

Phrasal dependency features as predictors of L2 Korean written proficiency: A topic-specific regression analysis

Bookyung Jung (Yale University; bookyung.jung@yale.edu)

Previous L2 studies have used NLP techniques to examine linguistic features associated with L2 proficiency (Lu, 2011; Kyle & Crossley, 2016). Although they have advanced NLP-informed L2 studies, it remains unclear to what extent features developed primarily for commonly studied languages generalize to others. In addition, research applying NLP techniques to L2 Korean remains underdeveloped. Accordingly, the present study investigates whether phrasal dependency parsing features predict proficiency across different essay-topics, with particular attention to the stability of these features' effects across topics.

Forty-three L2 Korean learners (mean age: 24.99, SD: 5.18) wrote three argumentative essays, with 20 minutes each. The prompts were adapted from TOPIK: T1) Early language education, T2) Necessity of history education, and T3) Competition versus cooperation? Participants' proficiency was measured using a Korean C-test (mean: 129.68, SD: 15.10, min:101, max:168, out of 188; Lee-Ellis, 2009). The essays were parsed using the spaCy (Honnibal & Montani, 2017). From all dependency features, phrasal dependency features were selected (Appendix1). The raw frequency of each index was normalized by dividing by the total number of *eojeol*, followed by removing features that violate assumptions (correlation $|r| < .10$, multicollinearity $|r| > .70$; Kyle & Crossley, 2016). Using the remaining features (Appendix2), we conducted three regression analyses to identify the model that best explained proficiency, building from a base that included three common features across topics (*advmod*, *amod*, *det*). Given the moderate sample size, the modeling prioritized stability and interpretability, using adjusted R^2 and stability of coefficient estimates as criteria.

The results indicate that the topic-specific models included different features: *nsubj* (T1; $R^2 = .006$), *obj* (T2; $R^2 = .105$), and *det:poss* (T3; $R^2 = .213$), in addition to three common features. Moreover, *advmod* showed a stable effect across all three topics, with the same direction of association. In contrast, some coefficients differed in sign across topics: negative *amod* in T3, and positive *det* in T1.

The findings show that some dependency features exhibit stable behavior, and their contribution to proficiency may be modest. At the same time, dependency features may have limitations in capturing the different linguistic features that learners employ depending on discourse demands (T1). These results underscore the importance of adopting topic-sensitive pedagogical support in L2 Korean instruction (e.g., emphasizing different phrase) and applying linguistic features cautiously when using NLP-based indices. The study aims to contribute to NLP-based L2 Korean research by introducing a stability-oriented modeling framework that prioritizes coefficient robustness and cross-topic generalizability in small-sample datasets. Moderate sample size and potential parser-extraction error for learner language are limitations.

References

- Honnibal, M., & Montani, I. (2017). spaCy 2: *Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*.
- Kyle, K., & Crossley, S. (2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing*, 23, 12-24.
- Lee-Ellis, S. (2009). The development and validation of a Korean C-Test using Rasch Analysis. *Language Testing*, 26(2), 245–274.
- Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly*, 45, 36-62.

Appendix1. Phrasal dependency features

[*advmod*] adverbial modifier, [*amod*] adjectival modifier, [*case*] case-marking,
[*det*] determiner, [*det:poss*] possessive determiner, [*flat*] flat multiword expression,
[*iobj*] indirect object, [*mark*] marker, [*nmod*] nominal modifier,
[*nsubj*] nominal subject, [*nsubj:pass*] passive voice nominal subject,
[*nummod*] numeric modifier, [*obj*] object, [*obl*] oblique

Appendix2. Remaining phrasal dependency features after statistical assumption check

T1: *advmod, amod, case, det, flat, mark, nsubj*

T2: *advmod, amod, case, det, det:poss, mark, nmod, obj*

T3: *advmod, amod, det, det:poss, flat, obj, obl*