Validating Automated Measures of Phraseological Competence in L2 Speaking: Evidence from Human Judgments

Tingting Wang[1], Magali Paquot[2], Dandan Zhou[1]

[1]Nanjing University; [2]FNRS-UCLouvain

Phraseological competence is crucial for language acquisition, processing, and fluency (Ellis *et al.*, 2008; Paquot *et al.*, 2020) but remains challenging for L2 learners (Laufer & Waldman, 2011; Paquot & Granger, 2012). While existing studies have developed automated indices to operationalize this construct and demonstrated their predictive validity by linking them to L2 proficiency, few have directly addressed whether the indices adequately capture the construct itself (Paquot & Naets, 2025). This raises concerns about their construct validity – the extent to which a measure accurately represents the theoretical construct it is intended to assess. Establishing construct validity requires evidence of alignment between automated measures and alternative methods, with human ratings serving as a critical benchmark in applied linguistics (Crossley *et al.*, 2013). However, few studies have evaluated the alignment between human judgments and automated measures of phraseological competence, particularly in L2 oral production, highlighting the need for further validation research.

Grounded in an argument-based framework (Kane, 2006), this study explores the following research questions to provide convergent evidence for automated measures of phraseological competence in L2 speaking:

RQ1. To what extent do automated measures of phraseological competence align with human ratings?

RQ2. What features of phraseological competence do human raters focus on, and how well do automated indices capture them?

A mixed-methods approach was adopted, consisting of two phases: (1) a corpus linguistic analysis using automated measures and (2) an experiment with human raters. In the first phase, oral performances from 98 test-takers of the TEM 8-Oral (Test for English Majors-Band 8) were analyzed using automated indices. Phraseological competence was operationalized along three dimensions–accuracy, diversity, and sophistication—and measured with respect to two key phenomena in phraseology: co-occurrence and recurrence (Granger & Paquot, 2008). Co-occurrence was examined through six grammatical relations (adjectival modifier (amod), direct object (dobj), adverbial modifier (advmod), adjectival complement (acomp), nominal subject (nsubj), and prepositional modifier (prep)). Recurrence was explored through the analysis of three-word lexical bundles. In the second phase, 30 human raters were recruited to evaluate the same performances using a comparative judgment method. They provided both ratings of phraseological competence and qualitative comments explaining their decisions. These human judgments served as external benchmarks for automated measures and offered insights into the features that influenced rater evaluations.

Based on a pilot study in which two raters evaluated 30 texts and provided comments on their decisions, we anticipate the following findings: 1) Automated measures of phraseological competence will demonstrate varying degrees of alignment with human ratings, with the sophistication measures for lexical bundles expected to show the strongest correlation and explanatory power; 2) Automated indices will effectively capture certain features prioritized by human raters, particularly phraseological diversity and sophistication. However, they may be less effective in adequately capturing some nuanced qualitative aspects that are equally emphasized in human evaluations of phraseological competence, such as idiomaticity and contextual appropriateness. The findings of this study will provide convergent validity evidence for the automated measures of phraseological competence and highlight areas where computational measures require refinement, particularly in capturing qualitative features emphasized by human raters. By bridging automated analysis and human judgment, the study aims to inform the development of more robust assessment tools and offer pedagogical insights for fostering phraseological competence in L2 learners.

References

Crossley, S., Salsbury, T., & McNamara, D. S. (2013). Validating lexical measures using human scores of lexical proficiency. In S. Jarvis & M. Daller (Eds.), *Studies in bilingualism* (Vol. 47, pp. 105-134). John Benjamins Publishing Company. https://doi.org/10.1075/sibil.47.06ch4

Ellis, N. C., Simpson-Vlach, R., & Maynard, C. (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly, 42*(3), 375–396. https://doi.org/10.1002/j.1545-7249.2008.tb00137.x

Granger, S., & Paquot, M. (2008). Disentangling the phraseological web. In S. Granger & F. Meunier (Eds.), *Phraseology: An Interdisciplinary Perspective* (pp. 27–49). Amsterdam & Philadelphia: John Benjamins. https://doi.org/10.1075/z.139.07gra

Kane, M. T. (2006). Validation. In R. Brennen (Ed.), *Educational measurement* (4th ed., pp. 17–64). Praeger and Greenwood Publishing.

Laufer, B., & Waldman, T. (2011). Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning, 61*(2), 647–672. https://doi.org/10.1111/j.1467-9922.2010.00621.x

Paquot, M., & Granger, S. (2012). Formulaic language in learner corpora. *Annual Review of Applied Linguistics, 32,* 130–149. https://doi.org/10.1017/S0267190512000098

Paquot, M., & Naets, H. (2025). Phraseological sophistication as a multidimensional construct: Exploring the relationship between association, register specificity and frequency of word combinations. In T. Larsson & D. Biber (Eds.), *Cumulative knowledge building and replication in Learner Corpus Research. International Journal of Learner Corpus Research, 11*(1). https://doi.org/10.1075/ijlcr.23033.paq

Paquot, M., Gries, S. Th., & Yoder, M. (2020). Measuring lexicogrammar. In P. Winke & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 223–232). Routledge. https://doi.org/10.4324/9781351034784-25