L2K-ARG: An open-access learner corpus of Korean argumentative writing

Hakyung Sung[1] & Gyu-Ho Shin[2]
[1]Second Language Studies, University of Hawaii, USA (hksung001@gmail.com)
[2]Linguistics, University of Illinois Chicago, USA (ghshin@uic.edu)

Keywords: Korean as a second language, argumentative writing, representativeness

Second language (L2) corpora are foundational resources for both L2-acquisition research and natural language processing, enabling advances in learner language modeling, automated assessment, and error detection. However, their availability remains heavily skewed toward high-resource languages, leaving languages such as Korean underrepresented in global learner corpora infrastructures (e.g., CLARIN, CECL). While several initiatives have produced L2-Korean datasets, their utility for the international research community is often constrained by restrictive access policies, limited documentation of data collection protocols, or incomplete learner metadata.

To address these gaps and support researchers worldwide working on L2-Korean corpora, this study introduces the L2-Korean ARGumentative Writing Corpus (L2K-ARG), an open-access learner corpus designed with explicit attention to corpus representativeness and transparency. For principled corpus construction, its design was guided by the framework proposed by Egbert et al. (2022), which foregrounds both domain and distributional considerations. The target domain was operationalized as university-level L2-Korean argumentative writing produced under controlled testing conditions, with systematic variation across task prompts and learner proficiency. Distributional considerations further informed sampling decisions to ensure that the corpus yields sufficiently precise estimates of frequently occurring morphosyntactic features.

At present, the corpus comprises 484 argumentative essays produced by 174 university-level learners from diverse first-language backgrounds (Chinese, Czech, English, Japanese, and Heritage Korean). To ensure comparability, all texts were elicited individually in proctored settings under a strict 20-minute time limit. Learners responded to a subset of five standardized prompts adapted from the Test of Proficiency in Korean, covering topics such as environmental conservation, competition versus cooperation, and early childhood education. A distinguishing feature of L2K-ARG is its morphosyntactic annotation based on the Universal Dependencies (UD) framework (de Marneffe et al., 2021), tailored for Korean learner data (Sung & Shin, 2024). The annotation pipeline decomposes Korean *eojeol* (i.e., a Korean-specific spacing unit that corresponds to a word-like unit) into constituent morphemes to reflect the language's agglutinative structure. Five core annotation layers are provided: (1) LEMMA (morphemic decomposition), (2) UPOS (universal part-of-speech tags), (3) XPOS (Korean-specific part-of-speech tags), (4) HEAD (dependency heads), and (5) DEPREL (dependency relations).

In addition, L2K-ARG is enriched with independent measures of general Korean proficiency and task-specific writing ability. First, participants' overall proficiency in Korean was assessed using the Korean C-test, in which syllable-level deletions are introduced within *eojeol* units and learners must restore the missing elements using contextual information. Second, because general proficiency measures do not directly assess writing production, a supplementary analytic writing assessment was conducted for a subset of the essays (n = 363). Three trained native Korean instructors evaluated each essay using a 10-point Likert-scale analytic rubric covering content and task fulfillment, essay organization and development, and language use (grammatical/lexical appropriateness and accuracy).

To conclude, L2K-ARG is released as an open, reusable corpus that integrates UD-based annotations with rich proficiency metadata. Future work will expand learner populations, task types, and corpus size to support more comprehensive analyses of various constructions and a wider range of L2-Korean usage contexts.

References

de Marneffe, M. C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal dependencies. *Computational linguistics*, *47*(2), 255-308.

Egbert, J., Biber, D., & Gray, B. (2022). *Designing and evaluating language corpora: A practical framework for corpus representativeness*. Cambridge University Press.

Sung, H., & Shin, G. H. (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)* (pp. 3747-3758).