Towards L2-friendly pipelines for learner corpora: A case of written production by L2-Korean learners

Hakyung Sung
Department of Linguistics
University of Oregon
hsung@uoregon.edu

Gyu-Ho Shin
Department of Linguistics
University of Illinois Chicago
gyuhoshin@gmail.com

Abstract

We introduce the Korean-Learner-Morpheme (KLM) corpus, a manually annotated dataset consisting of 129,784 morphemes from second language (L2) learners of Korean, featuring morpheme tokenization and part-of-speech (POS) tagging. We evaluate the performance of four Korean morphological analyzers in tokenization and POS tagging on the L2-Korean corpus. Results highlight the analyzers' reduced performance on L2 data, indicating the limitation of advanced deep-learning models when dealing with L2-Korean corpora. We further show that fine-tuning one of the models with the KLM corpus improves its accuracy of tokenization and POS tagging on L2-Korean dataset.

1 Introduction

The use of learner corpora has played a crucial role in understanding language learners' developmental aspects (e.g., Biber et al., 2011; Ellis and Ferreira-Junior, 2009; Gablasova et al., 2017). With the recent advancement of computational methods and techniques, automatic processing of learner corpora (together with sizeable datasets) is gaining momentum for a better understanding of the properties of learner language (e.g., Bestgen and Granger, 2014; Kyle and Crossley, 2017; Lu, 2010).

Despite the increasing interest in this approach, we identify two major caveats in the current research practice. One is the sampling bias towards dominant/hegemonic viewpoints and discourse, especially centering around a limited range of languages and language-usage contexts (e.g., L2 English) (c.f., Bender et al., 2021). This poses a threat to linguistic diversity, equity, and inclusion in the field, as well as weakening the generalizability of previous findings to other (and lesser-studied) languages.

The other caveat concerns the degree to which first language (L1)-based automatic processing pipelines work for L2 data. Indeed, a line of research has questioned the reliability of currently existing parsing/tagging models, which are trained and tested exclusively on the basis of L1 data, when applied to L2 corpora (e.g., Kyle, 2021; Meurers and Dickinson, 2017). This is because these L1-oriented models may not fully account for the characteristics of learner language, including spacing/spelling errors and novel combinations of words and phrases. These factors may negatively impact the performance of L1-based tools when analyzing linguistic features of L2 corpora, thus necessitating empirical investigation.

In this study, we aim to address these caveats by developing a sizable L2-Korean corpus, featuring enhanced morpheme tokenization and POS tagging of the open-access L2-Korean corpus dataset, which comprises 129,784 morphemes (7,527 sentences). Using this dataset, we evaluate the morpheme tokenization and POS-tagging accuracy of two language-general parsers incorporating cutting-edge algorithms (*Stanza, Trankit*) and two Korean-specific parsers commonly used by researchers in Korean studies (*Kkma, Komoran*).

This paper is structured as follows: We discuss the significance of morphological analysis in Korean studies and review relevant L2-Korean applied research. Next, we outline the annotation process employed in our study. We then elaborate on our methodology for evaluating the performance of the morpheme analyzers on our dataset, using an L1 corpus as a reference. Following this, we present a comprehensive analysis of the overall performance, including detailed comparisons across different proficiency levels and POS tags, as well as a re-evaluation of performance after training the L2 annotated corpus. Finally, we summarize our findings and propose future directions.

Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), pages 72–82
July 13, 2023 ©2023 Association for Computational Linguistics

2 Background

2.1 Linguistic properties of Korean

Korean, a language typologically distinctive from the major languages studied in the field (specifically English) is characterized by its agglutinative nature and `Subject Object Verb` word order. It features overt case-marking and active suffixation, allowing scrambling and omission of sentential components contingent upon contexts (Sohn, 1999). These characteristics collectively pose challenges to automatic processing of (L2-)Korean corpora (Shin and Jung, 2021), particularly for tokenization and POS tagging systems that are not entirely rely on white-space units such as English words (McDonald et al., 2013). Previous studies (e.g., Choi and Palmer, 2011; Park et al., 2013) have addressed word-level representation issues in Korean by utilizing linguistically motivated rules, highlighting the fact that words in Korean comprise both lexical and functional *morphemes* (i.e., the smallest meaningful unit of language). This necessitates considering morpheme-level parsing and tagging when handling Korean corpora automatically (Chen et al., 2022).

2.2 Application of morpheme tokenizers and POS taggers in L2-Korean research

In spite of the language-specific challenges associated with Korean for conducting automatic text processing, researchers have increasingly attempted to apply NLP techniques to L2-Korean research. Notably, however, most studies have not provided sufficient information about the tools they used or the reliability of the parsers/taggers for L2-text processing. An overview of this research practice is outlined below.

Error analysis: Kim et al. (2016) investigated the types of frequent errors from a sizable L2-Korean writing data (*n*=500) and identified rules for searching syntactic patterns by using a POS tagger (type not reported). Lee et al. (2016) proposed an automatic error-detection scheme for L2-Korean production involving functional morphemes (e.g., particles) in combination with a POS tagger (type not reported).

Lexico-grammatical token measurement: Lim et al. (2022) proposed an automated writing evaluation system by employing a transformer-based multilingual model and XLM-RoBERTa. They used a POS tagger (type not reported) to measure the number of morphemes as one of the complexity features of learner writing. Nam and Hong (2014) collected L2-Korean spoken data from storytelling, communications, and natural conversations and annotated the data based on the Sejong tag set. They employed a POS tagger (type not reported) to compare the number of particles across multiple proficiency groups.

Morpheme/construction extraction: Jung (2022) and Shin and Jung (2022) investigated the distribution of Korean particles in L2-Korean textbooks. Using *UDpipe* as a tagger, they developed a pipeline for automatically extracting the target particles. Likewise, Shin and Jung (2021) demonstrated how Korean passive constructions could be (semi-)automatically identified by using the same tagger and pipeline developed above.

Text similarity analysis: Cho and Park (2018) used various morphological analyzers (*Kkma*, *Okt*, *Hannanum*, and *Komoran*) to explore the text similarity (based on TF-IDF) of the writings produced by sixteen different L2-Korean learners.

3 Dataset

The Korean-Learner-Morpheme (KLM) corpus, as it currently stands, comprises 129,784 morphemes (67,284 *eojeols*, which are sequences of Korean characters separated by white-spaces) with morpheme tags grounded in the Sejong tag set (Appendix A). This corpus was sourced from the Kyung Hee Korean learner written corpus collected by Park and Lee (2016). The corpus encompasses data on classroom proficiency levels (ranging from 1 to 6 as a proxy for learner proficiency), nationality, gender, and writing topics. To create our dataset, we randomly extracted a total of 600 texts from the original corpus, with each proficiency level represented by 100 texts.

Despite the presence of morpheme tokenization and POS tags in the original corpus, several issues prevented its direct use for evaluation purposes, which ultimately led us to conduct manual annotations. First, without gold annotations for the data, we were not able to determine the accuracy of the automatic POS tagger (i.e., ESPRESSO) that Park and Lee (2016) used for morphological analysis. Additionally, we were uncertain whether the annotation scheme in the original corpus had been thoroughly tested, taking into account the language-specific properties of Korean. Second, we were unsure how the characteristics of learner language (e.g., spelling/spacing errors), which were not clearly indicated in the original corpus, were documented in the annotations (e.g., whether they were corrected or neglected during the automatic analysis). On top of these issues, the formatting proved difficult to process the data automatically.

To create our corpus, we first reformatted the texts into CoNLL-U format, following the Universal Dependencies (UD) formalism (c.f., Nivre et al., 2020). To ensure the metadata in the original dataset, we associated the respective # text_id attribute with the extracted metadata (e.g., # text_id = A100000_v01_중국_남자_사진기 빌리기) and incorporated the # sent_id attribute in an incremental manner (e.g., # sent_id = A100000_v01_중국_남자_사진기 빌리기_1, assigned to the first sentence of the text) for data management. Sentence- and eojeol- level segmentations were done using Stanza[1] as a tokenizer.

3.1 Annotation procedure

The corpus was annotated by two native Korean speakers: the first author of the paper and a graduate student who majored in Korean during their undergraduate studies. Before annotating the sentences, both annotators familiarized themselves with the Sejong tag set, its tokenization scheme[2], and the annotation guidelines from previous studies related to Korean UD guidelines (e.g., Chun et al., 2018; Park and Tyers, 2019) through two training sessions. The annotation process was carried out in the following steps: (1) the two annotators annotated 100 texts individually (both morpheme tokenization and POS tagging); (2) the annotators reviewed and discussed their disagreements; (3) if a disagreement was not resolved, the third annotator, the second author of this paper, reviewed the problematic tokens and POS tags and provided annotations; and (4) the third annotator commented on the entire annotation results, which were then discussed by the two main annotators before starting the next annotation round.

Although the annotators referred to previous studies for parsing/tagging guidance, there were a few instances in which making decisions proved challenging. Below are the major cases that we discussed, with the purpose of consistent annotations and better evaluation of morpheme tokenizers/taggers of interest. The full tagging guidelines and examples can be accessed here for related future projects: https://github.com/NLPxL2Korean/Korean_Learner_Morpheme_corpus.

Causative and passive markers: Causative and passive voices are often indicated by the voice markers (*-i/hi/li/ki/wu/kwu/chwu-* for morphological causative; *-i/hi/li/ki-* for suffixal passive; *-e/a ci-* for periphrastic passive; Sohn, 1999). These morphemes, when attached to a root, form causative or passive verbs and lead to changes in valence (i.e., the number of arguments controlled by a predicate in a clausal construction). We parsed all relevant morphemes and assigned them XSV (Suffix, verb derivative) POS tags (e.g., *mek+ta* "to eat" VV (Verb, main)+EF (Ending, closing); *mek+hi+ta* "to be eaten" VV+XSV+EF).

Auxiliary verbs: Verbs such as *iss-* "to be/exist/have", *ha-* "to do", and *toy-* "to become" function as both main verbs and auxiliary verbs. As main verbs, they typically operate independently, representing concepts of existence, activity, or possession (e.g., *ku-nun cha-ka iss-ta* "He has a car"). In these instances, we assigned a VV (Verb, main) tag. Conversely, when serving as auxiliary verbs, they work in conjunction with a main verb to convey grammatical meanings, such as continuous or progressive actions (e.g., *ku-nye-nun chayk-ul ilk-ko iss-ta* "She is reading a book"). In these cases, we assigned a VX (Verb, auxiliary) tag.

Copula, positive: The copula (-*i*) is a grammatical element that links the subject of a sentence with a predicate, often conveying a positive meaning (VCP). One complexity in parsing morphemes arises when the copula is combined with the ending *-lanun* in a compound form. This combination links the subject of a sentence to a noun or descriptive phrase while adding a nuance of specification, identification, or definition (translated as "called," "named," or "known as" in English). Interestingly, in some cases, the copula may be hidden, requiring the addition of *-i* before the ending *-lanun* to ensure accurate parsing (e.g.,

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

[2] publicly available from KoNLPy website https://konlpy.org/ko/v0.4.4/morph/

swukcey-lanun "(the thing) called homework" → *swukcey+i+lanun*, NNG+VCP+ETM).

Spelling errors: Instead of judging or omitting the annotation of misspelled words based on annotators' subjective interpretations, we opted for assigning three relevant tags from the Sejong tag set: NA (Undefined), NF (Undefined, but considered a noun), and NV (Undefined, but considered a verb). Following this annotation method, a total of 2,289 errors were marked (NA: 738, NF: 1,290, NV: 261).

3.2 Annotation review

Table 1 presents (1) the number and percentage of refined tokens and tags, and (2) the number and percentage of overall agreement rates between the two annotators in creating the corpus. The term "refined" tokens and tags refers to tokens and tags which were manually revised by the annotators against the tokens and tags used in the original corpus. Note that morpheme tokenization/POS tagging is not always a binary decision in Korean, as the morpheme boundary can be ambiguous. Therefore, we measured the reliability by calculating the ratio of the number of agreement items to the total number of tokens/tags, rather than by calculating Cohen's Kappa scores. Overall, the results indicate a high level of agreement between the annotators in both tasks.

Category	**Token**	**Tags**
# of refinement	19,481	20,987
% of refinement	15.01	16.17
# of agreement	128,890	128,243
% of agreement	99.31	98.81
Total		129,784

Table 1: Summary of annotation results

4 Analysis

4.1 Reference L1 corpus

We used the Google Korean Universal Dependency Treebank (UD Korean GSD) as a reference L1 corpus to establish a baseline for calculating accuracy. This dataset originally comprises around 6,000 sentences sourced from online blogs and news produced by Korean native speakers. The sentences were then annotated according to the UD guidelines (McDonald et al., 2013) and later enhanced by implementing a more refined morpheme tokenizations (Chun et al., 2018). For the purposes of this study, we employed 989 sentences from the UD Korean GSD test set.

4.2 Morphological analyzers

We employed four open-access morphological analyzers. They are based on various computational algorithms, ranging from statistical models[3], which have been widely used by L2-Korean researchers (e.g., Kkma[4], Komoran), to deep-learning models such as Stanza[1] and Trankit[5].

5 Results and Discussion

5.1 Overall performance

Table 2 displays the overall F1 scores[6] of the morphological analyzers for the L2 (target) and L1 (reference) datasets[7]. Figure 1 presents by-proficiency-level performance per analyzer.

	Token		**Tag**	
Analyzer	**L2**	L1	**L2**	L1
Stanza	**0.89**	0.92	**0.86**	0.93
Trankit	0.81	0.85	0.80	0.88
Kkma	0.86	0.88	0.80	0.81
Komoran	**0.89**	0.92	**0.86**	0.86

Table 2: F1 scores (overall)

We draw three main observations. First, all the analyzers exhibited reduced performance on the

[3] KoNLPy as an interface, see Park and Cho, 2014

[4] Kkma employs a more extensive tag set (52 tags) compared to the other three analyzers (45 tags from the Sejong tag set), necessitating an additional step for tag standardization prior to evaluating accuracy.

[5] `https://github.com/nlp-uoregon/trankit/`

[6] It is often the case that True Negatives apply to a binary classification problem in which tokenization is clearly based on white-space, such as English. Notably, tokenization in Korean does not always fall into binary classification because of unclear morpheme boundaries. We thus calculated the F1 scores using True Positives (the number of correct matches between the predicted and gold standard annotations), False Positives (the number of predicted annotations that do not match the gold standard annotations), and False Negatives (the number of gold standard annotations that do not match the predicted annotations). We acknowledge that our approach here should be further verified by future research with providing a *Perfect* matrix (Raman et al., 2022).

[7] Subtle differences in output representation arise when comparing the performance of Stanza/Trankit to that of Kkma/Komoran. Stanza/Trankit utilize word-level units based on white-space, facilitating a robust comparison between annotated and predicted tags, as their outputs are structured around these word-level units. On the other hand, Kkma/Komoran display morphemes without maintaining original word boundaries, necessitating the evaluation of accuracy strictly on a sentence-unit level.

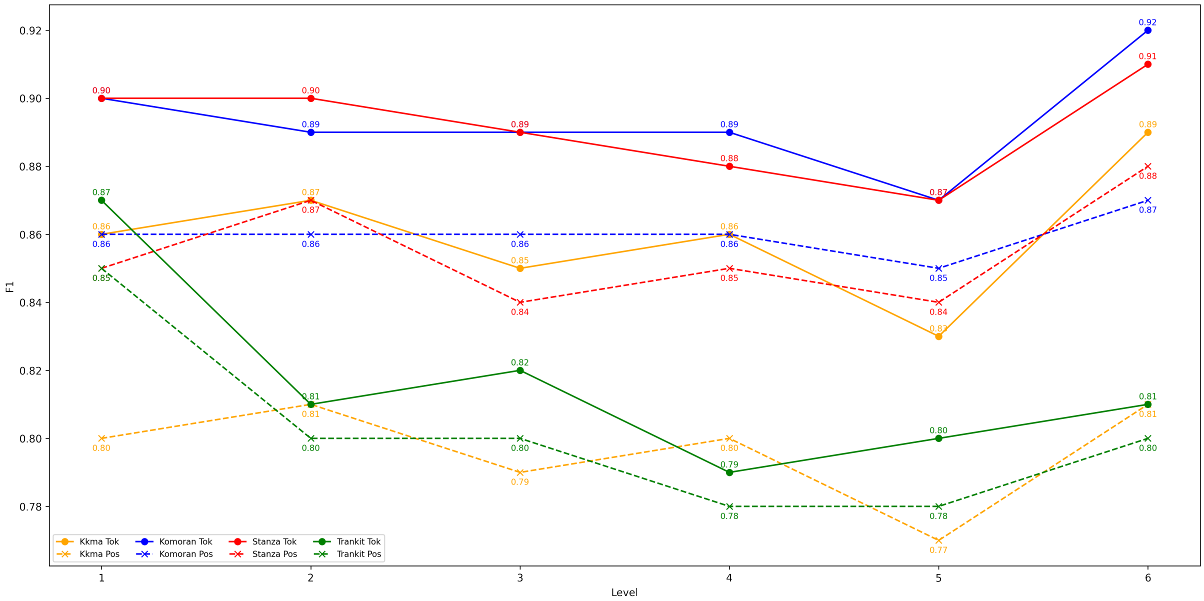


Figure 1: Comparison of analyzers (by-level) in L2 dataset

L2 data compared to their performance on the L1 data, indicating the challenges to automatic L2-data processing induced by learner language characteristics (and possibly in conjunction with the linguistic properties of Korean). Second, Stanza and Komoran achieved the highest F1 scores (tied) in morpheme tokenization and POS tagging on the L2 data. Given that Stanza and Trankit utilize state-of-art deep-learning algorithms, while Komoran is based on a comparatively basic probabilistic model, this finding indicates that even sophisticated models may suffer from coping with Korean learner corpora. Third, each analyzer demonstrated asymmetric patterns of performance by proficiency level. To illustrate, whereas the accuracy rates of Stanza and Komoran remained relatively stable across the levels, the accuracy rate of Trankit decreased notably after Level 2 (novice-intermediate). Of the four analyzers, Kkma showed the largest gap between the tokenization accuracy and the POS-tagging accuracy for all the levels.

5.2 By-tag performance

To examine the variation in performance across individual tags within the given datasets, we conducted a comparative analysis between the best-performing models (Stanza, Komoran) for each tag, as shown in the second and third columns of Table 3 (only includes results for the L2 data; see Appendix B for information on the L1 data). To calculate the by-tag accuracy, we included only the cases in which the number of predicted tags and the number of annotated tags were the same (within an eojeol unit for Stanza; within a sentence unit for Komoran). This approach ensures a fair comparison by maintaining an equal number of tags, avoiding any mismatch that could affect the evaluation process in an unexpected/uncontrollable way. Consequently, there was a discrepancy between the two tokenizers in terms of the number of tags ultimately included in the analysis.

To keep our analysis concise, we excluded tags related to punctuation, numbers, foreign languages, and errors, as well as tags with a low frequency count (overall counts below 10), resulting in a total of 29 tags for the main analysis. In the following section, we discuss tags with low accuracy or those that were of particular interest in previous studies. We also present the confusion matrix for these tags calculated by Stanza in Figure 2a, in which the off-diagonal elements indicate the number of incorrect predictions.

Predicate-related tags: The accuracy of VV (Verb, main), VX (Verb, auxiliary), and VA (Verb, adjective) was not satisfactory (except for VA in Komoran). This finding is surprising when we consider the status of verb and adjective as the primitive syntactic categories in human language and as one of the most significant content morphemes in Korean. Upon examining the confusion matrix (Figure 2a), we observed a considerable number of mismatches among these three groups, with a majority of the VX tags being predicted as the VV tags. The verb *iss-* emerged as one that requires further refinement in future research regarding its POS tags, because its classification as either a VV or VX, depending on its formal co-occurences with other morphemes, was not effective. Overall, these results suggest that the distinctions between main verbs, adjectives, and auxiliary verbs may not be clear-cut with the current taggers. These ambiguities could stem from linguistic complexities, overlapping grammatical features, or limitations in the underlying model's ability to discern the subtle differences between them.

Noun-related tags: XR (Noun, root) and NP (Pronoun) demonstrated notable by-analyzer asymmetries. Caution is needed, however, as their occurrences in the dataset were small. Considering language-specific properties of Korean (e.g. pronoun are underused), further investigation is required with a more sizeable dataset to fully reveal model performance on these tags.

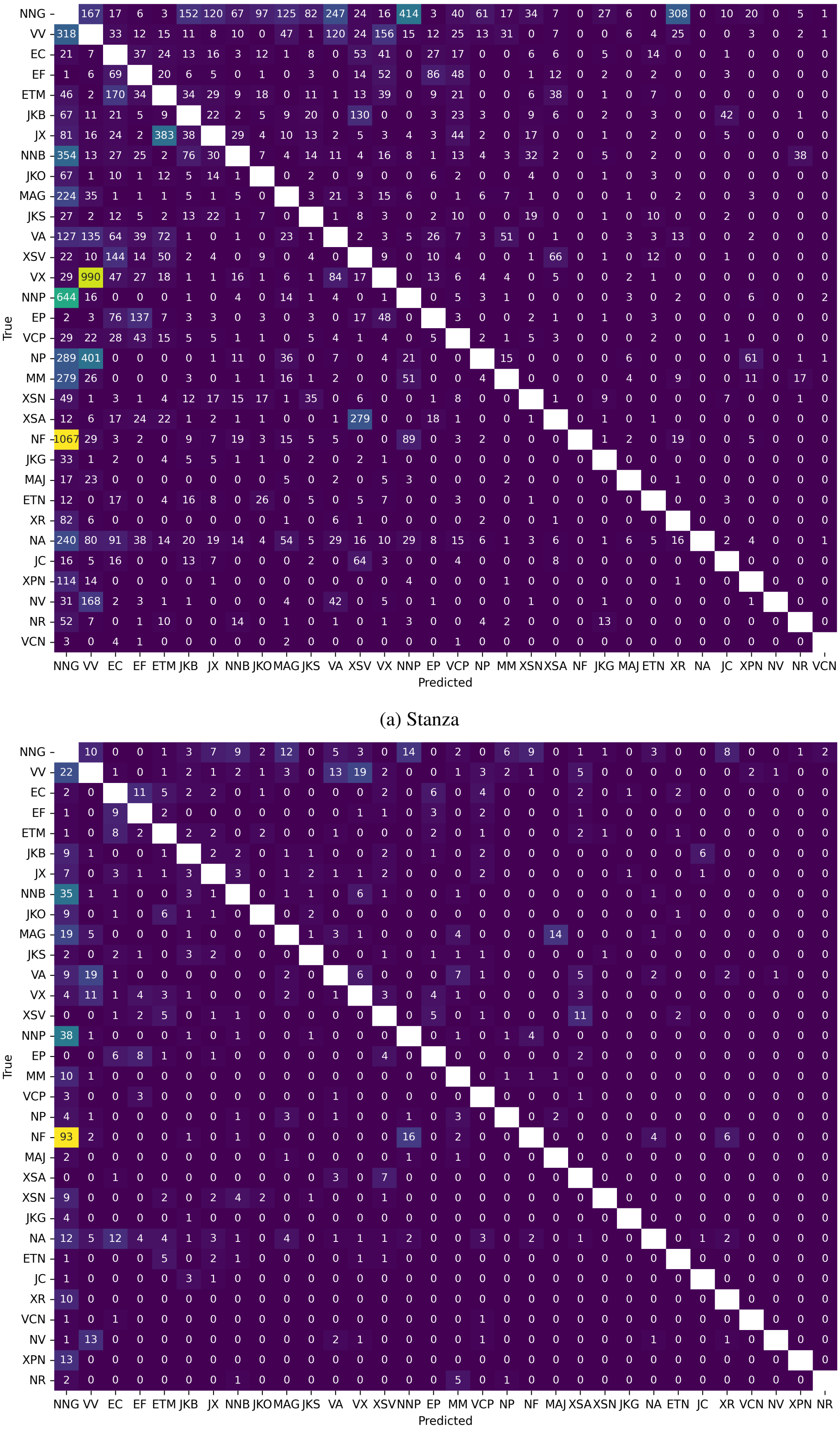
Particle- and suffix-related tags: Particles and suffixes are often considered challenging for the automatic processing of Korean (Shin and Jung, 2021). The results demonstrate that most particle-related tags (JKO, JKS, JKG, JKB, JX; but except for JC) and some suffix-related tags (predicate ending: EF, EC, EP) exhibited relatively high accuracy (mostly above 0.85) whereas tags comprising X (derivational suffixes: XSA, XSN, XSV) seemed not. The confusion matrix revealed that XSA was often tagged as XSV, and XSV as EC.

5.3 Model training through L2 data

Based on these observations, we trained a model on an L2 dataset and evaluated if model performance improved in comparison to a model trained solely on an L1 dataset. To construct the model, we split the KLM corpus into three datasets (80% for a training set; 10% for a development/validation set; 10% for a test set) and employed Stanza (pre-trained on the UD Korean GSD training set) to train morpheme tokenization (i.e., lemma) and tagging (i.e., XPOS) annotation models. For training the POS/morphological features tagger modules, we employed pre-trained embedding vectors from the L1-Korean-GSD model and integrated our L2 test dataset to the vector space. The accuracy evaluation was performed using the L1/L2 test sets with gold standard tokenization and POS tagging.

Analyzer	**Stanza**	**Komoran**	**Stanza+L2**
	(count)	(count)	(count)
JKO	0.94 (4705)	0.93 (2212)	**0.96** (454)
MAJ	**0.94** (1192)	**0.94** (668)	0.85 (143)
JKS	0.92 (4160)	0.91 (1874)	**0.95** (402)
JKG	0.92 (1257)	0.85 (423)	**0.95** (119)
EF	0.91 (7389)	**0.99** (3583)	0.93 (730)
VCN	0.91 (178)	**0.95** (75)	0.86 (26)
JKB	0.89 (6399)	0.89 (423)	**0.92** (634)
EC	0.88 (8871)	**0.90** (3920)	**0.90** (846)
MAG	0.87 (4628)	**0.90** (1885)	0.86 (446)
ETM	0.86 (6843)	0.90 (2753)	**0.91** (689)
JX	0.86 (5317)	**0.91** (2384)	**0.91** (543)
EP	0.86 (2984)	**0.98** (1299)	0.87 (289)
NNB	**0.85** (4685)	0.84 (1887)	0.84 (532)
XSN	0.84 (1557)	0.85 (581)	**0.87** (139)
ETN	0.83 (831)	**0.89** (326)	0.85 (83)
NNG	0.77 (30353)	0.82 (9682)	**0.83** (2866)
VCP	0.80 (2307)	**0.89** (744)	0.85 (216)
VV	0.74 (12704)	0.82 (4672)	**0.85** (1073)
MM	0.76 (1799)	**0.89** (733)	0.81 (223)
JC	0.77 (712)	0.63 (287)	**0.80** (61)
XSV	0.75 (3956)	**0.85** (1705)	**0.85** (364)
VA	0.73 (4028)	**0.92** (1547)	0.81 (392)
NP	0.68 (2260)	**0.91** (1010)	0.89 (201)
NNP	0.65 (3610)	0.47 (3476)	**0.77** (330)
XSA	0.68 (1353)	**0.71** (327)	**0.71** (142)
VX	0.62 (3624)	0.64 (1451)	**0.81** (369)
XR	0.41 (826)	**0.67** (318)	0.49 (52)
NR	0.27 (226)	**0.78** (73)	0.52 (18)
XPN	0.14 (283)	**0.40** (83)	0.35 (18)

Table 3: F1 scores (by-tag) in L2 dataset



(a) Stanza

(b) Stanza+L2 trained

Figure 2: Comparison of confusion matrix in L2 dataset

Re-evaluation results: Despite the small size of the training data, the Stanza+L2 model exhibited improvements in the F1 scores of **tokenization (0.93)** and **POS tagging (0.91)** compared to the best models trained exclusively on the L1 dataset (i.e., Stanza, Komoran), which had F1 scores of 0.89 for tokenization and 0.86 for POS tagging. However, when we compared the performance of the three models (i.e., Stanza, Komoran, Stanza+L2) on the L1 dataset, the performance of Stanza+L2 dropped (Token: 0.83; Tag: 0.82). The precise reason for this drop is unclear now; we speculate that it may be an example of "forgetting" (Kirkpatrick et al., 2017) in which neural networks abruptly forget what they have retained when learning a new task. In other words, it may be due to the detailed tagging scheme that our study adopts in comparison to the scheme of the L1 dataset (e.g., parsing causative/passive suffixes). Further research should clarify the interplay between the enhancement of parsing systems and the operation of neural networks in model training.

The by-tag performance of Stanza+L2 (as indicated in the final column of Table 3) shows that the accuracy of 15 out of 29 tags performed better than that for both of the L1 baseline models. The confusion matrix (Figure 2b) further showed that the locus of this improvement was predicate-related tags (VV, VA, VX) and error-related tags (NA, NF, NV). However, for the remaining 17 tags, Komoran still outperformed Stanza+L2. Considering the differences in the pre-training datasets of Stanza and Komoran, the disparity in training data size may have partially accounted for the observed performance discrepancies. Given this context, future research could explore the possibility of expanding Stanza's L2 training dataset, potentially incorporating a more diverse and comprehensive range of L2-Korean texts to improve its performance in areas in which the Stanza currently trails behind Komoran.

6 Conclusion

6.1 Summary of findings

In this study, we presented a manually annotated L2-Korean corpus and evaluated the performance of Korean morphological analyzers pre-trained on L1 datasets for tokenization and POS tagging on L2-Korean data. The KLM corpus and related resources are publicly accessible at: `https://github.com/NLPxL2Korean/Korean_Learner_Morpheme_corpus`.

The results revealed that morphological analyzers exhibited somewhat lower performance on L2-Korean data in comparison to their performance on L1 datasets. A detailed analysis of POS tags showed that several essential morphological tags, including predicate- and suffix-related tags, displayed relatively low accuracy. However, the study demonstrated that substantial improvements in morpheme tokenization and POS tagging performance for L2-Korean data could be attained by incorporating L2 data into the training sets, even with the relatively small dataset. Although no study has specifically focused on L2-Korean data so far, these findings align with previous studies on L2-English UD treebanks (e.g., Berzak et al., 2016; Kyle et al., 2022).

6.2 Future directions

To enhance computational resources for lesser-studied languages and improve their performance, carefully designed and validated data-processing pipelines hold great promise. This can be pursued through three primary directions. First, it is essential to expand the size of L2 corpora by (1) refining gold-standard annotation and tagging schemes, and (2) including informative metadata, such as learner proficiency. Second, incorporating syntactic treebanks into the KLM corpus or other available L2-Korean corpora could be considered, as previous research on L2 English has demonstrated promising outcomes. Third, both language-specific properties and learner language characteristics should be taken into account during the resource development process to ensure the interpretability of model results.

Limitations

Although our study offers empirical reports on the currently available Korean morphological parsers for processing L2-Korean texts, there are remaining areas which await further research. First, the KLM corpus that we proposed in this study consists of a relatively small dataset for training deep-learning models, so increasing the size of the dataset for training may be necessary to fully ensure model performance and generalize the result. Second, the proficiency levels in the original corpus seem unreliable because there was no separate test for proficiency measurement; instead, the developers used class levels as a proxy for learner proficiency.

This invites the need for re-evaluating individual learners' proficiency in Korean, ideally via holistic evaluation of learner essays by human raters. Finally, this work may need larger computing resources when applying cutting-edge deep-learning algorithms, especially with a larger training dataset.

Ethics Statement

We believe that future research should continue to consider linguistic diversity and give importance to the inclusion of underrepresented languages to research, while promoting equitable research practices in the field. Our findings thus have the potential to contribute to developing more effective and inclusive language-learning resources and tools for language learners. Specifically, connecting the currently available (and L1-based) morphological analyzers to language-specific properties and learner-language characteristics existing in L2 data, including the improvement of their performance, can enhance AI literacy, computer-assisted language learning, and educational materials to meet the unique and individualized needs of language learners with diverse backgrounds.

Acknowledgments

The authors gratefully acknowledge Youkyung Sung for her contributions to manual annotations and discussions for the enhancement of tokenizing and POS tagging guidelines on the L2 Korean dataset.

References

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

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.

Yige Chen, Eunkyul Leah Jo, Yundong Yao, Kyung-Tae Lim, Miikka Silfverberg, Francis M Tyers, and Jungyeul Park. 2022. Yet another format of universal dependencies for korean. *arXiv preprint arXiv:2209.09742*.

S. Cho and Y. Park. 2018. Characteristics of korean language writing by students at the university of sheffield (korean:sheffield tayhakkyo hankwuke haksupcauy cakmwun thukseng pwunsek). *Cakmwunyenkwu [Korean writing association]*, 38:149–172.

Jinho D Choi and Martha Palmer. 2011. Getting the most out of transition-based dependency parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 687–692.

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)*.

Nick C Ellis and Fernando Ferreira-Junior. 2009. Construction learning as a function of frequency, frequency distribution, and function. *The Modern language journal*, 93(3):370–385.

Dana Gablasova, Vaclav Brezina, and Tony McEnery. 2017. Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence. *Language learning*, 67(S1):155–179.

Janggyoon Jeong, Lee Min-Young, Kwon Minji, and Jung Woo-Sung. 2018. Classification of writing style by using a morpheme network analysis.

Boo Kyung Jung. 2022. The nature of l2 input: Analysis of textbooks for learners of korean as a second language. *Korean Linguistics*, 18(2):182–208.

JY. Kim, YH. Park, MJ. Kim, HN. Kim, SK. Choi, JH. Suh, and YJ Kwak. 2016. A study of developing usage searcher of grammar pattern in the korean learner's writing corpus (korean: hankwuke haksupcauy cakmwun malmwungchilul hwalyonghan mwunhyeng yonglyey kemsaykki kaypal yenkwu). *Teaching Korean as a Foreign Language*, 44:131–156.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

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

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

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.

Sun-Hee Lee, M Dickenson, and Ross Israel. 2016. Challenges of learner corpus annotation: Focusing on korean learner language analysis (kolla) system. *Language facts and perspectives*, 38:221–251.

KyungTae Lim, Jayoung Song, and Jungyeul Park. 2022. Neural automated writing evaluation for korean l2 writing. *Natural Language Engineering*, page 1–23.

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

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.

YJ. Nam and UP. Hong. 2014. Towards a corpus-based approach to korean as a second language (korean: L2loseuy hankwuke cayenpalhwa khophesuuy kwuchwukkwa hwalyong). *The Journal of the Humanities for Unification*, 57:193–220.

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. *arXiv preprint arXiv:2004.10643*.

Eunjeong L Park and Sungzoon Cho. 2014. Konlpy: Korean natural language processing in python. In *Annual Conference on Human and Language Technology*, pages 133–136. Human and Language Technology.

Jungyeul Park, Daisuke Kawahara, Sadao Kurohashi, and Key-Sun Choi. 2013. Towards fully lexicalized dependency parsing for Korean. In *Proceedings of the 13th International Conference on Parsing Technologies (IWPT 2013)*, pages 120–126, Nara, Japan. Assocation for Computational Linguistics.

Jungyeul Park and Jung Hee Lee. 2016. A korean learner corpus and its features. *En-e-hak [Linguistics]*, (75):69–85.

Jungyeul Park and Francis Tyers. 2019. A new annotation scheme for the sejong part-of-speech tagged corpus. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 195–202.

Karthik Raman, Iftekhar Naim, Jiecao Chen, Kazuma Hashimoto, Kiran Yalasangi, and Krishna Srinivasan. 2022. Transforming sequence tagging into a seq2seq task. *arXiv preprint arXiv:2203.08378*.

Gyu-Ho Shin and Boo Kyung Jung. 2021. Automatic analysis of passive constructions in korean: Written production by mandarin-speaking learners of korean. *International Journal of Learner Corpus Research*, 7(1):53–82.

Gyu-Ho Shin and Boo Kyung Jung. 2022. Input–output relation in second language acquisition: Textbook and learner writing for adult english-speaking beginners of korean. *Australian Review of Applied Linguistics*, 45(3):347–370.

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

A Sejong Tag Set

The table provides a Sejong Tag set. The description was sourced from Jeong et al., 2018.

Tag	Description
NNG	Noun, common (보통 명사)
NNP	Proper Noun (고유 명사)
NNB	Noun, common bound (의존 명사)
NR	Numeral (수사)
NP	Pronoun (대명사)
VV	Verb, main (동사)
VA	Adjective (형용사)
VX	Verb, auxiliary (보조 동사)
VCP	Copular, positive (긍정 지정사)
VCN	Copular, negative (부정 지정사)
MM	Determiner (관형사)
MAG	Common adverb (일반 부사)
MAJ	Conjunctive adverb (접속 부사)
IC	Exclamation (감탄사)
JKS	Postposition, nominative (주격 조사)
JKC	Postposition, complement (보격 조사)
JKG	Postposition, prenominal (관형격 조사)
JKO	Postposition, objectival (목적격 조사)
JKB	Postposition, adverbial (부사격 조사)
JKV	Postposition, vocative (호격 조사)
JKQ	Postposition, quotative (인용격 조사)
JC	Postposition, conjunctive (접속 조사)
JX	Postposition, auxiliary (보조사)
EP	Ending, prefinal (선어말 어미)
EF	Ending, closing (종결 어미)
EC	Ending, connecting (연결 어미)
ETN	Ending, nounal (명사형 전성 어미)
ETM	Ending, determinitive (관형형 전성 어미)
XPN	Prefix, nounal (체언 접두사)
XSN	Suffix, verbal (명사 파생 접미사)
XSV	Suffix, verb derivative (동사파생 -)
XSA	Suffix, adjective derivative (형용사 파생 -)
XR	Root (어근)
NF	Undecided (consider a noun) (명사 추정)
NV	Undecided (consider a verb) (용언 추정)
NA	Undecided (분석 불능)
SF	Period, Question, Exclamation (마침표 등)
SE	Ellipsis (줄임표)
SS	Quotation, Bracket, Dash (따옴표 등)
SP	Comma, Colon, Slash (쉼표,콜론, 빗금)
SO	Hyphen, Swung Dash (붙임표, 물결표)
SW	Symbol (기타기호)
SH	Chinese characters (한자)
SL	Foreign characters (외국어)
SN	Number (숫자)

B F1 scores (by-tag) in L1 dataset

The table provides the by-tag accuracies from a L1 reference corpus (UD Korean GSD).

Analyzer	**Stanza**	**Komoran**
	(count)	(count)
JKO	0.96 (653)	0.93 (246)
MAJ	0.77 (44)	0.68 (36)
JKS	0.94 (564)	0.95 (242)
JKG	0.93 (323)	0.94 (121)
EF	0.96 (758)	0.99 (328)
VCN	1.00 (10)	1.00 (3)
JKB	0.93 (1005)	0.91 (372)
EC	0.95 (1590)	0.94 (721)
MAG	0.90 (622)	0.95 (248)
ETM	0.97 (967)	0.92 (394)
JX	0.92 (871)	0.93 (382)
EP	0.94 (573)	0.95 (220)
NNB	0.91 (715)	0.82 (223)
XSN	0.88 (314)	0.89 (131)
ETN	0.82 (108)	0.86 (38)
NNG	0.91 (6136)	0.80 (1684)
VCP	0.86 (334)	0.90 (113)
VV	0.93 (1478)	0.88 (615)
MM	0.92 (189)	0.89 (78)
JC	0.85 (161)	0.81 (61)
XSV	0.93 (689)	0.90 (259)
VA	0.93 (458)	0.96 (228)
NP	0.88 (138)	0.87 (71)
NNP	0.75 (855)	0.37 (793)
XSA	0.87 (225)	0.88 (90)
VX	0.91 (390)	0.76 (168)
XR	0.83 (206)	0.94 (87)
NR	0.74 (107)	0.81 (28)
XPN	0.42 (66)	0.76 (14)