conversion process. In KTB, these forms had to be individually and lexically targeted to ensure parallel treatment. The Google Treebank, however, does not make such provision; as a matter of fact, it lacks AUX as a POS category altogether, which means this corpus remains disparate on this issue. This illustrates difficulty in achieving uniformity across multiple corpus resources by way of automatic and semi-automatic conversion.

7. Conclusion

In this paper, we have presented the manual assessment and revision process for the GKT, and the phrase-structure to UD conversion of Penn Korean and KAIST treebanks, discussing some of the statistics and the current issues relating the three presented treebanks. In the final version of the paper we also seek to provide examination of the correlation of the dependency relations to the grammatical necessities of Korean syntax, and an analysis of the strengths and shortcomings of applying the UD framework to the Korean language.

[2]We suspect these errors were present in the original annotation of the corpus and propagated to the current distribution of CoNLL'17 shared task data.

[3]This is true even in a noun-noun compound where one of the noun explicitly case marked such as "**샐러드 바-를** 9 D ˆ 다" (tr. **salad bar-acc** can eat), where *salad* is assigned the head even though *bar* is marked with the accusative case.

8. Bibliographical References

Choi, J. D. and Palmer, M. (2011). Statistical Dependency Parsing in Korean: From Corpus Generation To Automatic Parsing. In *Proceedings of IWPT workshop on Statistical Parsing of Morphologically Rich Languages*, SPMRL'11, pages 1–11.

Choi, K.-S., Han, Y. S., Han, Y. G., and Kwon, O. W. (1994). KAIST Tree Bank Project for Korean: Present and Future Development. In *In Proceedings of the International Workshop on Sharable Natural Language Resources*, pages 7–14.

Choi, J. D. (2013). Preparing Korean Data for the Shared Task on Parsing Morphologically Rich Languages. Technical Report 1309.1649, ArXiv.

Han, C.-H., Han, N.-R., Ko, E.-S., Palmer, M., and Yi, H. (2002). Penn Korean Treebank: Development and Evaluation. In *In Proceedings of the 16th Pacific Asia Conference on Language, Information and Computation*, PACLIC'02.

Han, N.-R., Ryu, S., Chae, S.-H., yun Yang, S., Lee, S., and Palmer, M. (2006). Korean Treebank Annotations Version 2.0. `https://catalog.ldc.upenn.edu/LDC2006T09`.

Hong, Y. (2009). 21st Century Sejong Project Results and Tasks (21세기 세 ... Ä 획 ¬ 업 성 ü 및 ü). In *New Korean Life (새국어생 \)*. National Institute of Korean Language.

Lee, D.-G. and Rim, H.-C. (2009). Probabilistic Modeling of Korean Morphology. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(5):945–955, July.

McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Täckström, O., Bedini, C., Bertomeu Castelló, N., and Lee, J. (2013). Universal Dependency Annotation for Multilingual Parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, ACL'13, pages 92–97.

Nivre, J., Bosco, C., Choi, J., de Marneffe, M.-C., Dozat, T., Farkas, R., Foster, J., Ginter, F., Goldberg, Y., Hajič, J., Kanerva, J., Laippala, V., Lenci, A., Lynn, T., Manning, C., McDonald, R., Missilä, A., Montemagni, S., Petrov, S., Pyysalo, S., Silveira, N., Simi, M., Smith, A., Tsarfaty, R., Vincze, V., and Zeman, D. (2015). Universal Dependencies 1.0. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.

Petrov, S., Das, D., and McDonald, R. (2012). A Universal Part-of-Speech Tagset. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, LREC'12, pages 2089–2096.

Zeman, D., Popel, M., Straka, M., Hajic, J., Nivre, J., Ginter, F., Luotolahti, J., Pyysalo, S., Petrov, S., Potthast, M., Tyers, F., Badmaeva, E., Gokirmak, M., Nedoluzhko, A., Cinkova, S., Hajic jr., J., Hlavacova, J., Kettnerová, V., Uresova, Z., Kanerva, J., Ojala, S., Missilä, A., Manning, C. D., Schuster, S., Reddy, S., Taji, D., Habash, N., Leung, H., de Marneffe, M.-C., Sanguinetti, M., Simi, M., Kanayama, H., dePaiva, V., Droganova, K., Martínez Alonso, H., Çöltekin, c., Sulubacak, U., Uszkoreit, H., Macketanz, V., Burchardt, A., Harris, K., Marheinecke, K., Rehm, G., Kayadelen, T., Attia, M., Elkahky, A., Yu, Z., Pitler, E., Lertpradit, S., Mandl, M., Kirchner, J., Alcalde, H. F., Strnadová, J., Banerjee, E., Manurung, R., Stella, A., Shimada, A., Kwak, S., Mendonca, G., Lando, T., Nitisaroj, R., and Li, J. (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*, CoNLL'17, pages 1–19.