Predicting Collocational Preferences: A Corpus-Based Approach to Grammatical Sensitivity

Bálint József Ugrin[1], Ágnes Lukács[1]
[1] Budapest University of Technology and Economics
balintugrin@edu.bme.hu

The use of large language corpora has been widely explored for predicting both lexical [1] and abstract grammatical categories [2,3]. In line with this research, we used corpus tools to predict speaker preferences for frequent two-word collocations (e.g., *"heated argument"* vs. *"intense argument"*).

We investigated whether grammatical sensitivity to common or uncommon expressions can be effectively assessed through two-word collocations using a forced-choice task, as previous studies have revealed systematic knowledge of collocations [4] and differences in processing speed for collocations embedded in context [5]. Our investigation was guided by two questions: (1) which corpus-driven co-occurrence measurement is the best predictor of speaker preferences, and (2) do speakers behave consistently enough across test items to support a unidimensional and reliable measurement of collocational sensitivity?

We tested Hungarian adjective–noun collocations varying in adjective choice but sharing the same noun head, allowing comparison within a shared semantic field. Each trial consisted of four target items (semantically similar collocations) differing in their association strength, and one distractor item to control for attentional lapses. Participants were instructed to select the most natural-sounding collocation from the five options. The participant sample included 100 native Hungarian speakers (median age = 22, range: 18–77).

Our first hypothesis concerned evaluating corpus-based metrics that capture word association strength. From several candidate measures, we tested mutual information (MI), which is commonly used in psycholinguistic research [3,5], and logDice, which is recommended in computational linguistic accounts [6]. While MI is better at capturing words that co-occur more than expected, logDice balance frequency and co-occurrence making it less prone to rare collocations. As the two association metrics strongly correlate, evaluating their predictive power using the same trials would pose difficulty. To compare the two, 25 trials featured items that varied in MI scores while maintaining similar logDice scores, and another 25 trials featured items that varied in logDice scores with similar MI scores. Each participant saw all 50 trials in a randomised order. Across all trials, association strength ranges and variability were held constant, however, mean association strength varied, resulting in differing trial difficulty. Collocational frequency was also collected and showed less explanatory power than association strengths. Our results showed that logDice scores better predicted speaker preferences than MI, and this effect held regardless of trial difficulty.

Our second hypothesis posited that speaker behaviour would be consistent across trials—i.e., participants who performed well on one trial would tend to do so on others. We analysed the data using exploratory factor analysis and item response theory which confirmed that participants behaved consistently in trials structured around logDice-based variability. This provides evidence that our task can serve as a unidimensional and reliable (internally consistent) instrument for assessing grammatical sensitivity in collocations. Furthermore, our findings illustrate how advances in computational linguistics, such as the preference of logDice over MI, can be successfully applied in investigating intermediate phenomena between vocabulary and grammar.

References

[1] Keuleers, E., Stevens, M., Mandera, P., & Brysbaert, M. (2015). Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment. *Quarterly Journal of Experimental Psychology*, *68*(8), 1665–1692. https://doi.org/10.1080/17470218.2015.1022560

[2] Baayen, R. H., Chuang, Y.-Y., Shafaei-Bajestan, E., & Blevins, J. P. (2019). The Discriminative Lexicon: A Unified Computational Model for the Lexicon and Lexical Processing in Comprehension and Production Grounded Not in (De)Composition but in Linear Discriminative Learning. *Complexity*, *2019*(1), 4895891. https://doi.org/10.1155/2019/4895891

[3] Futrell, R., Qian, P., Gibson, E., Fedorenko, E., & Blank, I. (2019). Syntactic dependencies correspond to word pairs with high mutual information. *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, 3–13. https://doi.org/10.18653/v1/W19-7703

[4] Dąbrowska, E. (2014). Words that go together: Measuring individual differences in native speakers' knowledge of collocations. *The Mental Lexicon*, *9*(3), 401–418. https://doi.org/10.1075/ml.9.3.02dab

[5] McConnell, K., & Blumenthal-Dramé, A. (2021). Usage-Based Individual Differences in the Probabilistic Processing of Multi-Word Sequences. *Frontiers in Communication*, *6*, 703351. https://doi.org/10.3389/fcomm.2021.703351

[6] Rychlý, P. (2008). A lexicographer-friendly association score. Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN.