Intercomprehension of Slavic Functional Multiwords: Translation Experiment Results

Maria Kunilovskaya, Iulia Zaitova, Wei Xue, Ira Strenger (University of Saarland)
maria.kunilovskaya@uni-saarland.de

This study reports the results of a free translation experiment as a probe for Slavic intercomprehension between Russian and Czech, Polish, Bulgarian, Belarusian, Ukrainian. The experiment focuses on non-compositional functional multiword expressions – or microsyntactic units (MSUs) (Avgustinova & Iomdin, 2019) from five word classes: prepositions, adverbial predicatives, conjunctions, particles, parentheses. MSUs require additional cognitive effort in cross-lingual comprehension because their meaning cannot be inferred from the components. The lack of transparency and discourse-organising functions make MSU a good case for intercomprehension studies. They reflect the ability of the participants to understand the message in a foreign language.

The data comes from an experiment, where native speakers of Russian translated contextualised MSUs from five Slavic languages into Russian. The study engaged 126 users without formal knowledge of the source languages (SLs). The translation tasks were based on at least 50 sentences containing a unique MSU stimulus, and each stimulus has generated at least 20 responses. The Slavic MSUs (and their contexts) were extracted as correspondences for Russian MSUs using bidirectional parallel corpora and lexicographic resources of the Russian and Czech National Corpora[1]. The participants' responses were annotated for seven types of translation solutions (paraphrase, correct, fluent literal, awkward literal, fantasy, noise, and empty), designed to capture the level of the cross-linguistic intelligibility of the stimuli. The annotation was used to calculate items' intelligibility scores.

The study aims to reveal factors that favour intercomprehension across Slavic languages based on a range of computational representations and modelling approaches. In particular, regression and correlation analysis are used to identify the most important intercomprehension predictors. The stimuli are represented by features that reflect the properties of the translation tasks and their outcomes, including a pronunciation-based variant of the point-wise Phonologically Weighted Levenshtein Distance (PWLD) motivated in our previous work (Zaitova et al., 2024), cosine similarities, surprisals, translation quality scores and translation solution entropy indices. Cosine similarities and surprisals are calculated based on ruRoBERTa-large model (Zmitrovich et al., 2024), a dedicated Russian language Transformer, which can process input in Latin script. Automatic translation quality scores were calculated using COMET models (Rei et al., 2022). These approaches utilise contexts for the targeted MSUs available in our dataset.

The experimental results from both annotation and computational models confirm the expected gradual increase of mutual intelligibility from West-Slavic to East-Slavic languages. We show that intelligibility is highly contingent on the ability of speakers to recognise and interpret formal similarities between languages as well as on the size of these similarities. For several Slavic languages, the context sentence complexity was a significant predictor of intelligibility.

[1] https://ruscorpora.ru/en/ and https://www.korpus.cz/

References

Avgustinova, T., & Iomdin, L. (2019). Towards a typology of microsyntactic constructions. In *Computational and Corpus-Based Phraseology: Third International Conference, Europhras 2019, Malaga, Spain, September 25–27, 2019, Proceedings 3* (pp. 15-30). Springer International Publishing.

Rei, R., C. de Souza, J. G., Alves, D., Zerva, C., Farinha, A. C., Glushkova, T., Lavie, A., Coheur, L., & Martins, A. F. T. (2022, December). COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In P.Koehn, L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussà, C. Federmann, M. Fishel, A. Fraser, M. Freitag, Y. Graham, R. Grundkiewicz, P. Guzman, B. Haddow, M. Huck, A. Jimeno Yepes, T. Kocmi, A. Martins, M. Morishita, . . . M. Zampieri (Eds.). In *Proceedings of the seventh conference on machine translation (WMT)* (pp. 578–585). Association for Computational Linguistics.

RNC. (2003–2023). Russian National Corpus [Retrieved September 28, 2023]. *Russian National Corpus Project*.

Zaitova, I., Stenger, I., Butt, M. U., & Avgustinova, T. (2024, May). Cross-Linguistic Processing of Non-Compositional Expressions in Slavic Languages. In *Proceedings of the Workshop on Cognitive Aspects of the Lexicon@ LREC-COLING 2024* (pp. 86-97).

Zmitrovich, D., Abramov, A., Kalmykov, A., Kadulin, V., Tikhonova, M., Taktasheva, E., ... & Fenogenova, A. (2024, May). A Family of Pretrained Transformer Language Models for Russian. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 507-524).