Constructing a Dependency Treebank for Second Language Learners of Korean

Hakyung Sung[1], Gyu-Ho Shin[2]
[1]Department of Linguistics, University of Oregon, hsung@uoregon.edu
[2]Department of Linguistics, University of Illinois at Chicago, ghshin@uic.edu

Abstract

We introduce a manually annotated syntactic treebank based on Universal Dependencies, derived from the written data of second language (L2) Korean learners. In developing this new dataset, we critically evaluated previous works and revised the annotation guidelines to better reflect the linguistic properties of Korean and the characteristics of L2 learners. The L2 Korean treebank encompasses 7,530 sentences (66,982 words; 129,333 morphemes) and is publicly available at: https://github.com/NLPxL2Korean/L2KW-corpus.

Keywords: Korean, learner corpus, dependency annotation

1. Introduction

The Universal Dependencies (UD) framework has rapidly gained prominence in the field of morpho-syntactic annotation, representing both a consistent method for cross-lingual annotations and a collection of annotated corpora from a variety of languages (De Marneffe et al., 2021; Nivre et al., 2016, 2020). These UD treebanks now support various research activities in natural language processing (NLP), from syntactic to semantic parsing, while also gaining traction in language science, especially in learner corpus research (Chinese: Lee et al., 2017; English: Berzak et al., 2016; Kyle et al., 2022; Italian: Di Nuovo et al., 2019; Swedish: Masciolini et al., 2023).

In learner corpus research, NLP tools (e.g., part-of-speech taggers, syntactic parsers) have been used to study the lexico-grammatical patterns of second language (L2) learners for over a decade (e.g., Bestgen and Granger, 2014; Biber et al., 2011; Lu, 2010). As the field advances, there emerges a need to develop more specialized resources that incorporate diverse learner languages. This necessity arises mainly from two observations: First, many learner corpus studies utilize NLP tools that were trained predominantly on first-language datasets when analyzing L2 data. Such an approach might not always capture the unique patterns of learner languages (Kyle, 2021; Meurers and Dickinson, 2017). Consequently, researchers have investigated the performance of these tools on gold-annotated L2 datasets to assess their reliability before directly using them for analyses (e.g., Kyle and Eguchi, 2023). Second, beyond mere evaluation, these gold-annotated datasets have been used to train domain-general taggers/parsers, thereby enhancing the accuracy of the tools in L2 contexts (Kyle et al., 2022; Sung and Shin, 2023a).

In response to these needs, we aim to construct a UD treebank for L2 learners of Korean. This involves adding head word indices and dependency relations to a publicly available L2 Korean learner corpus. In the process, we explore previous Korean UD datasets/studies (e.g., Korean-GSD, Korean-Kaist, Korean-PUD[1]; Kim et al., 2018; Lee et al., 2019; Seo et al., 2019) and develop revised annotation guidelines that align more closely with the language-specific properties of L2 Korean. Moreover, we establish guidelines to ensure consistent syntactic annotations, addressing common errors, such as spacing and spelling, often found in L2 datasets.

2. Dataset Overview

For this annotation project, we use the L2 Korean Learner Morpheme (KLM) corpus (Sung and Shin, 2023b), which is publicly and freely available. The data consists of manually tokenized and annotated morphemes, determined through thorough discussions and cross-validation by human annotators. Originating from six hundred texts written by L2 learners of Korean, the KLM corpus evenly represents six proficiency levels, with one hundred texts at each level. All example sentences in this paper (except for a few comparative ones from the Google UD Korean treebank [GSD]) are sourced from this dataset, although some sentences have been streamlined due to space limitations and minimally edited for clarity.

3. UD Guidelines for Korean

In this section, we design core guidelines that we establish before beginning the annotation process.

[1]These datasets are available at https://universaldependencies.org/ko/index.html.

LREC-COLING 2024, pages 3747–3758
20-25 May, 2024.

While we primarily adhere to the current UD guidelines (De Marneffe et al., 2021), we also consult earlier UD studies on head-final languages, specifically Korean and Japanese.

3.1. Head Word Index

The head-assignment principle of UD stipulates that each token must be connected to a single head. The main verb of a clause, or the pivotal word when a verb is absent, is labeled as `root`. The `root` then acts as the anchor, defining the syntactic relationships among other lexical elements in a clause, as illustrated in Figure 1.

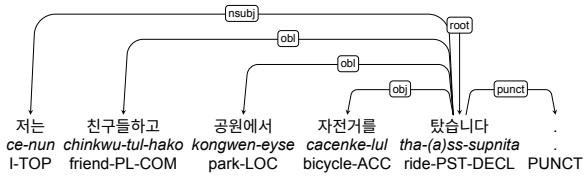

Figure 1: Example of an annotation
'I rode a bicycle in the park with my friends.'

While designating the head is straightforward in simple phrases/clauses, complexities arise in certain structures such as coordination (e.g., 사과와 바나나 *sakwa-wa panana* "apples and bananas") and multi-word nominals[2] (e.g., 학교 버스 *hakkyo pesu* "school bus"). In response to these complexities, the framework advocates for a unified approach across all languages, asserting that the word on the left (in the word combination) should serve as the head. Following this, Seo et al. (2019) suggested positioning the head on the left when annotating those coordination and multi-word nominals in Korean, emphasizing that the primary objective of the UD project is to construct cross-lingual corpora, rather than to address language-specific traits.

Although the left-head approach may align with the UD project's overarching goal, it is not without its critics. Several studies suggested that assigning the head to the rightmost element is more appropriate for Korean (Choi and Palmer, 2011; Chun et al., 2018; Han et al., 2020) and other head-final languages like Japanese (Asahara et al., 2018; Kanayama et al., 2018) in certain cases. In this study, we adhere to the latter approach for those particular cases. Subsequent sections delve into coordination and multi-word nominals, elucidating the advantages of our chosen approaches.

[2]We define 'multi-word nominals' as expressions where a head noun is modified by preceding nouns, which differs from the fixed grammaticized multiword expressions outlined in the UD guidelines. See here: https://universaldependencies.org/workgroups/newdoc/two_nominals.html

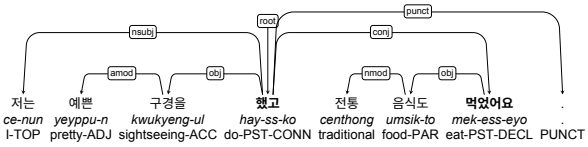

Figure 2: Coordination (Left-headed)

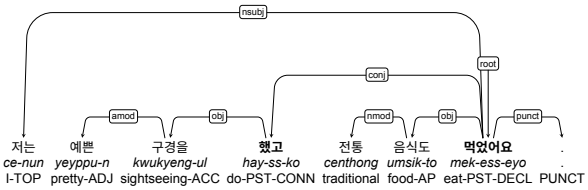

Figure 3: Coordination (Right-headed)
'I saw the beautiful sights (of Mt. Seorak) and ate traditional food.'

3.1.1. Coordination

Coordination refers to the combination of two or more grammatical elements (e.g., words, phrases, clauses) that are of equivalent syntactic type. These grammatical elements, known as conjuncts, are typically connected using coordinating conjunctions. In Korean, the connective (ending) marker of a predicate can be represented by -고 *-ko*, signifying conjuncts, and therefore should be tagged as `conj`. However, when using a left-head approach, an irregularity arises: the connective marker -고 is assigned the `root` tag, while the sentence-final predicate, which should likely be the `root`, is labeled with the `conj`. This configuration is also problematic as it creates an unusually long distance between the root and punctuation marks (2).

In contrast, when using a right-head approach, -고 is given the `conj` tag, while the sentence-final predicate receives the `root` tag. This representation appears to align more closely with the linguistic interpretations (Figure 3) (for further examples, see Chun et al., 2018, p. 2197).

3.1.2. Multi-word Nominals

Annotating dependency relations in multi-word nominals requires understanding of how Korean particles function in discerning the syntactic relationships between nouns (Sohn, 1999). Specifically, these particles help designate the roles of nouns as either core arguments (e.g., subject, object) or non-core arguments (e.g., oblique) within a clause. Syntactically, these particles, which are functional morphemes, commonly follow nouns, or content morphemes, without any word boundaries.

Case markers, a sub-category of particles, define the grammatical role of a noun in a sentence, marking a core argument role such as subject (-이/-가 *-i/-ka*), object (-을/-를 *-ul/-lul*), and posses-

sive (-의 *-uy*). On the other hand, postpositions, another subset of particles, express various spatial/temporal relationships, such as -에서 *-eyse* in 학교**에서** *hakkyo-**eyse***, meaning "**in/at/from** the school". When paired with a postposition, nouns often represent non-core arguments in the clause, like obliques[3].

In multi-word nominals, applying the left-head principle tends to diminish the importance of case markers and postpositions in establishing grammatical relationships between noun phrases. Figure 4 illustrates that the leftmost nouns, missing particles and modifying the subsequent noun, are categorized as `obl` or `obj`. This categorization does not confirm with linguistic interpretations of case markers. Therefore, we suggest that the rightmost nouns, which carry a case marker (e.g., -을 *-ul*) and a postposition (e.g., -에서 *-eyse*) should be assigned the dependency relation stemming from the `root` as in Figure 5.

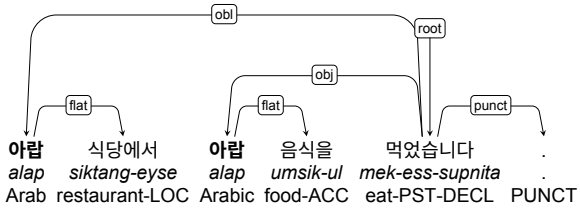

Figure 4: Multi-word nominal (Left-headed)

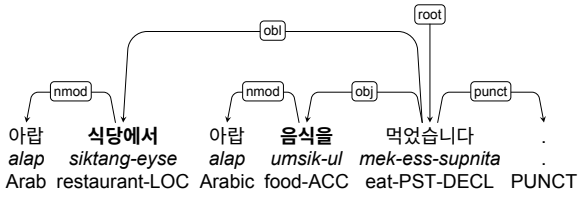

Figure 5: Multi-word nominal (Right-headed)
'(I) ate Arabic food at an Arab restaurant.'

3.2. Tags Excluded

In this section, we discuss two critical annotation tags that are excluded from the annotation scheme.

3.2.1. Indirect Object

The `iobj` tag is employed to denote an indirect object, a recipient of an event of transfer, as illustrated in the English example *I gave **him** a present* (i.e., a double-object construction). However, its applicability to the Korean context has been debated.

Prior research suggested that the `iobj` tag might be superfluous in Korean (Kim et al., 2018; Lee et al., 2019). While English differentiates an indirect object in a double-object construction, Korean recognizes a recipient primarily through particles or context. Consequently, there is no pressing necessity to syntactically differentiate a recipient nominal from other arguments using the `iobj` tag.

A contrasting viewpoint posited that any noun phrase accompanied by recipient-related particles (e.g., -에게 *-eykey*, -한테 *-hanthey*) should be labeled with the `iobj` tag (Seo et al., 2019). Nevertheless, this approach faces two challenges. First, not all arguments acting as a recipient are marked with the aforementioned particles (e.g., 영수가 선물을 **철수** 주었어 *Yengswu-ka senmwul-ul **Cheolsu** cwu-ess-e* "Yengswu gave **Chelswu** a gift."). Second, the presence of such markers does not necessarily identify a noun as a recipient role in some cases, the marked noun may function merely as a goal (e.g., 영수가 **철수에게** 편지를 썼다 *Yengswu-ka **Chelswu-eykey** phyenci-lul ssu-ess-ta* "Yengswu wrote a letter to **Chelswu**").

Based on these discussions, this study aligns with the view that disregards the `iobj` tag, rather than opting to label a recipient as `obl` (Figure 6).

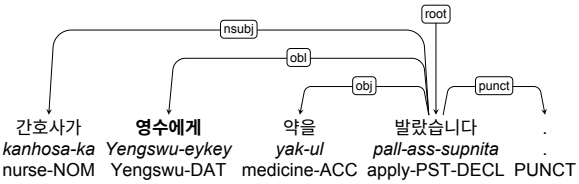

Figure 6: `iobj` replaced with obl
'The nurse applied medicine to Yengswu.'

3.2.2. Obligatory Control within Clausal Complements

In the UD annotation scheme, two dependency tags are related to clausal complement of a predicate: `ccomp` and `xcomp`. The `ccomp` tag denotes a clausal complement without an obligatory controlled subject (e.g., *She said **that she was tired***). In contrast, the `xcomp` tag is for a clausal complement with an obligatory controlled subject (e.g., *He wants **to eat ice cream***).

In Korean, Kim et al. (2018) and Lee et al. (2019) suggested merging `ccomp` and `xcomp` tags into a single `ccomp` tag. They pointed out that, while `ccomp` and `xcomp` have distinct syntactic properties in English, they do not in Korean. Specifically, in English, `ccomp` often includes a complementizer

[3]Postposition falls under the adposition category, a class of words articulating spatial/temporal relationships between a noun and other clause elements. In English, an adposition takes the form of a preposition (e.g., *in, on, under*), placed before the noun phrase it modifies as a distinct morpheme (e.g., ***in** the house*; ***on** the table*).

(e.g., *she said **that** she was tired*) and presents a finite clause, whereas `xcomp` does not. Conversely, in Korean, syntactic markers (e.g., -는, -냐고, -라고 *-nun, -nyako, -lako*) play a pivotal role in demarcating complement clauses, but these markers can be applied to both the English `ccomp` and `xcomp` cases, which makes the distinction between the two less clear in Korean. Following these observations, we only use the `ccomp` tag for marking the clausal complements (Figure 7).

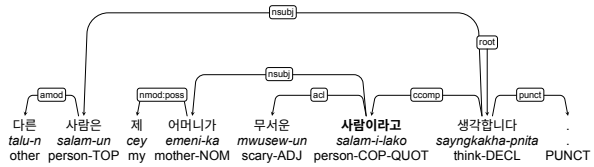

Figure 7: `ccomp`
'Others think my mother is a scary person.'

3.3. Controversial Tags

In this section, we discuss tags that have been contentious in previous Korean UD annotation work, with the aim of establishing clearer annotation guidelines for our study.

3.3.1. Oblique; Adverb Modifier

When determining how an argument modifies a predicate or another modifier, both `obl` and `advmod` tags are possible options. The `obl` tag is designated for a nominal acting as a modifier (i.e., oblique), while the `advmod` tag is for an adverb or adverbial phrase in a modifying role (i.e., adverb modifier).

In Korean, when noun phrases work as non-core arguments, differentiating them syntactically from adverbial modifiers becomes challenging. As a solution, Kim et al. (2018) suggested unifying these tags (`obl` & `advmod`) into oblique (`obl`), emphasizing the classification of core arguments (i.e., subject, object) versus non-core nominal arguments within a clause. In contrast, other studies (Lee et al., 2019; Seo et al., 2019) recommended adhering to the UD guidelines. These studies underscored the importance of determining whether the modifier originates from the noun phrase or the adverb itself, focusing on the part-of-speech of the word. In other words, if the modification arises from the noun phrase, it should be tagged as `obl`; if it comes from the adverb, the `advmod` tag should be applied, even though their roles as clause modifiers may be similar[4].

Our decision is to take the latter, a more fine-grained approach. By doing so, the `obl` tag encompasses two syntactic variations of non-core argument noun phrases in Korean: (1) those combined with a particle (Figures 8 & 9) and (2) those without a particle, in other words, a sole noun functioning as a modifier of a clause (Figure 10).

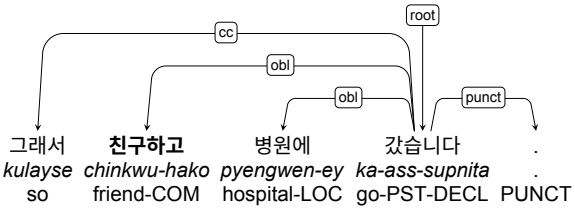

Figure 8: `obl` Case (1-1)
'So (I) went to the hospital with my friend.'

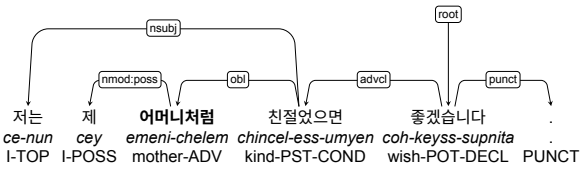

Figure 9: `obl` Case (1-2)
'I want to be kind like my mother.'

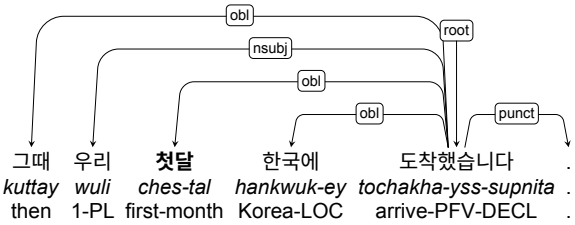

Figure 10: `obl` Case (2)
'At that time, we arrived in Korea on the first month.'

3.3.2. Auxiliary

The auxiliary tag (`aux`) is used for function words that express tense, aspect, mood, voice, direction, or evidentiality to a predicate. While English often represents these through auxiliary verbs (e.g., *be, can*), Korean employs multi-word verbal expressions instead. For instance, in expressions like 타고 가다 *tha-ko ka-ta* ("to take a ride and go"), a main verb is frequently succeeded by an auxiliary verb (e.g., 타고 **가다** *tha-ko **ka-ta***) (Kim et al., 2018; Lee et al., 2019; Seo et al., 2019).

Upon examining the Google Korean Universal

[4]For example, consider the following examples: 영수는 **자전거로** 학교에 갔다 *yengswu-nun **cacenke-lo** hakkyo-ey kass-ta* "Yengswu went to school **by bike**" (bold = `obl`); 영수는 **천천히** 학교에 갔다 *yengswu-nun **chenchenhi** hakkyo-ey kass-ta* "Yengswu went to school **slowly**") (bold = `advmod`)

Dependency treebank (GSD)[5] (Chun et al., 2018; Noh et al., 2018), we observe that auxiliary verbs were often assigned the `flat` tag. This tag is typically reserved for elements of headless semi-fixed multi-word expressions, such as names (Figure 11) — a categorization that may not align with a proper linguistic interpretation. Therefore, in our study, we have chosen to label auxiliary verbs that depend on the preceding `root` — to express features such as tense, aspect, mood, voice, direction, or even negation — with the `aux` tag (Figure 12).

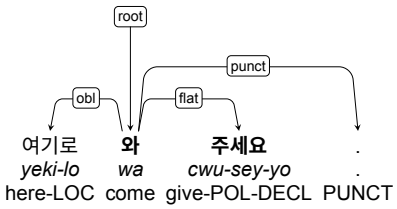

Figure 11: `root+flat` (Example-GSD)
'Please come here.'

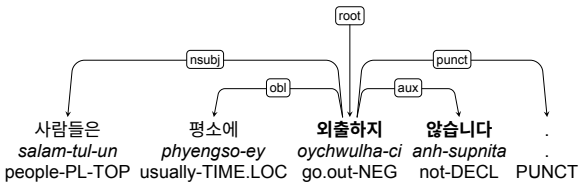

Figure 12: `root+aux`
'People don't usually go out.'

3.3.3. Copula

The `cop` tag denotes a copula, which is a function word that connects a subject to a non-verbal predicate. This tag is frequently observed in English attributive constructions (e.g., *He **is** a teacher*). A nominal or adjectival phrase follows a *be*-verb and describes an attribute of the subject.

In Korean, there is a copula morpheme -이- *-i-* that follows the preceding noun (e.g., 선생님 *sensaynghim* "teacher") and precedes a sentence-final functional morpheme (e.g., -다 *-ta*) without word-spacing boundaries (e.g., 그는 선생님**이**다 *ku-nun sensaynghim-**i**-ta* "He is a teacher"). This aspect challenges the assignment of the `cop` tag in Korean because the sentence-final predicate typically receives the `root` tag (Kim et al., 2018; Lee et al., 2019). Nevertheless, Seo et al. (2019) argued that this tag should still be recognized as it shows a unique relationship between the subject and a non-verbal predicate.

In this study, we follow Seo et al. (2019) but propose that both the `root` and the `cop` tags be retained. As a result, we use the alternative `root:cop` tag for a sentence-final ending with a copula (Figure 13).

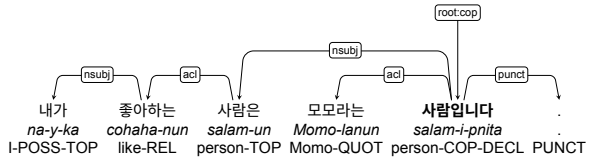

Figure 13: `cop`
'My favorite person is a person called Momo.'

3.3.4. Dislocated

The `dislocated` tag is designated for peripheral (either initial or final) nominals in a clause that do not serve as a core argument but work for other roles such as topic or afterthought. In the Japanese UD guidelines (Asahara et al., 2018), the tag is used for instances of double subject construction. Similarly, in Korean, some studies advocated for retaining the tag (Chun et al., 2018; Seo et al., 2019), emphasizing its importance in identifying a topic argument (i.e., an argument that introduces what the clause is about) within a clause.

Following this, we choose to retain the `dislocated` tag, particularly in the context of double subject constructions. We aim to tag the topic, which is typically positioned before the subject, while tagging the subsequent subject with the `nsubj` tag (Figure 14).

However, the use of this tag remains contentious. Some researchers (Kim et al., 2018; Lee et al., 2019) have suggested using double subject or object constructions instead, highlighting the challenges in assigning the `dislocated` tag to topic nouns. This is due to the instability of Korean topic markers (another subset of particles; e.g., -은/-는 *-un/-nun*) which can often be interchanged with subject/object case markers[6]. A similar discussion is noted in the Japanese UD project (Asahara et al., 2018, p. 1827).

[5]This corpus, sourced from online blogs and news produced by Korean native speakers, is commonly used to train Korean language model parsers. The dependency annotations were auto-converted from the original Google UD Korean treebank (McDonald et al., 2013) to align with the UDv2 guidelines (Chun et al., 2018).

[6]For instance, in the sentence provided in Figure 14, it is also grammatically correct to state 우리**가** 기분이 좋지 않다 *wuli-**ka** kipwun-i coh-ci anh-ta*, where the topic marker -는 *-nun* is replaced with the subject case marker -가 *-ka*.

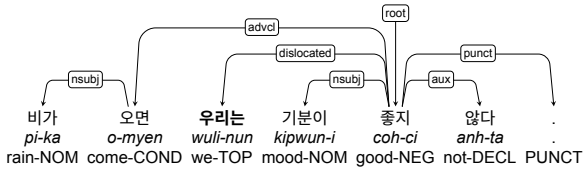

Figure 14: `dislocated`
'When it rains, we feel bad.'

3.4. Summary

Table 1 summarizes the dependency tags that we use in the corpus annotation[7].

Category	Tags
Nominals	
Clausal Core Arguments	`nsubj`, `obj`
Clausal Non-core Arguments	`obl`, `vocative`, `dislocated`
Nominal Dependents	`nmod`, `nmod:poss`, `appos`, `nummod`
Clauses	
Clausal Core Arguments	`csubj`, `ccomp`
Clausal Non-core Arguments	`advcl`
Nominal Dependents	`acl`
Modifier Words	
Clausal Non-core Arguments	`advmod`, `discourse`
Nominal Dependents	`amod`
Function Words	
Clausal Non-core Arguments	`aux`, `mark`
Nominal Dependents	`det`, `case`
Others	
Coordination	`conj`, `cc`
MWE	`fixed`, `flat`, `compound`
Loose	`list`, `parataxis`
Special	`goeswith`
Other	`punct`, `root`, `root:cop`, `dep`

Table 1: Summary of dependency tags

4. Annotation

4.1. Procedure

4.1.1. Automated Pre-tagging

Prior to the annotation process, the sentences were automatically tagged using Stanza[8], which provided information about the head of a word (either a value of ID or zero if it is the head of a clause) and a UD relation to the head. These correspond to the seventh and eighth columns in the CoNLL-U format (Nivre et al., 2020).

4.1.2. Manual Annotation Process

The corpus was primarily annotated by two native Korean speakers: the first author of this paper and a graduate student who specialized in Korean linguistics during their undergraduate studies. Initially, the annotators consulted the UD guidelines and related research, as outlined in the previous section. Both annotators independently reviewed the tags, either correcting those automatically tagged or adding missing tags due from the automated pre-tagging process. Given that the dataset already contained manual morpheme annotations from a prior project (Sung and Shin, 2023b), the annotators occasionally referred to this linguistic information for reference. Weekly meetings were held to adjudicate discrepancies in the tags and to update guidelines. After this, the third annotator, who is a native Korean speaker, a linguist, and the second author of this study, reviewed the annotations and provided feedback. We also had focused discussions to resolve disagreement on specific instances. Upon finalizing the annotations, they were reviewed once more for consistency, leading to some minor adjustments.

4.1.3. Inter-rater Reliability

To evaluate the initial inter-rater reliability between the two annotators (before the third check and adjudication), we used the Cohen's Kappa score. The score for the head index annotations was **0.93**, and that for the dependency tag was **0.94**. These scores indicate good and consistent agreement between the annotators.

4.2. Further Discussions on Tagging Scheme

In this section, we discuss the various issues and debates that emerged throughout the annotation process. The first issue (§4.2.1) relates to the syntactic nature of Korean, while subsequent issues (§4.2.2 and following) arise from the unique characteristics of the learner corpus.

4.2.1. Bound Noun

In Korean, bound nouns necessitate a sophisticated tagging approach due to their notable syntactic and semantic properties. Syntactically, there should be a space between a preceding element and a bound noun, causing them to appear as a single word[9]. Semantically, bound nouns resemble functional morphemes; unlike other nouns, they depend on preceding elements such as demonstratives, adjective phrases, or other nominals (Kim and Yang, 2007; Lee and Song, 2012; Martin and Shin, 2021; Sohn, 1999).

The concept of bound noun is not universally recognized across languages. Consequently, the UD guidelines do not specifically designate a tag for

[7] For detailed descriptions of each tag, please refer to De Marneffe et al., 2021, p. 266.

[8] https://github.com/stanfordnlp/stanza/

[9] See the 42nd clause of Article 5 in 한국어 어문 규범 *hankwuke emwun kyupem* "Korean Language Norms". Accessible at https://kornorms.korean.go.kr/m/m_regltn.do.

these elements (Seo et al., 2019, p. 98), and to our knowledge, no previous studies have clearly done so. According to linguistic references (Sohn, 1999, pp. 205-206; Martin and Shin, 2021), bound nouns should be annotated as dependent on their preceding noun or other aforementioned elements. In the following section, we outline three major cases involving this category.

Case 1: noun + bound noun (`obl + case`) When a noun phrase precedes a bound noun, we assign the `obl` tag to the nominal element and the `case` tag to the bound noun (Figure 16). By definition, the `case` tag is used to link a case-marking element to a nominal. This tag may not perfectly align with a bound noun but does partially resonate with how bound nouns are realized in a sentence, as they are often paired with case particles (e.g., -에 *-ey*).

Upon examining the GSD dataset, we notice that the preceding nominal element was frequently labeled as `nmod`, possibly considering the part-of-speech of the nominal element (Figure 15). However, this practice seems questionable because, according to its definition, the `nmod` tag should modify another nominal not a predicate. Therefore, we use the `obl` tag instead, characterized as a nominal modifying a predicate (Figure 16).

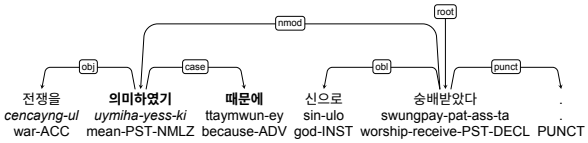

Figure 15: `nmod + case` (Example-GSD)
'(It) means war, so it was worshiped as a god.'

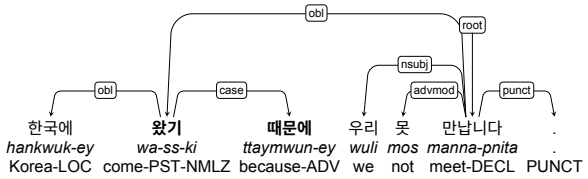

Figure 16: `obl + case`
'Since (I) came to Korea, we can't meet.'

Case 2: adjective clause + bound noun (`advcl + mark`) When an adjective clause precedes a bound noun, we assign the `advcl` tag to the adjective clauses (which is an adverbial clause modifying a predicate or a modifier word) and a `mark` tag (which is used for a function word linking a clause marked as subordinate to the predicate of that clause) (Figure 17). Even though the adjective clause is morphologically an adjective[10], the combination of the adjective clause and the bound noun functions as an adverb modifying a predicate. This approach is also consistent with the annotation scheme from the GSD.

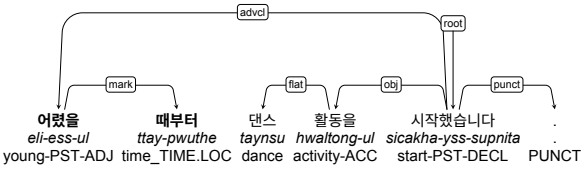

Figure 17: `advcl + mark`
'(I) started dancing at a young age.'

Case 3: bound noun as a core argument The two cases that we explained so far address bound nouns that act as modifiers of a predicate and are not core arguments within a clause. However, bound nouns can sometimes be merged with a subsequent case marker, thereby functioning as core arguments. In such scenarios, the preceding element modifies the bound noun and is dependent upon it. As a result, the bound noun is labeled with tags that represent its role as a core argument in the clause (e.g., `nsubj`). An example includes -**할 수가** *-hal swu-ka* as shown in Figure 18.

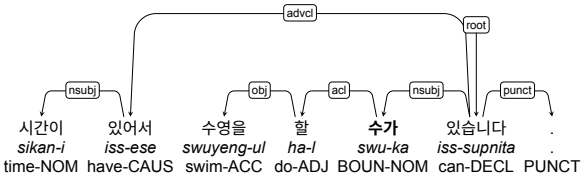

Figure 18: bound noun as a `nsubj`
'(I) have time to go swimming.'

4.2.2. Spacing Error

Case 1: redundant space Word spacing is the boundary between words that construct a sentence. L2 Learners often commit spacing errors, with one prevalent mistake being the insertion of a redundant space between a word and its particle. Despite the presence of these unnecessary spaces, our initial strategy involves identifying the dependency relations based on cues from particles, which are important for determining a word's function within a clause (as discussed in §3.1.2). Next, we link the word from a redundant space to the nearest tag, marking it as a dependent with the `goeswith` tag. This tag is intended to connect parts of a word that, according to standard orthography or linguistic

[10] In Korean, this is realized by adding -ㄹ *-liul* to the predicate.

conventions, should be treated as a single unit, suggesting that readers should perceive these tokens as one word (Figure 19).

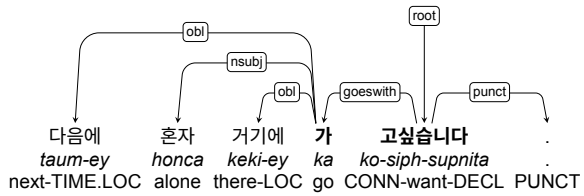

Figure 19: Space error 1
(corrected: 다음에 혼자 거기에 **가고 싶습니다**.)
'I want to go there by myself next time.'

However, in line with the UD guidelines that content words should primarily determine the main syntactic relationship (elaborated further in §4.2.4), when an extra space exists between a content word and its subsequent particles, we assign a dependency tag to the content word. The following particle is treated as dependent on the content word and annotated using the `case` tag (Figure 20).

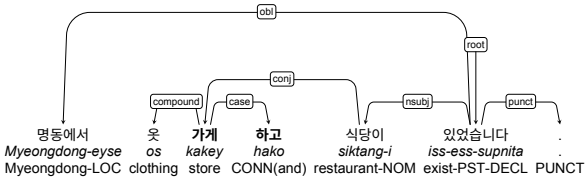

Figure 20: Space error 2
(corrected: 명동에서 옷 **가게하고** 식당이 있었습니다.)
'There was a clothing store and a restaurant in Myeongdong.'

Case 2: omitted space Conversely, instances of omitted spaces between word tokens frequently appear in the corpus. In these situations, the final morpheme within the merged token acts as a cue for determining the dependency relation (Figure 21). We avoid using a unique dependency tag to signify a missing space.

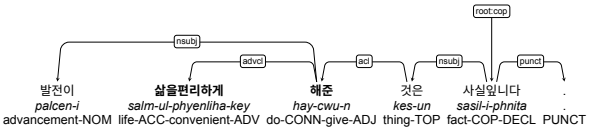

Figure 21: Space error 3
(corrected: 발전이 **삶을 편리하게 해 준** 것은 사실입니다.)
'It is true that advancements have made our lives more convenient.'

4.2.3. Spelling Error

Beyond spacing errors, spelling errors are another common mistake that L2 learners frequently make. We tag this type of error in consideration of the context in which they occurred, as showcased in Figures 22 and 23.

However, there were situations where discerning the syntactic relationship based on context became challenging due to the spelling error. In these cases, we apply the `flat` tag to the word, treating it as dependent on the subsequent word (Figure 24).

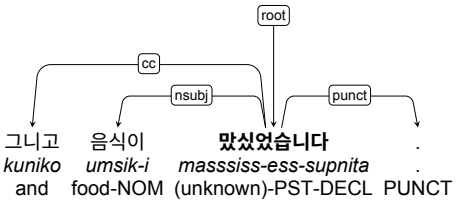

Figure 22: Spelling error 1
(corrected: 그리고 음식이 **맛있었습니다**.)
'And the food was delicious.'

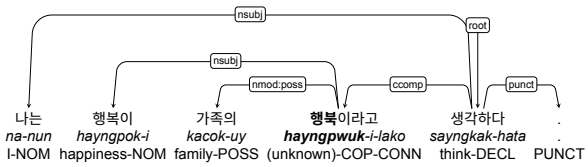

Figure 23: Spelling error 2
(corrected: 나는 행복이 가족의 **행복**이라고 생각하다.)
'I think happiness is the happiness of the family.'

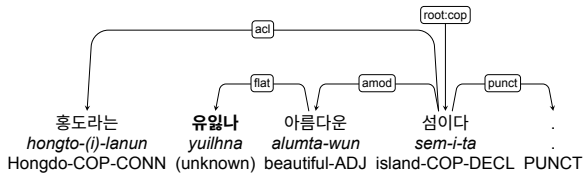

Figure 24: Spelling error 3
'(This) is an (unknown).beautiful island called Hong-do.'

4.2.4. Particle error and interpretation

Korean is notable in terms of its active use of particles in delivering grammatical relations and other types of syntactic cues (e.g., Kim and Ock, 2016). Literature on L2 Korean often reports learners' difficulty acquiring and using particles appropriately even when they achieve advanced proficiency in Korean (e.g., Kim, 2004; Kim and Guo, 2016). This poses a challenge to annotating dependency relations when a sentence involves atypical and/or erroneous use of particles.

To handle this issue, we rely on the original UD guidelines, which assume that primary dependency relations are anchored by content words such as verbs, nouns, and adjectives (De Marneffe et al.,

2021). That is, when we notice particle underuse/misuse in the dataset, we focus exclusively on the properties of content words such as the propositional meaning of a predicate.

For example, considering that the verb 많다 *manh-ta* "be.many-DECL" does not require an object, sentences produced by learners such as 책 많습니다 *chayk manh-supnita* (Figure 25) or 책을 많습니다 *chayk-ul manh-supnita* (Figure 26) can pose challenges in annotation. We labeled the term 책(을) *chayk(-ul)* as `nsubj` regardless of whether the noun is marked without particles or marked with the accusative case marker -을 *-ul*.

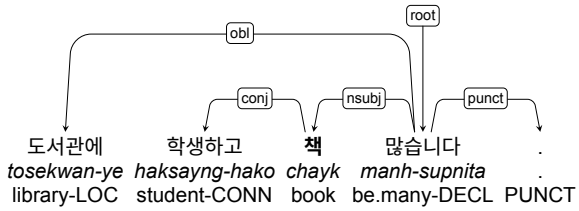

Figure 25: Particle underuse
(corrected: 도서관에 학생하고 **책이** 많습니다.)
'There are many students and books in the library.'

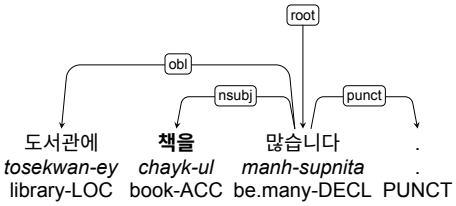

Figure 26: Particle error
(corrected: 도서관에 **책이** 많습니다.)
'There are many books in the library.'

On the other hand, there were instances in which grammatical errors (including errors in particles, spelling, etc.) made it challenging for the annotators to infer the obligatory arguments from a predicate. In such situations, we rely on unaffected particles among other cues to decide the dependency relationship between the nouns and the verbs in the given clauses (Figures 27, 28).

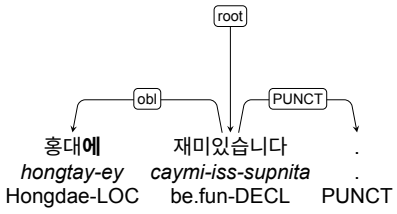

Figure 27: Particle interpretation 1
'(interpreted as) be fun to be in Hongdae.'

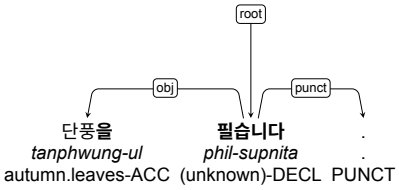

Figure 28: Particle interpretation 2
'(unknown).predicate autumn leaves.'

5. Conclusion

This study introduced a manually annotated UD treebank for the L2-Korean corpus, comprising 7,530 sentences, 66,982 words, and 129,333 morphemes (See Appendix B for the frequency of each dependency tag's occurrences). We have provided a detailed account of its development procedure, including the revisions made to the existing Korean UD annotation guidelines. Furthermore, we have incorporated language-specific properties and learner-language-specific characteristics into the dataset. The L2 Korean treebank is available for non-commercial use at the following repository: https://github.com/NLPxL2Korean/L2KW-corpus.

The process of crafting these annotation guidelines affirmed the efficacy of the L2 Korean UD annotation scheme. However, it also unveiled certain areas of under-specification, which points to a future research direction including the refinement of the tagging schemes as well as the quantitative extension of the annotated dataset. These improvements will ensure a better alignment with linguistic interpretations, ultimately enhancing computational resources for a variety of languages and linguistic registers.

6. Copyrights

7. Ethical considerations

We utilized the open-access KLM corpus (Sung and Shin, 2023), which is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. The KLM corpus is based on the open-access *Kyunghee* learner corpus (Park and Lee, 2021). We strictly adhered to ethical standards to ensure no negative impacts on societal or environmental domains. Every aspect of this research was conducted without any dishonest practices or misrepresentations.

8. Bibliographical 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)*.

Yevgeni Berzak, Jessica Kenney, Carolyn Spadine, Jing Xian Wang, Lucia Lam, Keiko Sophie Mori, Sebastian Garza, and Boris Katz. 2016. Universal dependencies for learner english. *arXiv preprint arXiv:1605.04278*.

Yves Bestgen and Sylviane Granger. 2014. Quantifying the development of phraseological competence in l2 english writing: An automated approach. *Journal of Second Language Writing*, 26:28–41.

Douglas Biber, Bethany Gray, and Kornwipa Poonpon. 2011. Should we use characteristics of conversation to measure grammatical complexity in l2 writing development? *Tesol Quarterly*, 45(1):5–35.

Jinho D. Choi and Martha Palmer. 2011. Transition-based semantic role labeling using predicate argument clustering. In *Proceedings of the ACL 2011 Workshop on Relational Models of Semantics*, pages 37–45.

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)*.

Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational linguistics*, 47(2):255–308.

Elisa Di Nuovo, Cristina Bosco, Alessandro Mazzei, Manuela Sanguinetti, et al. 2019. Towards an italian learner treebank in universal dependencies. In *CEUR workshop proceedings*, volume 2481, pages 1–6. CEUR-WS.

Ji Yoon Han, Tae Hwan Oh, Lee Jin, and Hansaem Kim. 2020. Annotation issues in universal dependencies for korean and japanese. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 99–108.

Eunkyul Leah Jo, Kyuwon Kim, Xihan Wu, Kyung-Tae Lim, Jungyeul Park, and Chulwoo Park. 2023. K-unimorph: Korean universal morphology and its feature schema. *arXiv preprint arXiv:2305.06335*.

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

Hansaem Kim et al. 2018. 2018년 국어 말뭉치 연구 및 구축 *[2018 Korean language corpus research and construction]*. National Institute of Korean Language, Republic of Korea.

JE Kim. 2004. The japanese learners misuse aspect for korean particles. *Journal of Korean Language Education*, 15(1):1–31.

Jong-Bok Kim and Jaehyung Yang. 2007. On the syntax and semantics of the bound noun constructions: With a computational implementation. In *Proceedings of the 21st Pacific Asia Conference on Language, Information and Computation*, pages 223–233.

Wan-su Kim and Cheol-young Ock. 2016. Korean semantic role labeling using case frame dictionary and subcategorization. *Journal of KIISE*, 43(12):1376–1384.

Young-joo Kim and Jin Guo. 2016. A study on the acquisition of korean adverbial case marker ey in spoken production by chinese korean l2 learners. *Kwukekyoyukyenkwu [Korean Education Research]*, 38:1–26.

Kristopher Kyle. 2021. Natural language processing for learner corpus research. *International Journal of Learner Corpus Research*, 7(1):1–16.

Kristopher Kyle and Masaki Eguchi. 2023. Assessing spoken lexical and lexicogrammatical proficiency using features of word, bigram, and dependency bigram use. *The Modern Language Journal*.

Kristopher Kyle, Masaki Eguchi, Aaron Miller, and Theodore Sither. 2022. A dependency treebank of spoken second language english. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 39–45.

Chanyoung Lee, Tae hwan Oh, and Hansam Kim. 2019. 한국어 보편 의존 구문 분석 (universal dependencies) 방법론 연구 [a study on universal dependency annotation for korean]. 언어사실과 관점 *[Language Facts and Perspectives]*, 47:141–175.

John SY Lee, Herman Leung, and Keying Li. 2017. Towards universal dependencies for learner chinese. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 67–71.

Sun-Hee Lee and Jae-young Song. 2012. Annotating particle realization and ellipsis in korean. In *Proceedings of the Sixth Linguistic Annotation Workshop*, pages 175–183.

Xiaofei Lu. 2010. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4):474–496.

J.R. Martin and Gi-Hyun Shin. 2021. Korean nominal groups: System and structure. *Word*, 67(3):387–429.

Arianna Masciolini, Elena Volodina, and Dana Dannlls. 2023. Towards automatically extracting morphosyntactical error patterns from l1-l2 parallel dependency treebanks. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 585–597.

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

Detmar Meurers and Markus Dickinson. 2017. Evidence and interpretation in language learning research: Opportunities for collaboration with computational linguistics. *Language Learning*, 67(S1):66–95.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Jan Hajič, Christopher D Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043.

Youngbin Noh, Jiyoon Han, Tae Hwan Oh, and Hansaem Kim. 2018. Enhancing universal dependencies for korean. In *Proceedings of the second Workshop on Universal Dependencies (UDW 2018)*, pages 108–116.

Saetbyol Seo et al. 2019. 한국어 보편 의존 관계 분석에 관한 제언 [a proposal on universal dependencies (v.2) annotation for korean]. 언어와 정보*[Language and Information]*, 23(1):91–122.

Ho-Min Sohn. 1999. *The Korean language*. New York, NY: Cambridge University Cambridge University Press.

Hakyung Sung and Gyu-Ho Shin. 2023a. Diversifying language models for lesser-studied languages and language-usage contexts: A case of second language korean. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11461–11473.

Hakyung Sung and Gyu-Ho Shin. 2023b. Towards l2-friendly pipelines for learner corpora: A case of written production by l2-korean learners. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 72–82.

9. Language resource references

Park, Jungyeul and Lee, Jung Hee. 2021. *Kyung Hee Korean Learner Corpus (v2)*. [link].

Sung, Hakyung and Shin, Gyu-Ho. 2023. *Korean Learner Morpheme Corpus*. [link].

Appendix

A. Glossing abbreviations

The table lists glossing abbreviations.

Abbreviation	Description
ADJ	Adjective
ADV	Adverb
BOUND	Bound noun
COM	Comitative
COND	Conditional
CONN	Connective
COP	Copula
DECL	Declarative
LOC	Locative
NEG	Negation
NOM	Nominative
PAR	Particle
PFV	Perfective
PL	Plural
POL	Polite
POSS	Possessive
POT	Potential
PST	Past
QUOT	Quotation
REL	Relative
TOP	Topic

B. Dependency tag counts

The table lists the counts for each dependency tag.

Dependency tag	Count
nsubj	8,767
punct	8,287
obl	7,332
root	6,866
obj	5,572
advmod	4,995
advcl	4,703
acl	4,501
nmod	2,059
aux	1,963
conj	1,860
amod	1,413
cc	1,306
nmod:poss	1,299
det	933
case	894
flat	854
root:cop	664
ccomp	642
dislocated	576
mark	509
list	303
goeswith	203
nummod	179
appos	128
compound	52
vocative	46
parataxis	37
csubj	22
discourse	6
fixed	6
dep	3
cop	2
Total	66,982