Predictive Potential of Linguistic Distances and Surprisal in Multilingual Intercomprehension Experiments

Iuliia Zaitova, Wei Xue, Irina Stenger, Tania Avgustinova

(Department of Language Science and Technology, Saarland University, Germany)

izaitova@lsv.uni-saarland.de

This study explores the predictive potential of linguistic distances and surprisal in multilingual intercomprehension experiments. Linguistic distances refer to the measurable differences between languages (Wichmann et al., 2010). They can be quantified in various domains, such as phonology and orthography (Gooskens and van Bezooijen, 2013), with each domain contributing differently to the overall distance between languages. Previous research showed that higher linguistic distances were associated with decreased intercomprehension (Gooskens and Swarte, 2017; Möller and Zeevaert, 2015; Vanhove and Berthele, 2015).

The difficulty in processing a linguistic unit is proportional to the metric of surprisal, as estimated by language models (Hale, 2001; Levy, 2008). Surprisal is defined as the negative log-likelihood of encountering a unit given its preceding context derived from language models (surprisal = $-\log P(w_i \mid \textit{context})$ for a given unit w_i in a sequence), and it effectively measures the unpredictability of that unit (Crocker et al., 2016).

Given the above background, we conducted two web-based experiments to examine the intercomprehension of microsyntactic units (specific constructions between the lexicon and the grammar, idiomatic properties of which are closely tied to syntax, see Avgustinova and Iomdin, 2019) in context under different input conditions: (1) spoken and (2) written. Each experiment included two tasks: free translation and multiple choice. Native Russian speakers participated in the experiments covering five closely related Slavic languages (Belarusian, Bulgarian, Czech, Polish, and Ukrainian). We examined the participants' intercomprehension performance through accuracy. We calculated Pearson correlations of the accuracy values with phonologically weighted Levenshtein distance (PWLD), orthography-based Jaccard similarity, and surprisal estimates from Automatic Speech Recognition models, namely Wav2Vec2-Large-Ru-Golos-With-LM (Bondarenko, 2022) and Whisper Medium Russian and language models, namely ruBERTa-large and ruGPT3large (Zmitrovich et al., 2023).

Figure 1 shows the accuracy results for both experiments. In general, we found that spoken input led to higher accuracy values in both tasks except those in the multiple choice task for Ukrainian and Bulgarian, suggesting that the written modality might introduce a confounding factor. As surprisal from ruBERTa-large and PWLD showed stronger correlations in both tasks, we only present those factors in relation to the free translation and the multiple choice tasks, as shown in Table 1. We observed significant correlation of free translation accuracy for all languages together and for Ukrainian individually. As for multiple choice accuracy, significant correlations with PWLD were observed when pooling all languages together, as well as for all languages individually except Belarusian. We also observed stronger correlations in the experiment with written input, especially for the multiple choice task. Overall, this study underscores the predictive potential of surprisal and linguistic distances in multilingual intercomprehension experiments, providing valuable insights for the field of computational linguistics. Future research should expand to diverse language groups to validate these findings and explore their broader applicability.

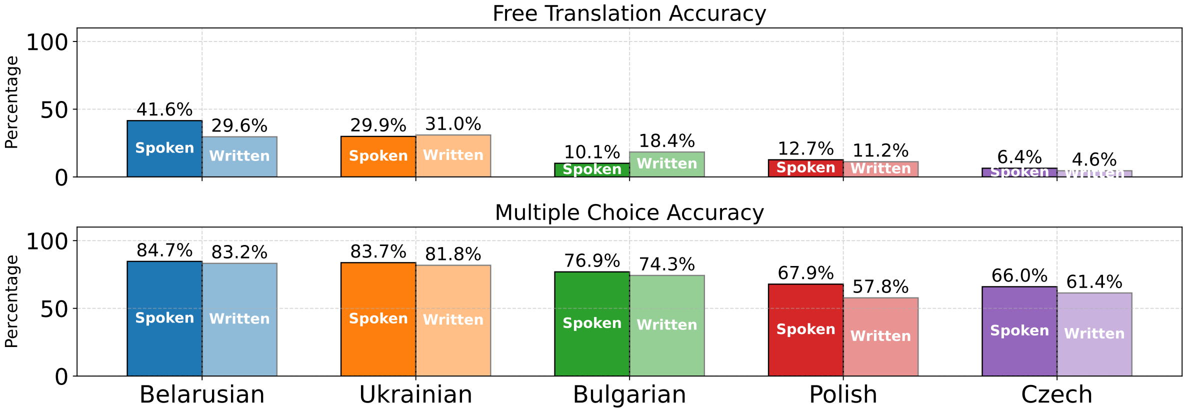

Figure 1: Experimental results for both tasks.

	Free Trans with ruBERTa-large surprisal		Multiple Choice with PWLD	
Language	**Written**	**Spoken**	**Written**	**Spoken**
Belarusian	-0.06	-0.03 (NS)	-0.2 (NS)	-0.15 (NS)
Bulgarian	-0.17 (NS)	0.00 (NS)	-0.42**	-0.33*
Czech	-0.23 (NS)	-0.06 (NS)	-0.28*	-0.25 (NS)
Polish	-0.21 (NS)	-0.16 (NS)	-0.38**	-0.45***
Ukrainian	-0.25*	-0.06 (NS)	-0.50***	-0.43***
All	-0.38***	-0.38***	-0.42***	-0.39***

Note: * = p $<$ .05, ** = p $<$ .01, *** = p $<$ .001, NS = Non-significant

Table 1: Pearson correlation of predictors with accuracy of participants' responses

References

Avgustinova, T., & Iomdin, L. (2019, September). Towards a typology of microsyntactic constructions. https://doi.org/10.1007/978-3-030-30135-4_2

Bondarenko, I. (2022). Xlsr wav2vec2 russian with 2-gram language model by ivan bondarenko.

Crocker, M., Demberg, V., & Teich, E. (2016). Information density and linguistic encoding (ideal). *Künstliche Intelligenz*, *30*, 77–81.

Gooskens, C., & Swarte, F. (2017). Linguistic and extra-linguistic predictors of mutual intelligibility between germanic languages. *Nordic Journal of Linguistics*, *40*, 123–147.

Gooskens, C., & van Bezooijen, R. (2013). Lexical and orthographic distances between germanic, romance and slavic languages and their relationship to geographic distance (wilbert heeringa, jelena golubovic, charlotte gooskens, anja schüppert, femke swarte & stefanie voigt). https://api.semanticscholar.org/CorpusID:6289144

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. *Second meeting of the north american chapter of the association for computational linguistics*.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Möller, R., & Zeevaert, L. (2015). Investigating word recognition in intercomprehension: Methods and findings. *Linguistics*, *53*.

Vanhove, J., & Berthele, R. (2015). Item-related determinants of cognate guessing in multilinguals. *Crosslinguistic Influence and Crosslinguistic Interaction in Multilingual Language Learning*, *95*, 118.

Wichmann, S., Holman, E., Bakker, D., & Brown, C. (2010). Evaluating linguistic distance measures. *Physica A: Statistical Mechanics and its Applications*, *389*, 3632–3639. https://doi.org/10.1016/j.physa.2010.05.011

Zmitrovich, D., Abramov, A., Kalmykov, A., Tikhonova, M., Taktasheva, E., Astafurov, D., Baushenko, M., Snegirev, A., Shavrina, T., Markov, S., Mikhailov, V., & Fenogenova, A. (2023). A family of pretrained transformer language models for russian.