# Towards L2-friendly pipelines for learner corpora: A case of L2-Korean learners

Hakyung Sung[†] and Gyu-Ho Shin[‡]

[†]Department of Linguistics, University of Oregon
[‡]Department of Linguistics, University of Illinois Chicago

## SUMMARY


- We introduce the Korean-Learner-Morpheme (KLM) corpus, a manually annotated dataset from second language (L2) learners of Korean, featuring morpheme tokenization and part-of-speech (POS) tagging ($n = 129,784$).
- 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.
- Model fine-tuning with the KLM corpus demonstrates improved tokenization and POS tagging accuracy on the L2-Korean dataset.


## DATASETS & ANNOTATION

**L2 corpus:** The KLM corpus, comprising 129,784 morphemes with morpheme tags grounded in the Sejong tag set

- The inclusion of data concerning classroom proficiency levels (ranging from 1 to 6 as a proxy for learner proficiency), nationality, gender, and writing topics
- The random extraction of 600 texts from the original corpus (Park & Lee, 2016), with each proficiency level represented by 100 texts
- The manual annotation of the corpus by three Korean native speakers, featuring detailed descriptions during Korean morpheme annotations and their evaluations

| 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** |

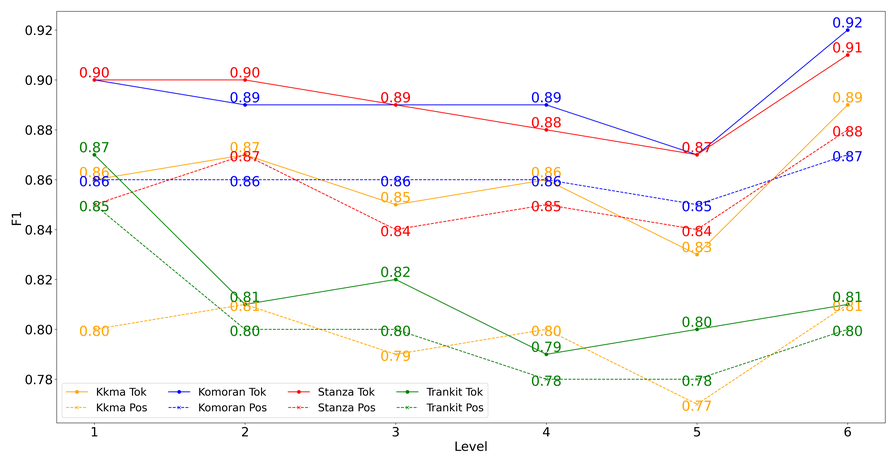
**Reference L1 corpus:** Google Korean Universal Dependency Treebank (UD Korean GSD) as a baseline

## OVERALL PERFORMANCE

**Morphological analyzers:** The employment of four open-access morphological analyzers, incorporating various computational algorithms ranging from statistical models to deep-learning models (i.e., Stanza: Deep Biaffine attention for neural network; Trankit: Transformer; Kkma & Komoran: Hidden Markov model)

| Analyzer | L2 TOK | L2 POS | L1 TOK | L1 POS |
|---|---|---|---|---|
| Stanza | **0.89** | **0.86** | 0.92 | 0.93 |
| Trankit | 0.81 | 0.80 | 0.85 | 0.88 |
| Kkma | 0.86 | 0.80 | 0.88 | 0.81 |
| Komoran | **0.89** | **0.86** | 0.92 | 0.86 |

## BY-LEVEL PERFORMANCE




## MODEL TRAINING & RE-EVALUATION

- **Training:** The Stanza model, further trained on an L2 dataset (KLM corpus), aiming to evaluate the potential improvement in performance compared to an exclusively L1-trained model
- **Results**: The F1 scores for the Stanza+L2 model showing improvements with **tokenization** $0.93 > 0.89$ and **POS tagging** $0.91 > 0.86$ (compared to the highest scores of models trained exclusively on the L1 dataset)

## BY-TAG PERFORMANCE

| POS Tags | Analyzer: Stanza | Analyzer: Komoran | Analyzer: Stanza+L2 |
|---|---|---|---|
| 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) |
| VCN | 0.91 (178) | **0.95** (75) | 0.86 (26) |
| JKB | 0.89 (6399) | 0.89 (423) | **0.92** (634) |
| MAG | 0.87 (4628) | **0.90** (1885) | 0.86 (446) |
| JX | 0.86 (5317) | **0.91** (2384) | **0.91** (543) |
| NNB | **0.85** (4685) | 0.84 (1887) | 0.84 (532) |
| XSN (Suffix, n.) | 0.84 (1557) | 0.85 (581) | **0.87** (139) |
| ETN | 0.83 (831) | **0.89** (326) | 0.85 (83) |
| NNG (Noun, common) | 0.77 (30353) | 0.82 (9682) | **0.83** (2866) |
| VCP | 0.80 (2307) | **0.89** (744) | 0.85 (216) |
| VV (Verb, main) | 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 (Suffix, v.) | 0.75 (3956) | **0.85** (1705) | **0.85** (364) |
| VA (Adjective) | 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 (Suffix, adj.) | 0.68 (1353) | **0.71** (327) | **0.71** (142) |
| VX (Verb, auxiliary) | 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) |

## DISCUSSIONS & CONCLUSION

- **Overall:** Observation of somewhat reduced performance by the morphological analyzers when applied to L2-Korean data compared to the L1 reference corpus
- **By-level:** Identification of asymmetric performance patterns by proficiency level for each analyzer
- **By-tag:** Through detailed POS tag analysis, disclosure of low accuracy in essential morphological tags, including predicate (highlighted in yellow) and suffix-related (highlighted in red) tags
- **Developing L2 domain-specific model:** Achievement of improved performance in morpheme tokenization and POS tagging for L2-Korean data by integrating L2 data into the training sets
- **Future direction:** Emphasis on the potential of carefully designed and validated data-processing pipelines for enhancing computational resources for lesser-studied languages and boosting their performance