

Enriching the Syntactic Annotation of Korean Treebanks for Higher-Level Processing
-A Comparative Study of the Penn Korean Treebank and the 21st Sejong Korean Treebank -

This paper explores several important issues in developing syntactically annotated Korean corpora for higher-level language processing, including semantic-discourse parsing, question-answering, machine translation, information retrieval, etc. In particular, we compare the Penn Korean Treebank (PTK) and the Korean Treebank of the 21st Century Sejong Project (ST) and discuss four critical issues in syntactic annotation. We argue for the use of more sophisticated morphosyntactic information, and based on our comparative study, we propose revisions in the syntactic annotation schemes of the existing Korean Treebanks in order to improve the quality of annotated corpora and their usability both for conducting theoretical research and for developing computational tools.

The results of our comparative study reveal four significant issues in syntactic annotations: the syntactic analysis of verbal complexes, the hierarchical structure of noun phrases, the representation of traces, and the marking of zero elements. These factors may trigger erroneous syntactic representations for certain linguistic phenomena and may increase difficulties in data search and lessen reliability in computational processing. Thus, evaluating and improving syntactic annotation of Treebanks is an important task for aspects of both theoretical and computational linguistics.

The first notable discrepancy appears with syntactic structures of verbal complexes. The PTK separates each component of a verbal complex and allows each auxiliary verb in that complex to project to a VP, as in (1a). This contrasts with the syntactic analysis of the ST, which combines verbal complexes under the same phrasal category, as in (1b). While considering the agglutinative properties of Korean, which license strong morphosyntactic dependencies among multiple verbal elements, we argue for a unified syntactic annotation for verbal clusters in Korean.

The second controversial factor relates to the hierarchical structure of NPs. In the PTK, all nouns appearing in an NP are licensed in a flat structure (as in (2)), whereas the ST brackets nouns that appear in a semantically close relation. The flat structure approach is potentially problematic because it may assign incorrect modification relations to examples like (2). In (2), *choykun-ey sellipton* ‘recently established’ modifies *sausuwest hangkong* ‘Southwest Airline’ but not the entire NP corresponding to ‘one airplane that belongs to Southwest Airline’. Furthermore, the flat structure analysis tends to increase computational complexity by allowing too many tokens of noun complexes. For example, it is possible to provide multiple analyses for the unambiguous example in (3a). While the correct analysis should be (3b), the flat structure analysis also allows (3c-e).

Another issue involves the representation of traces. While the PTK assumes traces for certain long-distance dependency constructions, the ST simply does not assume traces at all. The former approach overgenerates trace constructions by assigning empty *wh*-operators to relative clauses. In contrast, the latter undergenerates traces and fails to capture the syntactic and semantic dependency between a trace and its filler. We specify advantages and disadvantages of both approaches and evaluate them. In doing so, we consider how information about traces is essential to semantic parsing and machine translation, and we consider the unique properties of long-distance dependency constructions in Korean.

The final point relates to the syntactic marking of zero elements, which are different from traces. Issues regarding zero elements in the PTK have been already discussed in Lee et al. (2004). In line with this, we show that the ST approach to empty categories is problematic because it loses all the information required for the retrieval of semantic interpretations. We claim that syntactic annotations need to be based on the classification of zero elements and systematic predicate-argument relations.

In addition to suggesting revised syntactic annotations, we argue for adding more sophisticated morphosyntactic classifications. For example, specifying verbal nouns that require arguments is useful for the correct analysis of argument structure and for aiding extraction of event nouns for event tagging. As we examine the four issues mentioned above, we develop an approach to each one that allows the treebank annotation to capture linguistic phenomena correctly and to facilitate the application of computational linguistic technology.

(1) a. PTK: 대대장-의 허가 없이 쓰지 못하게 되어 있습니다.
 battalion commander-GEN permission without use cannot become be

‘(It) is not supposed to be used without permission from the battalion commander’

(VP (VP (VP (ADVP (NP-COMP (NP 대대장/NNC+의/PCA)
 (NP 허가/NNC))
 없이/ADV+는/PAU)
 (VP (NP-OBJ *T*-1)
 쓰/VV+지/EAU))
 못하/VX+게/EAU)
 되/VX+어/EAU)
 있/VX+습니다/EFN))

b. ST: 악몽-의 순간을 되새기-고 싶어하-지 않았다.
 bad dream-GEN moments-ACC remember-END want-END don't

‘(I) didn't want to remember the moments of the bad dream.’

(VP (NP_OBJ (NP_MOD 악몽/NNG + 의/JKG)
 (NP_OBJ 순간/NNG + 을/JKO))
 (VP (VP (VP 되새기/VV + 고/EC)
 (VP 싶/VX + 어/EC + 하/VX + 지/EC))
 (VP 않/VX + 았/EP + 다/EF))

(2) (최근 설립된) (NP 미국/NPR 사우스웨스트/NPR 항공/NNC 소속/NNC 여객기/NNC 1/NNU 대 /NNX+어/PCA))

choykun selliptoy-n mikwuk sauswest hangkong sosok yekaykki 1 tay
 recently establish-REL American Southwest airline belong to airplane 1 CLASSIFIER

‘One airplane that belongs to Southwest Airline, which has been recently established’

(3) a. (NP 우리 엄마 가죽 지갑 속)
 wuli emma kacwuk cikap sok

‘the inside of my mother's leather wallet’

- b. (NP (우리 엄마) ((가죽 지갑 속))) c. (NP 우리 (((엄마 가죽 지갑 속)))
 d. (NP ((우리 엄마) 가죽) (지갑 속)) e. (NP (우리 엄마) (가죽 (지갑 속))) etc.

References

- Chung-hye Han, Na-Rae Han, Eon-Suk Ko, Heejong Yi and Martha Palmer (2002) Penn Korean Treebank: Development and Evaluation, *Proceedings of the 16th Pacific Asia Conference on Language, Information and Computation*. The Korean Society for Language and Information. (2002)
- Han Chung-hye, Na-Rae Han, & Eon-Suk Go (2001) Bracketing Guidelines of Penn Korean Treebank. *Technical Report, IRCS-01-10*.
- Korean Treebank Guidelines of 21st Sejong Project (2003)
- Lee Sun-Hee, Donna K. Byron, Whitney Gegg-Harrison (2004) Annotating Zero Anaphors in Korean Using the Penn Korean Treebank. *The Third Workshop on Treebank and Linguistic Theories*, Tuebingen, Germany