UD-KSL Treebank v1.3: A semi-automated framework for aligning XPOS-extracted units with UPOS tags

Hakyung Sung[1]   **Gyu-Ho Shin**[2]
Chanyoung Lee[3]   **You Kyung Sung**[4]   **Boo Kyung Jung**[5]

[1]Linguistics, University of Oregon
[2]Linguistics, University of Illinois Chicago
[3]Korean Language and Literature, Konkuk University
[4]Library and Information Science, Chung-Ang University
[5]East Asian Languages and Literatures, Yale University

Abstract

The present study extends recent work on Universal Dependencies annotations for second-language (L2) Korean by introducing a semi-automated framework that identifies morphosyntactic constructions from XPOS sequences and aligns those constructions with corresponding UPOS categories. We also broaden the existing L2-Korean corpus by annotating 2,998 new sentences from argumentative essays. To evaluate the impact of XPOS-UPOS alignments, we fine-tune L2-Korean morphosyntactic analysis models on datasets both with and without these alignments, using two NLP toolkits. Our results indicate that the aligned dataset not only improves consistency across annotation layers but also enhances morphosyntactic tagging and dependency-parsing accuracy, particularly in cases of limited annotated data.

1 Introduction

Ongoing efforts to develop linguistic annotations for learner corpora have produced valuable resources that support quantitative, targeted analyses of specific linguistic features (e.g., argument structure constructions: Sung and Kyle, 2024, stance-taking features: Eguchi and Kyle, 2023, grammatical errors: Dahlmeier et al., 2013, sign language: Mesch and Schönström, 2018). One such initiative focuses on morphosyntactic features, including part-of-speech (POS) categories and dependency relations, thereby allowing for more fine-grained investigations on linguistic structures produced by learners (Gries and Berez, 2017). These investigations can inform theoretical models of language development and improve empirical approaches to evaluating learner performance. In parallel, many learner corpora follow the Universal Dependencies (UD) framework, providing cross-linguistic consistency in grammatical structures via universal POS and dependency tags (Berzak et al., 2016;

Di Nuovo et al., 2019; Lee et al., 2017; Kyle et al., 2022; Rozovskaya, 2024).

Notably, second language (L2) Korean has recently been incorporated into this growing body of UD-annotated learner corpora (Sung and Shin, 2023, 2024, 2025). Previous research on UD annotations for L2 Korean has produced expert-curated resources with detailed XPOS tags from the Korean-specific Sejong set, enabling fine-grained morphosyntactic feature extraction. In contrast, the corresponding universal POS (UPOS) tags in these corpora were typically generated automatically—using a domain-general Korean analysis package (e.g., Stanza-GSD; Qi et al., 2020)—with minimal human validation (Sung and Shin, 2025). This disparity in annotation procedures may lead to inconsistencies, potentially undermining the dataset's internal reliability and reducing the accuracy of downstream applications.[1]

To address this gap, this study extends recent L2-Korean UD work (Sung and Shin, 2025) by introducing a semi-automated framework that aligns XPOS tags with UPOS categories, combining automation with targeted human validation. This framework is informed by the structure of Korean *eojeol*—a morphosyntactic unit defined by whitespace segmentation—and explains how different morphemes combine to form specific morphosyntactic categories.We also expand the L2-Korean corpus with 2,998 newly annotated sentences from argumentative essays. To assess the benefits of XPOS–UPOS alignment on model performance, we fine-tune L2-Korean morphosyntactic analysis models on datasets with and without this alignment using two NLP toolkits. Results show that alignment improves tagging and dependency parsing accuracy, particularly in low-resource settings—likely due to greater consistency

---

[1]According to Kanayama et al., 2017 (p. 270), UPOS tagging errors can negatively impact dependency parsing, one of the downstream tasks sensitive to annotation inconsistencies.

between UPOS tags and syntactic dependencies.

2 Datasets

2.1 L2-Korean UD treebank v1.2

We built upon the latest L2-Korean UD treebank (UD-KSL v1.2; Sung and Shin, 2025), which contains 12,984 manually annotated sentences. In its previous iterations, each sentence was annotated by trained linguists across three annotation layers: (1) Each eojeol was segmented into individual morphemes—the minimal meaning-bearing units, including both lexical roots and grammatical affixes (e.g., case particles, verbal morphology); (2) Each morpheme was tagged with its lexical or grammatical category using XPOS tags based on the Sejong tag set (Appendix A); (3) Dependency relations between eojeols indicating grammatical functions (e.g., subject, object) were annotated according to the UD framework (de Marneffe et al., 2021).

2.2 Data collection

Participant profiles and essay prompts We collected argumentative essays from 153 L2-Korean learners with diverse linguistic backgrounds, including Czech ($n = 40$; mean age = 24.3, SD = 2.8), English ($n = 49$; mean age = 23.7, SD = 4.5), Mandarin Chinese ($n = 36$; mean age = 25.5, SD = 3.2), and Korean as a heritage language ($n = 28$; mean age = 24.0, SD = 2.0). All texts were elicited through a genre-controlled writing tasks designed to assess learners' linguistic ability to construct and support claims in Korean.[2] Essay prompts were adapted from the official Test of Proficiency in Korean. For Mandarin Chinese-speaking learners, two prompts were used: (1) "Which do you think is more important, conservation of nature or development of nature?" (2) "Which do you prefer, competition or cooperation?"; for the other learner groups, three prompts were used (1) "Is early language education necessary for children?", (2) "Do we need to learn history?", (3) "Which do you prefer, competition or cooperation?".

Data elicitation and transcription Participants wrote argumentative essays by hand during individual Zoom sessions, with 20 minutes allocated per topic. Prompts were presented on the spot in both Korean and the participant's native language, and reference materials were not allowed. Handwritten

---

[2]The texts included in UD-KSL v1.2, which lacked genre control, consisted primarily of descriptive or narrative texts.

Figure 1: Example of the Korean C-test (Lee-Ellis, 2009)

essays were submitted as image files and manually transcribed into machine-readable texts by native Korean speakers with advanced linguistic expertise, preserving all original errors (i.e., no *a priori* corrections were made, nor was technical assistance applied, during manual transcription). All personally identifying information was anonymized.

Proficiency evaluation While collecting the samples, we measured participants' general Korean language proficiency using the Korean C-test (Lee-Ellis, 2009), which serves as a proxy for overall language ability by assessing comprehension of Korean sentences of varying lengths and complexity. The test comprises five passages with blanks inserted at the syllable level (Figure 1); each blank corresponds to a syllable and may appear in various positions within an eojeol. For testing efficiency, only the first four passages were used, as recommended by Lee-Ellis (2009). Participants received one point for each correctly restored blank, with a maximum possible score of 188. The test took approximately 20 minutes to complete, and participants' scores ranged from 37 to 181 ($M = 114$, $SD = 32.9$). These proficiency scores were included as metadata in the dataset. Although they were not used in the current analysis, we believe they may serve as a valuable resource for future studies.

2.3 Manual annotations: XPOS & deprel

Following the UD-KSL treebank v1.2 annotation procedure, we manually lemmatized eojeols, annotated XPOS tags, and marked dependency relations, using the three-layer approach described in Section 2.1. Four native Korean speakers served as annotators. Raw data were first auto-tagged using a *Stanza* Korean (GSD) model (Qi et al., 2020) fine-tuned on UD-KSL, and then reviewed and corrected by two primary annotators. Disagreements were resolved by a third annotator, with a fourth intervening if no consensus was reached. In total, 2,998 sentences were annotated and updated.

Annotation guideline Alongside the annotations, we developed an open-source annotation guideline covering 43 XPOS tags and 31 UD tags used in constructing the UD-KSL treebank.[3] Each tag was described in four categories: (1) Definition provided a brief explanation of the tag's core meaning; (2) Characteristics outlined its syntactic roles and functions in Korean, along with tagging guidelines; (3) Clarifications addressed ambiguous instances, distinctions from similar tags, exceptions, and rules for compound or derived forms (for XPOS only); and (4) Examples illustrated usage through representative examples drawn from the treebank.

3 XPOS-UPOS alignment

3.1 Motivation

The alignment between XPOS and UPOS tags is essential for capturing Korean's morphological richness while preserving the UD framework's cross-linguistic consistency. UPOS tags are intentionally coarse-grained to support cross-linguistic comparison by abstracting away from language-specific details (de Marneffe et al., 2021). While this abstraction serves the goals of universality, it also introduces challenges for morphologically rich languages such as Korean, where multiple grammatical elements are often agglutinated within a single spacing unit (Sohn, 1999). In such cases, the coarse granularity of UPOS may obscure important morphosyntactic information that is relevant for fine-grained linguistic analysis or learner language annotation (Han et al., 2020).

To illustrate this issue, consider the eojeol 학 생이 (glossed as student.NOM), which consists of two morphemes: (1) 학생 'student,' a lexical morpheme tagged as NNG (common noun), corresponding to the UPOS category NOUN; and (2) -이, a grammatical morpheme tagged as JKS (nominative case marker), which could map to the UPOS category PART. However, in the UD framework, UPOS tagging in Korean is applied at the eojeol level, requiring a single UPOS tag for the entire unit. In this case, it is typically labeled as NOUN, since the lexical noun functions as the syntactic head (cf. Noh et al., 2018).

When XPOS annotations are available, identifying the head morpheme within an eojeol enables more accurate and consistent mapping from XPOS to UPOS categories. This alignment preserves the

syntactic abstraction offered by UPOS while retaining key morphological details from the XPOS layer (e.g., Kanayama et al., 2018, Figure 3).[4]

3.2 Process and rationale

To construct reliable alignments between XPOS and UPOS tags, we used the gold-standard XPOS annotations from the UD-KSL v1.2. We first extracted all eojeol-level constructions,[5] each annotated with a sequence of XPOS tags. This yielded 2,080 unique constructions in the latest treebank, each representing a distinct morphological structure within an eojeol. We also recorded their frequencies to identify recurring patterns.

To focus manual review on common constructions, we applied a frequency threshold of five. Constructions that appeared more than five times were manually examined for XPOS–UPOS alignment, while those with five or fewer occurrences were assigned UPOS tags using default mapping heuristics. Notably, the manually reviewed constructions accounted for 96.41% (64,583 out of 66,989) of all eojeols in the treebank.

Using this frequency-screened dataset, we aligned each XPOS sequence with a corresponding UPOS tag. For example, NNG+JKO was mapped to NOUN, as it includes a common noun followed by an accusative case marker. Similarly, VA+EF was mapped to ADJ, reflecting a descriptive adjective followed by a sentence-final ending. Two Korean linguists independently performed the initial alignment using a double-blind procedure. Disagreements were adjudicated by a third linguist with relevant expertise. Table 1 presents representative constructions, their UPOS mappings, and corpus frequency counts.

3.3 Challenges

While direct alignment from XPOS to UPOS is currently the most practical approach, it inevitably sacrifices the rich, language-specific distinctions that XPOS encodes in favor of UPOS's universal categories (Lee et al., 2019). In Korean, where a single eojeol can encapsulate multiple morphemes

---

[3] https://nlpxl2korean.github.io/UD-KSL/

[4] To our knowledge, no fixed standard exists for mapping XPOS to UPOS in existing Korean UD treebanks. According to official UD guidelines, if an XPOS field is included, the treebank's README must specify how each XPOS tag maps to a UPOS value. This mapping may depend on additional contextual or annotated information (cf. https://universaldependencies.org/format.html).

[5] Drawing on a usage-based constructionist approach, we define *constructions* as morphosyntactic sequences within an eojeol that instantiate dedicated form-function mappings.

Eojeol	Composition	Gloss	XPOS tag	UPOS tag	Frequency
학교에	학교+에	school+LOC	NNG+JKB	ADP	2706
곳에	곳+에	place+LOC	NNB+JKB	ADP	284
이	이	DEM.PROX	MM	DET	176
정말	정말	really	MAG	ADV	4077
빠르게	빠르+게	be.fast+ADV	VA+EC	ADV	326
예쁘다	예쁘+다	be.pretty+DECL	VA+EF	ADJ	615
예쁜	예쁘+ㄴ	be.pretty+ADN	VA+ETM	ADJ	589
책을	책+을	book+ACC	NNG+JKO	NOUN	3679
책	책	book	NNG	NOUN	2546
학생이	학생+이	student+NOM	NNG+JKS	NOUN	2536
내가	나+가	I+NOM	NP+JKS	PRON	326
나도	나+도	I+FOC	NP+JX	PRON	759
먹고	먹+고	eat+CNJ	VV+EC	VERB	3553
먹는	먹+는	eat+RL	VV+ETM	VERB	2553
싶다	싶+다	want+DECL	VX+EF	AUX	639
싶어서	싶+어서	want+CNJ	VX+EC	AUX	303

Table 1: Examples of XPOS-to-UPOS alignment within Korean eojeols. Glosses follow the Leipzig Glossing Rules (see Appendix B for detailed descriptions).

with different syntactic functions, this one-to-one mapping cannot fully preserve grammatical nuance. Below, we list the UPOS labels that lacked direct XPOS equivalents during alignment; such labels are more likely to require case-by-case evaluation to ensure annotation accuracy.

Adverbial construction (ADV) Adverbial functions in Korean arise in two main ways: (1) through inflectional suffixes that attach to adjectival or verbal stems (e.g., the adverbializing suffix -게), and (2) through adverbial postpositions attached to nominal forms (e.g., the adverbial postpositions -에게). In our alignment scheme, the UPOS tag ADV is assigned only when explicit adverbial morphology is present. For example, 빠르게 (parsed_XPOS tagged as 빠르_VA+게_EC; 'fast' + adverbial suffix) is tagged ADV because -게 makes the stem function adverbially. Likewise, nominal forms with adverbial postpositions, such as 학교에서 (parsed as 학교_NNG+에서_JKB; 'school' + adverbial postposition), receive the ADV tag only if the XPOS sequence explicitly includes a recognized adverbial postposition.

Auxiliary verb construction (AUX) In Korean, auxiliary predicates, including both auxiliary verbs (e.g., 하려고 하다 and auxiliary adjectives (e.g., 예뻐 보이다), convey rich grammatical meanings and differ significantly from their Indo-European counterparts (Cho and Whit-

man, 2022). Under the UD framework, AUX typically denotes a closed class of verbs expressing tense, aspect, or modality.[6] However, many auxiliary verbs in Korean—tagged as VX under the XPOS scheme—retain substantial lexical meaning, complicating a purely functional classification. For example, in 먹어보다 (parsed as 먹_VV+어_EC+보_VX+다_EF; 'eat' + connective ending + 'try' + sentence-final ending), the auxiliary 보다 ('try') manifests its own lexical nuance rather than simply marking aspect or modality.

Auxiliary constructions can appear either within a single eojeol (e.g., 먹어보다) or split across multiple eojeols (e.g., 먹어 보았다). This variation depends on factors such as orthographic convention, formality, and speaker preference. When the construction appears as a single eojeol, our alignment process poses no difficulty: all morphemes are housed within one spacing unit, and the UPOS tag is determined by the syntactic head (typically the main verb) resulting in a VERB tag.

However, when the main and auxiliary verbs are split across two eojeols, additional analysis is needed to determine their syntactic roles. Predicate constructions were tagged as VERB or ADJ based on the lexical root, while accompanying auxiliaries were labeled AUX, following a predefined list (cf. Sung and Shin, 2025, Section 3.1.2). For example:

---

[6] https://universaldependencies.org/ko/index.html

- 가고 싶다 (가_VV+고_EC 싶_VX+다_EF, 'want to go'), the lexical verb 가고 ('to go') is tagged as VERB, and the auxiliary 싶다 ('to want') is tagged as AUX.

- 좋지 않다 (좋_VA+지_EC 않_VX+다_EF, 'to not be good'), the adjectival verb 좋지 ('to be good') is tagged as ADJ, and the negation expression 않다 ('not') is tagged as AUX.

While we followed UD guidelines for auxiliary constructions as closely as possible, the following cases required annotation adjustments due to syntactic constraints or gaps in the existing auxiliary inventories:

- 먹을 수 있다 (먹_VV+을_ETM 수_NNB 있_VX+다_EF, 'can eat'): In this construction, the main verb 먹다 ('to eat') is tagged as VERB, and the modal auxiliary 있다 ('can/be able to') ideally fits AUX. However, because 있다 functions as the clausal-level predicate, it was annotated as the syntactic root. As AUX cannot serve as a clause root under UD guidelines,[7] we tagged 있다 as ADJ—a compromise that preserves its predicative role while conforming to UD constraints.

- One exception to the AUX tagging scheme involved the verb 되다 ('to become'), which occurs in various clausal types including passive, aspectual, and modal constructions (e.g., 하게 되다, 'end up doing'). While 되다 functions grammatically as an auxiliary, it is not included in the closed list of auxiliaries under the current UD Korean guidelines. We thus annotated the entire construction as VERB. Nevertheless, based on its auxiliary-like morphosyntactic behavior, we suggest that 되다 in such contexts should be reconsidered as AUX for future annotation consistency.

Determinative ending for predicate (VERB, ADJ) In Korean, predicates (including verbs and adjectives) can combine with ETM morphemes to form noun-modifying clauses, serving a similar function to English participial or relative clauses. For instance, in 책을 읽은 사람 'the person who read a book', the verb 읽다 'to read' takes the ETM ending -은 to modify the noun 사람 'person.'

We assigned UPOS tags based on the lexical categories of predicates: forms derived from verbal stems (VV) were tagged as VERB, and those from adjectival stems (VA) were tagged as ADJ. For instance, in (책을) 읽는 사람 (parsed as 읽_VV+는_ETM 사람_NNG, 'who read the [book]'), the predicate 읽는 was tagged as VERB; in 예쁜 꽃 (parsed as 예쁘_VA+ㄴ_ETM 꽃_NNG 'a pretty flower'), the predicate 예쁘 was tagged as ADJ.

Case particle (NOUN, ADP) Case particles, attached morphologically to noun stems, play a crucial role in indicating grammatical functions such as subject, object, or adverbial modifiers. However, the UPOS tag set provides only a limited range of functional categories (e.g., ADP, PART), which cannot fully capture the morphosyntactic diversity found in Korean particles. In earlier UD annotations, noun phrases with different case particles were uniformly tagged as NOUN, masking their syntactic roles. In our alignment, we addressed this limitation by utilizing XPOS information to differentiate noun phrases based on particle type. For instance, noun phrases ending in topic markers (e.g., -은/는) or nominative case markers (e.g., -이/가) were retained as NOUN, as in 학생은 (학생_NNG+은_JX) ('the student [topic]') or 고양이가 (고양이_NNG+가_JKS) ('the cat [subject]'). In contrast, phrases marked with adverbial postpositions, such as -에서 ('at/from') or -로 ('by/with'), were classified as ADP where appropriate, as in 학교에서 (학교_NNG+에서_JKB) ('at school') or 버스로 (버스_NNG+로_JKB) ('by bus').

3.4 Semi-automatic alignment

We aligned XPOS and UPOS through a semi-automatic, two-phase process that combined rule-based alignment with manual validation and iterative refinement. First, we developed an automatic alignment script by using a predefined lookup table that mapped each Sejong XPOS tag to its corresponding UPOS tag. This step corrected 3,063 UPOS tags in the annotated texts of the current work (Section 2.2) and 11,691 tags in the existing UD dataset (Section 2.1). Next, a principal annotator conducted three rounds of manual verification. In the first round, a random 10% of corrected tokens were reviewed to flag mismatches and ambiguous cases. In the second round, the lookup table was modified based on common errors (e.g., auxiliary versus main predicates, adverbial postpo-

---

[7] https://universaldependencies.org/bm/pos/AUX_.html

UPOS tag	UD-KSL v1.2			UD-KSL working set		
	Unaligned	Aligned	Δ (A–U)	Unaligned	Aligned	Δ (A–U)
ADJ	4952	9267	+4315	2580	3810	+1230
ADP	1176	1015	-161	290	106	-184
ADV	19545	18864	-681	6332	6237	-95
AUX	1993	1968	-25	754	747	-7
CCONJ	9	7	-2	—	—	—
DET	1265	1421	+156	589	596	+7
NOUN	29481	29835	+354	9669	9720	+51
NUM	418	453	+35	95	104	+9
PART	1	1	0	2	1	-1
PRON	2771	3107	+336	713	747	+34
PROPN	19	—	-19	—	—	—
PUNCT	13032	13030	-2	3342	3342	—
SYM	2	—	-2	—	—	—
VERB	26117	21822	-4295	7825	6787	-1038
X	189	180	-9	79	73	-6

Table 2: Changes in UPOS tag frequencies before and after the alignment process applied to the UD-KSL v1.2 and UD-KSL working set.

sitions) and the script was re-run. In the final round, spot checks were performed on all remaining corrected tokens, and any remaining issues were resolved by consensus.

Table 2 presents the distribution of UPOS tags after completing the entire process across two datasets: (1) the original dataset from the previous L2-Korean UD treebank project (*UD-KSL-v1.2*), and (2) the annotated dataset developed in the current work (*UD-KSL working set*).

4 Experiments

We conducted experiments to assess the impact XPOS-UPOS alignment on model performance using a 2×2×2 design. The factors were: dataset type (*UD-KSL v1.2* vs. *UD-KSL working set*); refinement type (*aligned* [a dataset in which UPOS tags were aligned with corresponding XPOS tags] vs. *unaligned*); and toolkit type (*spaCy* vs. *Trankit*). L2-Korean morphosyntactic analysis models were fine-tuned on both dataset versions with both toolkits to determine whether the XPOS-UPOS alignment enhance the accuracy of morphosyntactic parsing and tagging in L2-Korean data.

4.1 Model training and evaluation

We used two open-source NLP toolkits—*spaCy* (Honnibal et al., 2020) and *Trankit* (Van Nguyen et al., 2021)—to train morphosyntactic analysis models. Both toolkits support fine-tuning on local

machines, offer robust performance, and provide user-friendly interfaces suitable even for users with minimal programming experience.

Each parser was trained and evaluated on two datasets: *UD-KSL v1.2* and the *UD-KSL working set*. These datasets include gold-standard UPOS, XPOS, and dependency labels, and were divided into training, validation, and test sets using an 8:1:1 split. The larger *UD-KSL v1.2* set comprised 10,323 training, 1,327 validation, and 1,327 test sentences, while the smaller *UD-KSL working set* contained 2,386 training, 311 validation, and 301 test sentences. Both datasets were provided in fixed and unfixed versions to evaluate the impact of data refinement on model performance.

During training, the toolkits were provided with full morphosyntactic input: lemmatized (i.e., all morphemes parsed in an eojeol text along with UPOS tags, XPOS tags, and dependency labels. During evaluation, the models predicted lemma, UPOS, XPOS, and dependency relations from raw text input. Performance was assessed using standard linguistic metrics: F1-scores for UPOS and XPOS tagging, lemma accuracy for base form identification, and Labeled and Unlabeled Attachment Scores (LAS/UAS) for dependency parsing.

To ensure consistency and isolate the effect of our aligned training data, we used default hyperparameter settings for both toolkits. This allowed us to evaluate model performance under standardized

Dataset	Metric	spaCy			Trankit		
		Unaligned	Aligned	Δ (A–U)	Unaligned	Aligned	Δ (A–U)
UD-KSL v1.2	UPOS	84.55	**90.86**	+6.31	95.74	**96.21**	+0.47
	XPOS	82.54	**82.78**	+0.24	90.25	**90.41**	+0.16
	LEMMA	87.53	87.53	0.00	84.50	**84.51**	+0.01
	UAS	**81.53**	81.29	-0.24	**91.06**	90.83	-0.23
	LAS	**75.08**	74.79	-0.29	88.24	88.24	0.00
UD-KSL working set	UPOS	89.05	**89.28**	+0.23	92.02	**96.06**	+4.04
	XPOS	81.21	**81.68**	+0.47	87.43	**90.94**	+3.51
	LEMMA	86.35	**86.38**	+0.03	76.41	**81.63**	+5.22
	UAS	**79.99**	79.43	-0.56	83.14	**87.81**	+4.67
	LAS	**72.21**	72.02	-0.19	80.07	**84.99**	+4.92

Table 3: Performance metrics from unfixed to fixed configurations. The Δ column indicates the performance change from the unfixed to the fixed configurations for each model.

configurations without introducing optimization-related variance. Neither model was trained on additional data beyond our manually annotated UD-KSL working set. While Trankit leverages multilingual representations from XLM-RoBERTa (Conneau et al., 2020), spaCy's tok2vec model was trained from scratch using only the subword features extracted from our Korean dataset.

4.2 Results

Table 3 summarizes model performance of each toolkit on the two datasets. Our current work mainly inquired into the benefits of UPOS alignments. In the following discussions, we explore the improvements brought by this alignment.

Performance on UPOS tagging Aligning UPOS tags improved the accuracy of both spaCy and Trankit, although the degree of improvement varied across datasets and models. For spaCy, alignments led to a substantial improvement on UD-KSL v1.2 ($\Delta = +6.31$) and a slight increase on the UD-KSL working set ($\Delta = +0.23$). Trankit also benefited from alignments, showing a modest gain in accuracy on UD-KSL v1.2 ($\Delta = +0.47$) and a more notable improvement on the UD-KSL working set ($\Delta = +3.51$). These results suggest that alignment contributes to more accurate UPOS predictions across models and datasets.

Performance on XPOS tagging Similar patterns were observed for XPOS tagging, although improvements varied by model. For spaCy, aligning UPOS tags resulted in marginal gains on both UD-KSL v1.2 ($\Delta = +0.24$) and the UD-KSL working set ($\Delta = +0.47$). In contrast, Trankit

showed clearer benefits for the UD-KSL working set ($\Delta = +3.51$) compared to UD-KSL v1.2 ($\Delta = +0.16$). These results suggest that UPOS alignment may be especially beneficial for XPOS tagging in low-resource settings, where training data is limited, as in the UD-KSL working set.

Performance on dependency parsing The impact of UPOS alignment on dependency parsing varied by model. For spaCy, alignment did not lead to improvements; parsing accuracy slightly declined on both UD-KSL v1.2 (UAS: $\Delta = -0.24$, LAS: $\Delta = -0.29$) and the UD-KSL working set (UAS: $\Delta = -0.56$, LAS: $\Delta = -0.19$). In contrast, Trankit showed clear gains on the working set, with increases in UAS ($\Delta = +4.67$) and LAS ($\Delta = +4.92$), while the effect on UD-KSL v1.2 was negligible (UAS: $\Delta = -0.23$, LAS: $\Delta = 0.00$). These findings indicate that the influence of UPOS alignment on parsing performance was asymmetric, likely shaped by both model architecture and data characteristics. Further research is needed to identify the underlying factors and assess their relative contributions to dependency parsing performance.

Performance by toolkit Clear differences emerged between spaCy and Trankit in terms of the benefits gained from UPOS alignment. Trankit consistently showed greater improvements across tasks, particularly in low-resource settings. This may reflect architectural differences: Trankit leverages a transformer-based model capable of capturing long-distance dependencies and contextual information, while spaCy's tok2vec model relies on subword-level features and more

localized lexical representations.

Performance by dataset size   Data size appeared to influence the effectiveness of the alignment. The smaller dataset benefited substantially more from the alignment, particularly when trained on *Trankit*. This suggests that alignment can serve as a compensatory strategy in low-resource settings by enhancing label consistency. In contrast, the larger dataset—likely benefiting from stronger baseline performance due to more training data—showed smaller gains, indicating diminishing returns from alignment as data availability increases.

Additional finding: Discrepancies in lemmatization performance   Although lemmatization was not a primary focus of this study, our results reinforce Trankit's relatively low lemmatization accuracy, as previously reported by Sung and Shin (2025). We tested whether UPOS alignment might mitigate this issue, but observed no substantial improvement, suggesting that architectural refinements are still needed.

spaCy, which integrates the rule-based morphological analyzer *MeCab* (Kudo, 2005) for Korean, leverages token-level embeddings from its tok2vec layer to capture local morphological patterns while minimizing interference from broader context. In contrast, Trankit's transformer-based seq2seq lemmatizer, adapted from Stanza (Qi et al., 2020), may place undue emphasis on long-distance dependencies, potentially introducing irrelevant context or overfitting—especially when data are limited. Further investigation is needed to validate these hypotheses and explore strategies for improving transformer-based lemmatization for L2 Korean.

5   Conclusion

Building upon prior L2-Korean UD annotation efforts (Sung and Shin, 2023, 2024, 2025), the present work introduced a semi-automatic framework for aligning fine-grained XPOS tags with UPOS tags for (L2-)Korean treebanks. We also augmented the UD-KSL treebank by annotating 2,998 new sentences from an argumentative writing domain. To support reproducibility and promote further research in L2 Korean NLP, all relevant resources have been made publicly available via the UD-KSL treebank: `https://github.com/UniversalDependencies/UD_Korean-KSL/tree/dev`.

We evaluated the effect of XPOS-UPOS align-

ment by training models both with and without alignment across two open-access NLP toolkits. Alignment consistently improved tagging accuracy for UPOS, XPOS, and LEMMA. However, dependency-parsing gains varied by toolkit and dataset size: on the smaller annotated dataset, the transformer-based Trankit showed more pronounced improvements than spaCy; on the larger dataset, alignment yielded minimal parsing gains for both toolkits, although Trankit still outperformed spaCy overall. These results suggest that the alignment enhances tagging robustness, while transformer architectures strengthen contextual parsing. Conversely, spaCy's dictionary-driven hybrid lemmatizer outperformed Trankit in lemma generation, suggesting that integrating lexicon-based methods could further improve lemmatization accuracy. Overall, this semi-automated alignment supports more consistent UPOS annotations and robust morphosyntactic analysis in L2 Korean NLP research.

Limitations

One limitation of the current approach may lie in its level of granularity. While the proposed method adopts a linguistically informed alignment strategy, more nuanced or hierarchical frameworks may be better suited to capturing the full complexity of Korean morphosyntax. In particular, certain constructions that did not lend themselves to straightforward mapping between XPOS and UPOS tags remain underexplored. Additional edge cases beyond those discussed in Section 3.3 warrant further investigation to enhance alignment consistency and coverage.

Another limitation is the continued reliance on human annotators despite the use of automated tools for initial tagging. Variability in annotator expertise and training may affect the consistency and accuracy of annotation outputs.

Acknowledgments

References

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. In *Proceedings*

of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 737–746.

Sungdai Cho and John Whitman. 2022. *The Cambridge handbook of Korean linguistics*. Cambridge University Press.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*.

Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a large annotated corpus of learner english: The nus corpus of learner english. In *Proceedings of the eighth workshop on innovative use of NLP for building educational applications*, pages 22–31.

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, and 1 others. 2019. Towards an italian learner treebank in universal dependencies. In *CEUR workshop proceedings*, volume 2481, pages 1–6. CEUR-WS.

Masaki Eguchi and Kristopher Kyle. 2023. Span identification of epistemic stance-taking in academic written english. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 429–442. Association for Computational Linguistics.

Stefan Th Gries and Andrea L Berez. 2017. Linguistic annotation in/for corpus linguistics. *Handbook of linguistic annotation*, pages 379–409.

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.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python.

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.

Hiroshi Kanayama, Masayasu Muraoka, and Katsumasa Yoshikawa. 2017. A semi-universal pipelined approach to the conll 2017 ud shared task. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 265–273.

Taku Kudo. 2005. Mecab: Yet another part-of-speech and morphological analyzer. *http://mecab. source-forge. net/*.

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 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.

Sunyoung Lee-Ellis. 2009. The development and validation of a korean c-test using rasch analysis. *Language Testing*, 26(2):245–274.

Johanna Mesch and Krister Schönström. 2018. From design and collection to annotation of a learner corpus of sign language. In *8th Workshop on the Representation and Processing of Sign Languages, Miyazaki, Japan, 12 May, 2018*, pages 121–126. European Language Resources Association.

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.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108.

Alla Rozovskaya. 2024. Universal dependencies for learner russian. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17112–17119.

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

Hakyung Sung and Kristopher Kyle. 2024. Annotation scheme for english argument structure constructions treebank. In *Proceedings of The 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 12–18.

Hakyung Sung and Gyu-Ho Shin. 2023. 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.

Hakyung Sung and Gyu-Ho Shin. 2024. Constructing a dependency treebank for second language learners of korean. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3747–3758.

Hakyung Sung and Gyu-Ho Shin. 2025. Second language korean universal dependency treebank v1.2: Focus on data augmentation and annotation scheme refinement. In *Proceedings of the Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025)*.

Minh Van Nguyen, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. 2021. Trankit: A light-weight transformer-based toolkit for multilingual natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 80–90.

A Sejong tagset

Tag	Description	Tag	Description
NNG	Noun, common	EP	Ending, prefinal
NNP	Noun, proper	EF	Ending, closing
NNB	Noun, bound	EC	Ending, connecting
NR	Numeral	ETN	Ending, nounal
NP	Pronoun	ETM	Ending, determinative
VV	Verb, main	XPN	Prefix, nounal
VA	Adjective	XSN	Suffix, noun derivative
VX	Verb, auxiliary	XSV	Suffix, verb derivative
VCP	Copular, positive	XSA	Suffix, adjective derivative
VCN	Copular, negative	XR	Root
MM	Determiner	NF	Undecided (considered as a noun)
MAG	Adverb, common	NV	Undecided (considered as a predicate)
MAJ	Adverb, conjunctive	NA	Undecided
IC	Exclamation	SF	Period, Question, Exclamation
JKS	Case particle, nominative	SE	Ellipsis
JKG	Case particle, prenominal	SP	Comma, Colon, Slash
JKO	Case particle, objectival	SO	Hyphen, Swung Dash
JKB	Case particle, adverbial	SW	Symbol
JKC	Case particle, complement	SS	Quotation, Bracket, Dash
JKV	Case particle, vocative	SH	Chinese characters
JKQ	Case particle, conjunctive	SL	Foreign characters
JX	Case particle, auxiliary	SN	Number

B Gloss

Gloss tags and their definitions are taken from the
Leipzig Glossing Rules.[8]

Gloss	Description
ACC	accusative case
ADN	attributive modifier
ADV	adverbial
CNJ	conjunctive suffix
DECL	declarative ending
DEM	demonstrative
FOC	focus particle
LOC	locative case
NOM	nominative
PROX	proximal demonstrative
RL	relativizer

[8] https://www.eva.mpg.de/lingua/pdf/
Glossing-Rules.pdf