Parser agreement and disagreement in L2 Korean UD: Implications for human-in-the-loop annotation

Hakyung Sung[1] **Gyu-Ho Shin**[2]
[1]Psychology, Rochester Institute of Technology
[2]Linguistics, University of Illinois Chicago
hksgla@rit.edu ghshin@uic.edu

Abstract

We propose a simplified human-in-the-loop workflow for second language (L2) Korean morphosyntactic annotation by leveraging agreement between two domain-adapted parsers. We first evaluate whether parser agreement can serve as a proxy for annotation correctness by comparing it with independent human judgments. The results show strong correspondence between parser and human judgments, supporting the feasibility of semi-automatic L2-Korean UD annotation. Further analysis demonstrates that parser disagreements cluster in linguistically predictable domains such as grammatical-relation distinctions and clause-boundary ambiguity. While many disagreement cases are tractable for iterative model refinement, others reflect deeper representational challenges inherent in parsing and tagging L2-Korean corpora.

1 Introduction

Second language (L2) learner corpora consist of language samples produced by individuals acquiring an L2. Natural language processing (NLP) provides computational methods for extracting linguistic features (e.g., part-of-speech tags, grammatical relations). Together, these fields share the goal of empirically modeling L2 production at scale (Meurers, 2015). However, applying NLP tools to L2 corpora has traditionally been considered problematic for two reasons. First, general-purpose models trained on well-edited native/first-language corpora are assumed to perform poorly on non-canonical learner language (Plank, 2016). Second, this assumption has been difficult to verify due to the lack of learner-language benchmarks.

In response, the Universal Dependencies (UD) framework has emerged as a principled foundation for annotating morphosyntactic features in L2 corpora (Masciolini et al., 2025; Zeman, 2025). UD offers cross-linguistic comparability while maintaining relative simplicity in morphosyntactic annotation and parsing (de Marneffe et al., 2021).

As UD-annotated learner corpora have become available, subsequent work has evaluated morphosyntactic models on such data, yielding mixed results. In high-resource languages such as L2 English, UD-based transformer models achieve over 90% F1 in part-of-speech (POS) tagging and dependency parsing (Kyle and Eguchi, 2024). In relatively under-resourced languages such as L2 Korean, performance varies by layer: morpheme-level tagging reaches F1 88%, while dependency parsing remains substantially lower (LAS 57%) (Sung and Shin, 2025b). Performance also varies by genre and proficiency. For instance, written learner data yield lower parsing accuracy than spoken data in L2 English (Kyle and Eguchi, 2024). In L2 Korean, higher-proficiency learners tend to produce longer and more syntactically complex sentences, which increase parsing difficulty; accordingly, proficiency is negatively correlated with dependency head accuracy ($r = -0.26$) (Sung and Shin, 2025b). Importantly, both studies report improved performance after fine-tuning on L2-annotated data, suggesting that domain adaptation partially mitigates these limitations.

Taken together, for researchers seeking to leverage NLP-based annotations in L2 research, the question is increasingly no longer whether morphosyntactic models can be applied to learner corpora, but how they can be effectively integrated into annotation workflows. Because performance varies across languages, proficiency levels, and genres, fully automatic annotation may remain insufficient. However, automatic pre-annotation paired with human verification offers a practical alternative. In this context, the present study explores a simplified human-in-the-loop workflow for L2-Korean UD annotation, testing whether parser agreement can guide selective human review without compromising annotation reliability.

Proceedings of the 20th Linguistic Annotation Workshop (LAW XX), pages 12–21

2 Related work

2.1 Use of morphosyntactic annotation in L2 learner corpora research

Part-of-Speech tags: Part-of-Speech (POS) tags have been widely employed in learner corpora research. In L2 English, for example, phraseological competence (i.e., the use of semi-/prefabricated expressions) has been examined through the automatic extraction of patterns from POS-tagged corpora (e.g., Granger and Bestgen, 2014). POS tags have also been used to measure lexical richness and disambiguate homographs in L2 Spanish (Díez-Ortega and Kyle, 2024). In the case of L2 Korean, language-specific POS tagsets enabled the representation of fine-grained morphemic distinctions within words (Sung et al., 2024).

Dependency relations: Syntactic information derived from dependency relations has likewise supported diverse corpus-based investigations. For instance, dependency-based phraseological units have been analyzed in L2 Dutch to examine the lexis–grammar interface (Rubin et al., 2025), while dependency representations have been used to assess lexical and syntactic complexity in L2 Russian (Kisselev et al., 2022). In L2 English, prior work examined n-grams within specific dependency relations (e.g., Paquot, 2019), verb–argument constructions and related predicate–argument patterns (e.g., Kyle and Crossley, 2017), and broader measures of syntactic complexity (e.g., Kyle and Crossley, 2018). Similarly, Hao et al. (2024) employed dependency parsing to investigate syntactic complexity in L2-Chinese writing.

2.2 Reliability of morphosyntactic annotation on L2 corpora

Although previous studies (as exemplified in Section 2.1) have reported important empirical findings based on extracted morphosyntactic features, their validity depends in part on annotation reliability. If automatic analyses are inaccurate, resulting conclusions may be compromised. While several studies evaluated the performance of NLP models on L2 corpora (e.g., Berzak et al., 2016), findings have been mixed and often limited in scope. As noted by Kyle and Eguchi (2024), many investigations have focused on isolated components (e.g., selected POS tags or dependency relations) rather than overall morphosyntactic performance. Moreover, earlier studies relied on neural architectures trained primarily on well-edited standard-language data (e.g., news articles), without adaptation to L2 learner language.

A recent advance has been the adoption of domain adaptation techniques, in which annotated L2 treebanks are incorporated into model training to improve annotation quality (Kyle and Eguchi, 2024; Sung and Shin, 2025b). Although effective, this approach presupposes the availability of reliable L2 annotations. This aspect makes it important to examine how such annotations are produced in existing L2 corpora.

2.3 UD annotation practices in L2 corpora

Over the past decade, an increasing number of L2 corpora have been annotated within the UD framework for diverse research purposes (Zeman, 2025). Our review identified eight such corpora to date. Masciolini et al. (2025) provide a detailed comparison, outlining their design characteristics (e.g., modality, size, annotation status) and the strategies adopted to address L2-specific phenomena, including ill-formed or non-canonical constructions.

Here, we examine the annotation methodologies underlying these corpora, focusing on whether morphosyntactic annotation was conducted either fully manually or through semi-automatic procedures (i.e., automatically annotated and then corrected by humans; see Table 1). Most UD-based L2 corpora relied on fully manual annotation, with relatively few adopting automatic approaches supplemented by human correction. While manual annotation supports quality control (e.g., through inter-annotator reliability), it is resource-intensive, difficult to scale, and challenging to replicate consistently across projects and annotator teams. Hybrid approaches that combine automatic processing with human oversight may therefore provide a more efficient and reproducible alternative.

2.4 Human-in-the-Loop annotation via model agreement

Human-in-the-Loop (HITL) machine learning broadly refers to workflows in which human expertise is intentionally integrated into automated systems to guide, validate, or correct model behavior (Mosqueira-Rey et al., 2023). Rather than replacing automation, such approaches strategically combine machine efficiency with human judgment, enhancing scalability while preserving reliability. They have been increasingly adopted in domains where full automation is unreliable or where high-

Language (domain)	Annotation method(s)	Reference
Chinese (written)	Manual	Lee et al. (2017)
English (written)	Manual	Berzak et al. (2016)
English (spoken)	Manual	Kyle et al. (2022)
Italian (written)	Semi-automatic	Di Nuovo et al. (2019)
Korean (written)	Manual; Semi-automatic	Sung and Shin (2024); Sung et al. (2025)
Russian (written)	Manual (single annotator)	Rozovskaya (2024)
Spanish (written)	Manual	Pulido et al. (2025)
Swedish (written)	Semi-automatic	Volodina et al. (2025)

Table 1: Overview of UD annotation practices in L2 learner corpora

stakes decisions require human oversight (Amershi et al., 2014; Holzinger, 2016).

Within the broader HITL taxonomy (Holmberg et al., 2020), active learning represents a prominent paradigm in which models select informative or uncertain instances for human annotation (Settles, 2009). Here, annotators function as oracles, and their feedback is used to iteratively refine the model. Active learning has been widely applied in NLP tasks such as POS tagging, dependency parsing, and text classification to reduce annotation cost while maintaining performance (Ringger et al., 2008).

Beyond uncertainty-based selection, prior work in dependency parsing has shown that simple ensemble strategies (i.e., agreement-based voting across multiple parsers) can produce robust predictions without requiring complex meta-modeling (Surdeanu and Manning, 2010). This suggests that model agreement can serve as a complementary signal of confidence. Building on this idea, the present study leverages parser agreement to guide selective human intervention, using disagreement cases as candidates for targeted review.

3 Experiment

In this exploratory study, we examine a simplified HITL workflow for L2-Korean UD annotation. Specifically, we compare the outputs of two independently fine-tuned parsers, treating agreement as a proxy for reliable annotation and disagreement as a signal for targeted human review. We assess whether such a setup can support more efficient annotation in future L2-Korean corpora. The study addresses the following research questions:

1. Can parser agreement reliably serve as a proxy for human annotation agreement?
2. How much manual correction is required to resolve parser disagreements?
3. Which morphosyntactic categories exhibit disagreement, and how can these patterns inform annotation refinement?

Metric	Stanza	Trankit
LEMMA	95.64	88.84
XPOS	89.72	91.81
UAS	85.53	92.28
LAS	80.36	89.13

Table 2: Performance comparison (F1 scores) of fine-tuned Stanza and Trankit models on the L2K-UD test dataset

3.1 Proposed framework

The proposed annotation framework consists of three steps.

Step 1: Automatic annotation. Two domain-adapted parsers—*Stanza* (Qi et al., 2020) and *Trankit* (Van Nguyen et al., 2021)—were applied.[1] We examined four layers: LEMMA, XPOS, HEAD, and DEPREL.[2] Table 2 reports in-domain performance on the UD-KSL test set (Sung and Shin, 2025a).

Step 2: Cross-model comparison. Parser outputs were compared at the token level. Tokens with identical outputs were provisionally accepted, whereas any disagreement triggered human review.

[1] Stanza is a neural pipeline that performs joint tokenization, POS tagging, lemmatization, and dependency parsing using BiLSTM-based and transition-based components, while Trankit is a transformer-based multilingual pipeline built on XLM-R representations for joint morphosyntactic analysis. Both models were fine-tuned on the UD-KSL training set (Sung and Shin, 2025a), a learner corpus of L2 Korean writing annotated with morpheme-level segmentation, XPOS tags, and dependencies.

[2] UPOS was excluded because it is deterministically derived from XPOS under the current annotation scheme (Sung et al., 2025).

Step 3: Human adjudication. Two trained annotators independently reviewed the flagged tokens. If they assigned identical annotations at the token level, their decision was adopted as the gold label. In cases of disagreement, a third annotator (one of the authors) adjudicated by reviewing both model outputs and the independent annotations to assign the final label.

3.2 Evaluation of parser agreement

Given the exploratory nature of this study, we first collected fully independent annotations from both human annotators for all sentences before restricting review to parser-disagreement cases. This design allowed us to assess whether parser agreement could serve as a proxy for human annotation agreement. Specifically, we measured how often parser agreement coincided with inter-annotator agreement. Under the assumption that alignment is sufficiently strong (i.e., human agreement exceeds 90% within parser-agreement cases), parser-agreement cases could be retained automatically in future annotation rounds, enabling human effort to focus primarily on disagreement cases.

3.3 Pipeline validation

Prior to full-scale annotation of the target corpus, we conducted a small-scale validation experiment to evaluate whether the proposed semi-automatic pipeline introduced annotation noise. We randomly sampled 500 L2-Korean sentences from the KoLLA corpus (Lee et al., 2009), none of which had been used to train the fine-tuned models. Annotation proceeded incrementally. In each round, a batch of 100 sentences was annotated using the framework described in Section 3.1, with human review limited to tokens where the two models disagreed. Adjudicated annotations were then incorporated into the training data, and both models were fine-tuned for 10 epochs on the expanded dataset. After each round, performance was evaluated on the UD-KSL test set to determine whether accuracy improved, stabilized, or declined. This procedure was repeated for five rounds.

Figure 1 presents model performance on the test set across incremental fine-tuning rounds. Overall, performance remained stable: both Stanza and Trankit maintained accuracies of approximately 85% or higher, with no observable decline. These results indicate that the proposed pipeline did not introduce substantial degradation in annotation quality during incremental fine-tuning.

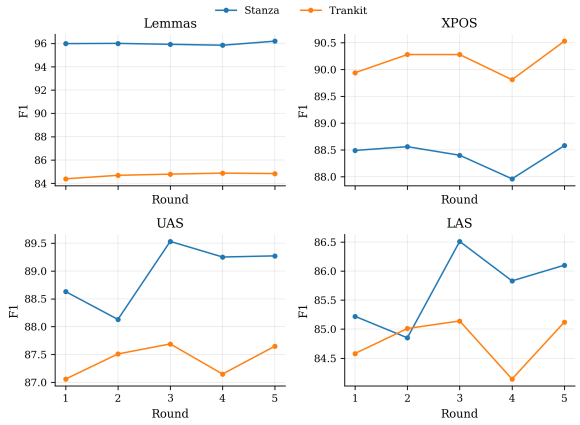

Figure 1: Model performance on the test set across fine-tuning rounds

4 Results

4.1 Annotations

Two trained annotators annotated a total of 2,208 sentences drawn from argumentative essays written by Japanese and English learners of Korean over a one-month period. We adopted the morphosyntactic annotation scheme developed in prior L2 Korean UD annotation work, most recently detailed in Sung et al. (2025).

4.2 Parser agreement as a proxy for human annotation agreement

To evaluate whether parser agreement can serve as a proxy for human annotation agreement, we calculated token-level agreement rates, excluding punctuation. Table 3 summarizes the results. Across all features, the two parsers agreed on 82% of token-level decisions. Within these parser-agreement cases, the human annotators also agreed on 93% of instances, with agreement exceeding 90% for all four features. These findings indicate that parser agreement closely corresponded to independent human agreement, supporting its use as a practical proxy for human annotation agreement in the HITL workflow.

Feature	Model agreement	Human agreement
LEMMA	78.09	96.97
XPOS	84.83	91.97
HEAD	78.62	90.22
DEPREL	88.78	92.55
Average	82.58	92.93

Table 3: Token-level agreement between the two parsers and corresponding human annotator agreement rates (punctuation excluded)

4.3 Human intervention following model mismatch

Based on the agreement results, we next evaluated a workflow in which human annotators intervened only when the two models disagreed.[3] At the token level, 7,994 out of 25,814 tokens (31%) required human correction in at least one morphosyntactic feature.[4]

Feature-level correction counts are provided in Table 4, and feature-level adjudication corrections are summarized in Table 5. A total of 2,019 out of 25,814 tokens (8%) required further modification after initial review. These cases reflect instances in which the two annotators did not converge, necessitating adjudication by a third annotator.

Feature	Human	Total	Rate (%)
LEMMA	4,485	25,814	17.37
XPOS	2,263	25,814	8.77
HEAD	3,798	25,814	14.71
DEPREL	1,713	25,814	6.64

Table 4: Feature-level human corrections following model mismatches

Feature	Fixed	Total	Rate (%)
LEMMA	1,125	25,814	4.36
XPOS	1,461	25,814	5.66
HEAD	581	25,814	2.25
DEPREL	1,153	25,814	4.47

Table 5: Feature-level third-annotator corrections

Overall, these findings demonstrate how parser disagreement can structure a tiered annotation workflow. Restricting human review to model-disagreement cases substantially reduces effort, with nearly 70% of tokens requiring no intervention after alignment. These tokens likely represent morphosyntactic categories that are relatively stable and well captured by the fine-tuned models. Most remaining disagreement cases were resolved through agreement between two annotators, suggesting that they are tractable.

In contrast, the 8% of tokens requiring third-annotator adjudication represent persistent disagreement across both models and trained annotators. These instances often involved structurally complex or potentially ambiguous linguistic units. Although tentative, such residual disagreement at high overall accuracy may reflect not only model limitations but also indeterminacy in linguistic categories or variation in annotation conventions (Manning, 2011).

4.4 Disagreement analysis

To further characterize these disagreement patterns, we analyzed where and why the models diverged.

4.4.1 Distribution of dependency-relation disagreements

We first conducted a focused analysis of disagreement patterns in the dependency-relation (DEPREL) layer, as dependency-relation labeling exhibited relatively high disagreement rates. Disagreements were classified according to the primary syntactic decision involved: (1) grammatical relation identification, (2) clause-boundary and clause-type differentiation, (3) discourse-level structural organization, and (4) modifier attachment.

- **Grammatical relation identification**: This category captures instability in assigning core grammatical relations (e.g., subject, object, oblique), including contrasts such as `nsubj–obj`, `nsubj–obl`, and `obj–obl`.
- **Clause boundary**: This category reflects uncertainty in clause typing and hierarchical embedding, including distinctions among adjectival (relative), adverbial, and complement clauses (e.g., `acl–advcl`, `advcl–ccomp`, and `advcl–root`).
- **Discourse-level organization**: This category involves higher-level decisions at the syntax–discourse interface, such as coordination scope, root status, and left dislocation (e.g., topic-marked elements). Recurrent contrasts included `dislocated–nsubj`, `root–conj`, and `conj–advcl`.
- **Modifier attachment**: This category captures ambiguity in determining the structural status or attachment site of modifiers, with contrasts such as `amod–acl` and `nmod–obl`.

Table 6 summarizes mismatch frequencies across categories. To illustrate these patterns more concretely, representative examples from the annotated texts are provided below.[5]

[3] Prior to analysis, tokenization mismatches were resolved to ensure proper token-level alignment. All reported agreement and intervention rates are based on the aligned data.

[4] Because a single token may require correction in multiple features, counts are not mutually exclusive.

[5] Sentences have been streamlined for clarity.

Mismatch type	Count
Grammatical relation	263
Clause boundary	206
Discourse / structure	235
Modifier attachment	152

Table 6: Major categories of dependency-relation disagreements between the two parsers. Counts indicate the total number of mismatches in each category.

Grammatical-relation ambiguity (e.g., `nsubj–obj`, `nsubj–obl`) accounts for the largest share of mismatches. These alternations typically arise when case-marking is underspecified or omitted, obscuring whether a nominal is analyzed as subject, object, or oblique. For example, when nominative (Figure 2) or accusative case markers (Figure 3) are dropped, a preverbal noun phrase may be ambiguously interpreted, leading to divergent grammatical relation identifications across parsers.

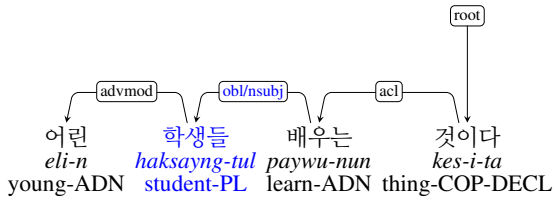

Figure 2: Grammatical-relation ambiguity under case marker omission 1. The nominal 학생들 'students' was tagged as either `obl` or `nsubj` across the parsers; contextual interpretation favors `nsubj`. (Translated as '(It) is that young students learn.')

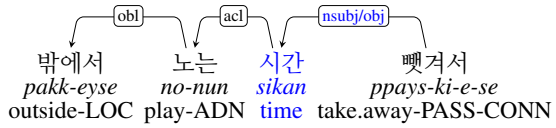

Figure 3: Grammatical-relation ambiguity under case marker omission 2. The nominal 시간 'time' received conflicting tags (`nsubj` vs. `obj`); contextual interpretation favors `obj`. (Translated as '(When) time spent playing outside is taken away.')

Clause-boundary ambiguity (e.g., `acl–advcl`, `advcl–ccomp`) reflects uncertainty in clause typing and hierarchical embedding. Figure 4 illustrates a case in which a clause can be analyzed either as an adnominal modifier (top) or as a subordinate clause within the predicate domain (bottom).

Compared to grammatical-relation ambiguity, clause-boundary ambiguity poses greater challenges for two reasons. First, the polyfunctionality

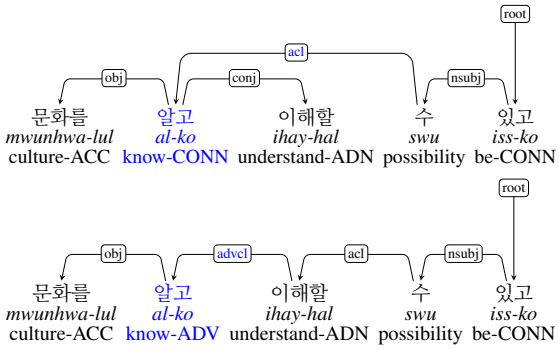

Figure 4: Clause-boundary ambiguity 1. The connective verb 알고 ('know-CONN/ADV') is ambiguous. In the first analysis, the clause is treated as an adnominal modifier (`acl`); in the second, it is analyzed as an adverbial clause (`advcl`); the appropriate annotation cannot be determined from the sentence in isolation. (Translated as Top: '[One] can know and understand the culture.' Bottom: 'After knowing the culture, [one] can understand it.')

of the connective 고 *-ko* in Korean frequently triggers disagreement, as its interpretation depends on discourse-semantic cues rather than overt syntactic marking. Second, clause-typing uncertainty often interacts with higher-level structural mismatches (e.g., `conj–advcl`), corresponding to the third category of disagreement. Such cases therefore require adjudication informed by broader sentential or discourse context.

Meanwhile, not all clause-type disagreements arise from context-dependent ambiguity. In `advcl–ccomp` mismatches, some cases instead appear to reflect model difficulty in learning complement structures headed by the -지 (*-ci*) complementizer (Figure 5).

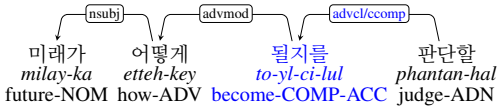

Figure 5: Clause-type disagreement 2. The embedded clause 될지를 ('how [it] will become') functions as a clausal complement (`ccomp`) of the matrix predicate 판단할 ('judge'), a pattern that one parser consistently failed to capture; morphosyntactic cues (i.e., the complementizer 지 and accusative marker 를) favor a `ccomp`.(Translated as: '[One] can judge how the future will turn out.')

Discourse-related ambiguity (e.g., `dislocated–nsubj`) reflects instability at the syntax-discourse interface, particularly in topic-prominent constructions where left-dislocated elements may be misanalyzed as canonical subjects.[6]

[6]Examples of structural mismatches (e.g., `root–conj`)

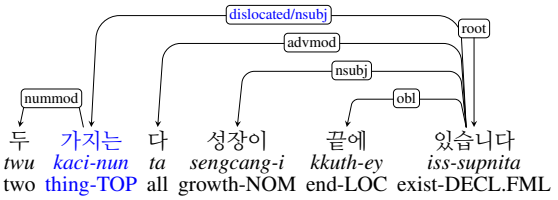

Figure 6: Discourse-level misanalysis. The topic-marked noun phrase 가지는 (‘things-TOP’) functions as `dislocated`, a pattern that one parser consistently failed to capture. (Translated as ‘As for the two things, both ultimately result in growth.’)

Finally, modifier attachment ambiguity (e.g., `amod–acl`, `nmod–obl`) reflects uncertainty in hierarchical scope, particularly when linear proximity does not clearly determine attachment. As illustrated in Figure 7, an adnominal form can be analyzed either as a lexical adjectival modifier (`amod`) or as a reduced relative clause (`acl`). Similarly, Figure 8 shows that a locative phrase can attach either to a noun phrase (`nmod`) or to the predicate as a clausal oblique (`obl`), depending on its interpreted scope. Such modifier-attachment ambiguity is likewise context-dependent, requiring broader interpretive information beyond local morphosyntactic cues.

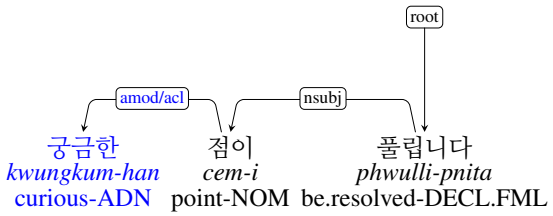

Figure 7: Modifier attachment ambiguity 1. The form 궁금한 (‘curious-ADN’) can function either as a lexical adjectival modifier (`amod`) or as a reduced relative clause (`acl`) modifying *cem* (‘point’). (Translated as ‘The questions are resolved.’)

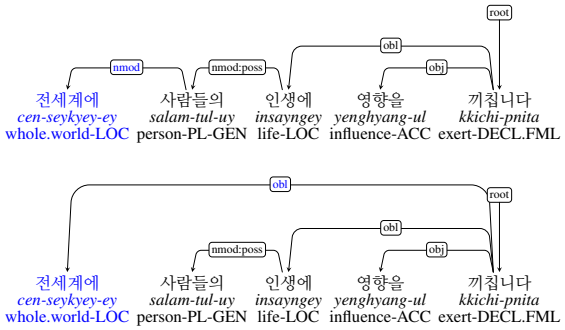

Figure 8: Modifier attachment ambiguity in a locative phrase. The locative 전세계에 (‘whole.world-LOC’) attaches either to 사람들 (‘people’) as `nmod` or to 끼칩니다 (‘exert’) as `obl`. Despite the locative marker (-에), the first analysis yields a possessive-like reading (‘people of the whole world’). (Translated as Top: ‘[It] affects the lives of people in the whole world.’ Bottom: ‘In the whole world, [it] affects people’s lives.’)

Rank	Stanza	Trankit	Count
1	NNG+**JKS**	NNG+**JKC**	54
2	NNG+**JC**	NNG+**JKB**	48
3	**NNG**+XSA+ETM	**XR**+XSA+ETM	40
4	**VV**+ETM	**NV**+ETM	28
5	**NNG**+VCP+EF	**NNB**+VCP+EF	27
6	**VV**+EC	**NV**+EC	25
7	–	JX	23
8	MAG	NNG	23
9	**VV**+ETM	**VX**+ETM	22
10	**NNG**+JKB	**NF**+JKB	22
11	**VX**+EC	**VV**+EC	21
12	**NNG**+XSA+EC	**XR**+XSA+EC	20
13	**NF**+JKO	**NNG**+JKO	20
14	NNG	MAG	19
15	NNG+NNG+NNG+**JKG**	NNG+NNG+NNG+**JX**	18
16	**NNG**+JKO	**NF**+JKO	18
17	VV+ETM+NNB	VV+ETM+NNB+**JX**	18
18	VV+EC+VX	VV+EC+VX+**EC**	17
19	**VA**+EC	**VV**+EC	17
20	VV+EC+**VX**	VV+EC	17

Table 7: Twenty most frequent XPOS disagreement pairs between Stanza and Trankit. Counts indicate the number of tokens assigned different morpheme-level POS analyses.

4.4.2 Morphological-level disagreements

Table 7 presents the twenty most frequent morpheme-level XPOS mismatches. Similar to dependency relations, these disagreements form recurrent patterns rather than occurring randomly. First, many mismatches involved case particles (e.g., `JKS`, `JKB`, `JKO`, `JKC`, `JX`, `JKG`). These contrasts often reflect functional ambiguity, particularly in distinguishing structural case markers from auxiliary or semantic/discourse particles. Such morphological ambiguity closely parallels the dependency-level ambiguities observed in grammatical-relation identification and discourse structure.

Second, high-frequency mismatches often arose from differences in lexical decomposition and root identification (e.g., `NNG` vs. `XR`; `NNG` vs. `NF`; `VV` vs. `NV`). Two issues are implicated. First, the distinction between common nouns (`NNG`) and lexical roots (`XR`) is not always clear-cut in Korean, as root classification can be inherently ambiguous. Second, this ambiguity is further amplified in learner language, where non-canonical forms are frequent. In our annotation scheme, tags such as `NF`, `NV`, and `NA` mark ill-formed or irregular forms. While parsers may recover intended lexical items for recurring spelling

are not presented separately, as they typically co-occur with clause-boundary ambiguities discussed above.

errors through dictionary-based matching, novel or idiosyncratic errors often lack lexical support, resulting in divergent analyses. These patterns underscore the need for more systematic approaches to learner-specific morphological variation.

Finally, some disagreements involved segmentation differences, including the insertion or omission of functional morphemes (e.g., additional JX or EC). These cases reflect variation in morphological parsing strategies rather than simple tagging errors.

5 Conclusion

The purpose of this study was to evaluate a simplified HITL workflow for L2-Korean UD annotation. The findings provide three main implications, which may be relevant for researchers working on morphosyntactic annotation in learner corpora.

First, across all annotation features, the two independently domain-adapted parsers agreed on 82% of token-level decisions. Within these consensus cases, human annotators also agreed on 93% of instances. This strong alignment suggests that parser agreement reliably predicts correspondence with independent human judgments. For scalable L2 annotation, parser consensus may therefore serve as an effective filtering mechanism, substantially reducing the need for exhaustive manual verification. In addition, as noted by one reviewer, this binary agreement approach could be extended within an ensemble framework (e.g., Surdeanu and Manning, 2010), which may enable more robust consensus estimation.

Second, despite high overall agreement, 31% of tokens required human review in at least one feature, and 8% required adjudication after initial correction. These findings suggest that morphosyntactic disagreement operates at multiple levels. Some cases are readily resolved through annotator agreement and are amenable to iterative model refinement, whereas others reflect deeper representational challenges in assigning L2-Korean forms to discrete morphosyntactic categories.

Third, parser disagreements clustered in linguistically predictable domains rather than occurring randomly. Our analysis showed that many involved argument-role distinctions and complement structures headed by complementizers, suggesting that targeted sampling and focused retraining could improve performance. In contrast, clause-boundary and modifier-attachment ambiguities were often context-dependent, indicating that some disagreements cannot be resolved through local morphosyntactic cues alone and may require broader contextual modeling. At the morphological level, frequent mismatches involved root identification, learner-specific spelling variation, and segmentation differences. These patterns highlight the need for systematic strategies to handle learner-generated forms, particularly when such non-standard forms are not attested in the training data or lexical resources.

In conclusion, we examined whether parser agreement can serve as a principled triaging mechanism in L2 annotation. The results point to its potential, while also highlighting the multi-level nature of morphosyntactic disagreement. Distinguishing between tractable modeling limitations and deeper representational ambiguities remains important for achieving efficient yet reliable analysis of learner language.

Limitations

First, the dataset consists exclusively of argumentative writing by adult L2-Korean learners. Because parser performance may vary by proficiency, age, genre, and language background, the generalizability of these findings is limited.

Second, although we briefly noted issues related to spelling-error tags, this study did not systematically examine learner-specific morphological variation. Developing principled approaches to modeling such variation is therefore an important direction for future research.

Third, the proposed workflow reflects a simplified HITL design rather than a fully integrated, interactive system. For instance, it did not incorporate dynamic confidence estimation, active learning, or real-time model updating. Research in this line would benefit from incorporating these aspects into the HITL design.

Acknowledgments

This study was supported by the 2024 Korean Studies Grant Program of the Academy of Korean Studies (AKS-2024-R-012). The authors gratefully acknowledge Youkyung Sung and Chanyoung Lee for their contributions to manual annotation, and Jeong Eun Shin for providing the data.

References

Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *AI magazine*, 35(4):105–120.

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, Berlin, Germany. Association for Computational Linguistics.

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, and Manuela Sanguinetti. 2019. Towards an italian learner treebank in universal dependencies. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*, pages 151–158, Bari, Italy. CEUR Workshop Proceedings.

María Díez-Ortega and Kristopher Kyle. 2024. Measuring the development of lexical richness of l2 spanish: A longitudinal learner corpus study. *Studies in Second Language Acquisition*, 46(1):169–199.

Sylviane Granger and Yves Bestgen. 2014. The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study. *International Review of Applied Linguistics in Language Teaching*, 52(3):229–252.

Yuxin Hao, Xuelin Wang, Shuai Bin, Qihao Yang, and Haitao Liu. 2024. How syntactic complexity indices predict chinese l2 writing quality: An analysis of unified dependency syntactically-annotated corpus. *Assessing Writing*, 61:100847.

Lars Holmberg, Paul Davidsson, and Per Linde. 2020. A feature space focus in machine teaching. In *2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pages 1–2.

Andreas Holzinger. 2016. Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain informatics*, 3(2):119–131.

Olesya Kisselev, Rossina Soyan, Dmitrii Pastushenkov, and Jason Merrill. 2022. Measuring writing development and proficiency gains using indices of lexical and syntactic complexity: Evidence from longitudinal russian learner corpus data. *The Modern Language Journal*, 106(4):798–817.

Kristopher Kyle and Scott Crossley. 2017. Assessing syntactic sophistication in l2 writing: A usage-based approach. *Language Testing*, 34(4):513–535.

Kristopher Kyle and Scott A Crossley. 2018. Measuring syntactic complexity in l2 writing using fine-grained clausal and phrasal indices. *The Modern Language Journal*, 102(2):333–349.

Kristopher Kyle and Masaki Eguchi. 2024. Evaluating nlp models with written and spoken l2 samples. *Research Methods in Applied Linguistics*, 3(2):100120.

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, Seattle, Washington. Association for Computational Linguistics.

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)*, volume 135 of *Linköping Electronic Conference Proceedings*, pages 67–71. Linköping University Electronic Press.

Sun-Hee Lee, Seok Bae Jang, and Sang-Kyu Seo. 2009. Annotation of korean learner corpora for particle error detection. *Calico Journal*, 26(3):529–544.

Christopher D Manning. 2011. Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In *International conference on intelligent text processing and computational linguistics*, pages 171–189. Springer.

Arianna Masciolini, Aleksandrs Berdičevskis, Maria Irena Szawerna, and Elena Volodina. 2025. Annotating second language in universal dependencies: a review of current practices and directions for harmonized guidelines. In *Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025)*, pages 153–163.

Detmar Meurers. 2015. Learner corpora and natural language processing. In Sylviane Granger, Gaëtanelle Gilquin, and Fanny Meunier, editors, *The Cambridge Handbook of Learner Corpus Research*, pages 537–566. Cambridge University Press, Cambridge.

Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. 2023. Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review*, 56(4):3005–3054.

Magali Paquot. 2019. The phraseological dimension in interlanguage complexity research. *Second language research*, 35(1):121–145.

Barbara Plank. 2016. What to do about non-standard (or non-canonical) language in nlp. *arXiv:1608.07836 [cs]*.

Emiliana Pulido, Robert Pugh, and Zoey Liu. 2025. I speak for the árboles: Developing a dependency treebank for spanish l2 and heritage speakers. In

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 814–822.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Eric Ringger, Marc Carmen, Robbie Haertel, Kevin Seppi, Deryle Lonsdale, Peter McClanahan, James Carroll, and Noel Ellison. 2008. Assessing the costs of machine-assisted corpus annotation through a user study. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco. European Language Resources Association (ELRA).

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, Torino, Italia. ELRA and ICCL.

Rachel Rubin, Bram Bulté, Magali Paquot, and Alex Housen. 2025. Exploring complexity at the lexis-grammar interface: Diversity and sophistication of verb-argument structures in l2 dutch writing. *Journal of Second Language Writing*, 67:101183.

Burr Settles. 2009. Active learning literature survey. (1648). Computer Sciences Technical Report.

Hakyung Sung, Sooyeon Cho, and Kristopher Kyle. 2024. An empirical evaluation of lexical diversity indices in l2 korean writing assessment. *Language Assessment Quarterly*, 21(2):159–180.

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, Torino, Italia. ELRA and ICCL.

Hakyung Sung and Gyu-Ho Shin. 2025a. Second language korean universal dependency treebank v1. 2: Focus on data augmentation and annotation scheme refinement. In *Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025)*, pages 13–19.

Hakyung Sung and Gyu-Ho Shin. 2025b. Towards robust morphosyntactic analysis of L2 korean: Evaluating and fine-tuning a korean language model. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 24(4):1–20.

Hakyung Sung, Gyu-Ho Shin, Chanyoung Lee, You Kyung Sung, and Boo Kyung Jung. 2025. UD-KSL treebank v1.3: A semi-automated framework for aligning XPOS-extracted units with UPOS tags. In *Proceedings of the 19th Linguistic Annotation Workshop (LAW-XIX-2025)*, pages 115–125, Vienna, Austria. Association for Computational Linguistics.

Mihai Surdeanu and Christopher D Manning. 2010. Ensemble models for dependency parsing: cheap and good? In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 649–652.

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.

Elena Volodina, Arianna Masciolini, Beáta Megyesi, Julia Prentice, Lisa Rudebeck, Gunlög Sundberg, and Mats Wirén. 2025. SweLL with pride: How to put a learner corpus to good use. *Huminfra Handbook (forthcoming)*.

Daniel Zeman. 2025. Corpus-based language comparison: From morphology to dependencies and beyond. *Estudos Linguísticos (São Paulo. 1978)*, 54(1):259–275.

A Gloss abbreviations

Abbreviations used in interlinear glosses follow standard Leipzig conventions.

Abbreviation	Meaning
ACC	Accusative
ADN	Adnominal
ADV	Adverbial
CONN	Connective ending
COP	Copula
DECL	Declarative
FML	Formal speech level
LOC	Locative
NOM	Nominative
PASS	Passive
PL	Plural
TOP	Topic