Predictability of Microsyntactic Units across Slavic Languages: A Translation-based Study

Maria Kunilovskaya, Iuliia Zaitova, Wei Xue, Irina Stenger, and Tania Avgustinova

University of Saarland
izaitova@lsv.uni-saarland.de

Abstract

The paper presents the results of a free translation experiment, which was set up to explore Slavic cross-language intelligibility. In the experiment, native speakers of Russian were asked to read a sentence in one of the five Slavic languages and return a Russian translation of a highlighted item. The experiment is focused on microsyntactic units because they offer an increased intercomprehension difficulty due to opaque semantics. Each language is represented by at least 50 stimuli, and each stimulus has generated at least 20 responses. The levels of intercomprehension are captured by categorising participants' responses into seven types of translation solutions (paraphrase, correct, fluent_literal, awkward_literal, fantasy, noise, and empty), generally reflecting the level of the cross-linguistic intelligibility of the stimuli. The study aims to reveal linguistic factors that favour intercomprehension across Slavic languages. We use regression and correlation analysis to identify the most important intercomprehension predictors and statistical analysis to bring up the most typical cases and outliers. We explore several feature types that reflect the properties of the translation tasks and their outcomes, including point-wise phonological and orthographic distances, cosine similarities, surprisals, translation quality scores and translation solution entropy indices.

The experimental data confirms the expected gradual increase of intelligibility from West-Slavic to East-Slavic languages for the speakers of Russian. 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 Introduction

Cross-linguistic intercomprehension (receptive multilingualism) is defined as a phenomenon where speakers of different but related languages can communicate without studying each other's language (Trudgill, 2003). It can be viewed as specific cognitive conditions that tap into the mechanisms of human language processing (Meulleman and Fiorentino, 2018). Previous studies have focused on various aspects of intercomprehension within different language groups (Gooskens and Swarte, 2017; Stenger et al., 2017; Jagrova et al., 2018).

Some studies (Zaitova et al., 2024b,a) have looked at cross-linguistic intelligibility of functional multiword expressions with non-compositional semantics, called microsyntactic units (MSUs) (Avgustinova and Iomdin, 2019). MSUs can be grouped with prepositions, conjunctions, particles and other such word classes based on their function in the sentence. They are an interesting object for language processing studies because they are often important as discourse structuring items, signalling relations between clauses or conveying the speaker's attitude. Their intelligibility implies at least some understanding of the underlying proposition. Besides, MSUs present an additional difficulty for comprehension, especially across languages, because their meaning cannot be inferred from the components. An example of MSU in English is *all the same* or in Russian тем не менее (translit.: "tem ne menee", "nevertheless").

The exact mechanisms of intercomprehension

Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), pages 313–322

employed to process MSUs under various cross-linguistic conditions are still under-researched. To address this gap, our study presents the analysis of a free translation experiment in which native speakers of Russian translated MSUs from five Slavic languages (Czech, Polish, Bulgarian, Belarusian, and Ukrainian, hereinafter referred to as source languages) into Russian. We only used the data if the participants reported no training or exposure to the respective Slavic language.

The study aims to assess the level of intelligibility of the five Slavic languages for Russian speakers and to reveal the factors contributing to it. To this end, the translation solutions offered by native Russian speakers when rendering foreign MSUs into Russian are analysed. We employ several computational features (phonological distance, cognitive metrics, and translation quality scores) and provide a quantitative and qualitative description of Slavic MSU intelligibility as manifested by the participants' responses in the translation experiment.

It is expected that the East-Slavic languages (Belarusian and Ukrainian) would return the highest degree of intercomprehension, i.e., they would have the lowest difficulty in translation because Russian also belongs to the East-Slavic languages, followed by the South-Slavic Bulgarian (due to the use of Cyrillic script), with the Latin script-based West-Slavic languages (Czech, Polish) demonstrating the highest difficulty for the participants. Generally, translation difficulty indicators are expected to be reliable predictors of translation quality, i.e., of the outcomes of the translation task in this study[1].

2 Free Translation Experiment

Data collection: Platform, task and participants. The free translation experiment was held online[2] and aims to measure the degree of intelligibility of the targeted MSUs in the Slavic languages for Russian native speakers. The targeted MSU items come from a multi-parallel set, centred on Russian, which makes them comparable across the languages involved. In total, the experiment involved 126 unique participants without prior knowledge of the Slavic language they were translating from and 6,579 responses. The translation tasks for each Slavic language include between 50 and 60 unique sentences containing one of the target items. The study engaged from 101 to 121 native Russian participants per Slavic language who did not have any formal knowledge of that language. Table 1 provides basic descriptive statistics of the experimental data and participants. As can be seen from the table, the data is well-balanced across the languages in terms of the number of phrases and their part-of-speech category (PoS). There is approximately the same number of unique participants per language and the same number of responses per phrase.

	MSUs	ppt.	ppt./task	MSUs/PoS
CS	60	121	24.2±4.7	12.0±0.0
PL	50	116	23.1±5.3	10.0±1.3
BG	56	122	24.4±6.5	11.2±0.4
BE	57	121	24.4±5.4	11.4±0.5
UK	59	101	20.5±4.0	11.8±0.4

Table 1: Quantitative parameters of the free translation experiment. Abbreviations: ppt. (participants), CS (Czech), PL (Polish), BG (Bulgarian), BE (Belarusian), UK (Ukrainian)

Annotation of translation solutions and intelligibility scores. The participants' responses from the free translation experiment were categorised into seven groups of translation solutions reflecting the types of linguistic behaviour as well as the degree of understanding. These categories are explained below (in the order of decreasing intelligibility of the annotated response):

correct: a translation variant, which coincides with the reference ('gold') translation in cases where the available literal translation is different from the gold translation (otherwise, the response is categorised as 'fluent_literal'); it is the most expected standard solution that signals good understanding of the source phrase or even sentence,

fluent_literal: an acceptable translation variant, which coincides with both gold and literal translations; the cases where exploiting the cross-linguistic parallels yields good results,

paraphrase: a translation variant, which does not coincide with either gold or literal translation but faithfully renders the meaning of the

[1] Our code and datasets are available at https://github.com/SFB1102/b7-c4-slavic-translation-nodalida2025.

[2] https://intercomprehension.coli.uni-saarland.de/en/

source phrase; this can be a less expected descriptive response,

awkward_literal: this is a type of literal translation which is neither fluent_literal nor semantically incorrect, a translation technique to fall back to perceived cross-lingual similarities,

fantasy: a translation variant, which misrepresents the content of the source in the target language signalling lack of understanding,

noise: an irrelevant input, which does not allow to infer any specific translation solution; noisy solutions sometimes include comments like 'I have no idea' and 'I don't understand',

empty: no input provided indicating that the participant could not come up with a translation solution in the given time.

Note that this categorisation is developed for the purposes of this study and does not reflect translation quality of the participants' responses.

As can be seen from the description, the categorisation relied on existing gold and literal translations. The gold translations for the MSUs were extracted from the parallel subcorpora of the Russian National Corpus[3] and of the Czech National Corpus[4] with Russian as a target language (for more details see Zaitova et al., 2024a). The literal translations were generated by GPT-4 (22 July 2024) for isolated MSUs, i.e., for MSU outside of their context. To obtain literal translations, we used a prompt that included the task description "Return a literal word-for-word translation for a phrase in one of the Slavic languages into Russian.", a one-shot example in Czech and the task itself containing the name of the stimulus language and the phrase to translate. Automatic literal translations were preferred to human-generated literal translations to avoid subjective biases with regard to what was a literal translation. The sanity of the GPT-4 literal translations was controlled manually on an approximately 20% sample from each of the stimulus languages. The participants' responses were first pre-annotated for 'empty', 'correct', 'fluent_literal' and 'awkward_literal' categories because these annotations could have been filled in automatically based on matching gold and/or literal translations (see their description above). Two human annotators – trained linguists specialising in the Slavic languages and native speakers of Russian – contributed annotations for the remaining categories following formal and exemplified annotation guidelines. The annotators had access to gold and literal translations, as well as to the source language contexts. Conflicting annotations were resolved in a post-annotation discussion session.

To represent the overall intelligibility of the MSUs in a stimulus language for a Russian speaker, we assigned intelligibility weights to the annotated translation solutions on the following scheme: 'correct': 7, 'fluent_literal': 6, 'paraphrase': 5, 'awkward_literal': 4, 'fantasy': 2, 'noise': 0, 'empty': 0. The higher weights indicate greater intelligibility. The aggregate **intelligibility score** for each MSU item was calculated as a sum of weighted response probabilities across all responses for that stimulus. For example, the probabilities of responses for the Belarusian particle ледзьве не [hardly] had probabilities of the translation solutions distributed as follows: 0.0625, 0.0, 0.0625, 0.03125, 0.40625, 0.125, 0.3125. The sum of weighted probabilities is 1.6875.

3 Feature Extraction and Regression Analysis

Feature extraction. Generally, we explored four types of features: (a) surprisal values and (b) cosine similarities, both based on a pre-trained Transformer model, (c) Phonologically Weighted Levenshtein Distance (PWLD), and (d) automatic translation quality scores. These features were extracted for every source language items using gold and literal translations. We provide additional details on feature calculation below. Note that contextualised items were required when extracting some of the features, namely surprisals, cosine similarities, and automatic quality scores. Recall that the literal translations from GPT-4 were isolated phrases, not entire sentences. Therefore, we generated sentence-level contexts for these items by replacing them with the GPT-4 literal translations in the contexts from the parallel corpora.

(a-b) Transformer-based features: Surprisal and cosine similarity values reported in this study were generated using ruRoBERTa-large model (Zmitrovich et al., 2024)[5], a dedicated Russian language

[3] https://ruscorpora.ru/en/

[4] https://www.korpus.cz/

[5] https://huggingface.co/ai-forever/ruRoberta-large

Transformer[6]. To get a surprisal value for an MSU, we summed up surprisals of its components. The sentence-level surprisals are averaged across all words in a sentence, with the word-level surprisal being a sum of subword token surprisals. Cosine similarities were calculated using MSU embeddings that were mean-pooled across word-level embeddings of MSU components. The word embeddings were generated from subword representations using *minicons* python library.[7] Care was taken to minimise the number of extraction errors caused by mismatching tokenisation for isolated and contextualised MSUs, and by overmatching MSU components in a sentence. Specifically, we extracted surprisal values for the source, gold and literal MSUs themselves (surprisal_stim, surprisal_gold and surprisal_lit) and average surprisals for the sentences containing them (surprisal_stim_sent, surprisal_gold_sent and surprisal_lit_sent). The cosine similarity was calculated between (1) the stimulus source items in the five Slavic languages and their gold translations (cosine_stim_gold), and (2) the stimuli and their literal translations (cosine_stim_lit).

(c) PWLD: PWLD is a metric of weighted phonological similarity based on the Levenshtein distance between two phonemic sequences (Fontan et al., 2016). It takes into account the cost of each phoneme substitution given their phonemic features. We use an adaption of the PWLD proposed in Abdullah et al. (2021). PWLD is more suitable for cross-linguistic analysis than Levenshtein Distance because PWLD can catch more fine-grained phonological similarities. For example, in the pair of Czech and Russian cognates *ucho /u x o/* and ухо */u x ɔ/*, where phonemes */o/* and */ɔ/* are very similar to each other, PWLD would capture this similarity more effectively compared to Levenshtein Distance. To obtain the IPA transcriptions of all stimuli, we used *CharsiuG2P*, a transformer-based tool for grapheme-to-phoneme conversion (Zhu et al., 2022). We extracted PWLD scores between (1) the stimulus items and their literal translations (pwld_stim_lit), (2) the stimulus items and their gold translations (pwld_stim_gold), and (3) gold and literal translations (pwld_gold_lit).

(d) Automatic translation quality scores: We use scores from the reference-based and reference-free pre-trained COMET models[8]. The reference-based score was used to generate translation quality scores for literal translations, with the gold translation as reference. Additionally, we used reference-free quality scores (translation quality estimation scores) for the gold (qe_gold), literal (qe_lit), and participants' translations (eval_lit).

MSU translation entropy as an alternative to intelligibility score. The intelligibility score is based on annotated translation solutions, and thus takes into account **types of responses** abstracting from the individual choices. A more straightforward approach to judge about translation difficulty of an item is to calculate the its translation entropy from the distribution of valid translation variants seen in the data. We used the Shannon entropy formula:

$$H = -\sum_{i=1}^{n} p_i \log_2(p_i) \quad (1)$$

where p_i denotes the probability of the i-th unique response, and n denotes the total number of unique responses. The responses annotated as noise or empty were considered as having a `None` value. Shannon entropy captures the unpredictability of responses and can be interpreted as a measure of translation task difficulty: the higher the entropy, the more difficult the translation task is (Wei, 2022). It can also be views as a measure of literality: low entropy signals conditions for more automated literal translation (Carl and Schaeffer, 2017).

In sum, the analysis is based on 14 features shown in Appendix A. The Appendix reflects Pearson correlation of each feature with the entropy and intelligibility score for the source MSUs in each language, highlighting indicators that returned significant results. It can be seen that at least in terms of univariate analysis intelligibility scores are better aligned with the proposed features than entropy.

Regression analysis. The relevance of the features for intercomprehension was explored through their ability to predict the intelligibility score in a regression setup. The regression

[6]We also tried other Russian transformers such as `https://huggingface.co/ai-forever/rugpt3large_based_on_gpt2`, which returned similar results (omitted here for brevity).

[7]`https://pypi.org/project/minicons/`

[8]`https://huggingface.co/Unbabel/wmt22-comet-da` and `https://huggingface.co/Unbabel/wmt22-cometkiwi-da` respectively, described in Rei et al., 2022

was performed using Support Vector Machine algorithm (SVR) as implemented in *scikit-learn*.[9] The performance of SVR is reported in terms of Pearson's correlation coefficient (r) and Mean Absolute Error (MAE) with corresponding two-sided standard deviation across the 10 runs of the experiment ($\pm$). The error reported for intelligibility score as the response variable across all languages was lower than can be obtained by predicting the mean of the scores. This was not the case for entropy as an alternative response variable. Feature selection was performed using the Recursive Feature Elimination (RFE) technique, which iteratively applied a linear regressor to the feature space, eliminating the least important feature in each iteration until the desired number of features (here, N was arbitrarily set to 5) was reached.

4 Results and Discussion

4.1 Translation Solutions and Intelligibility

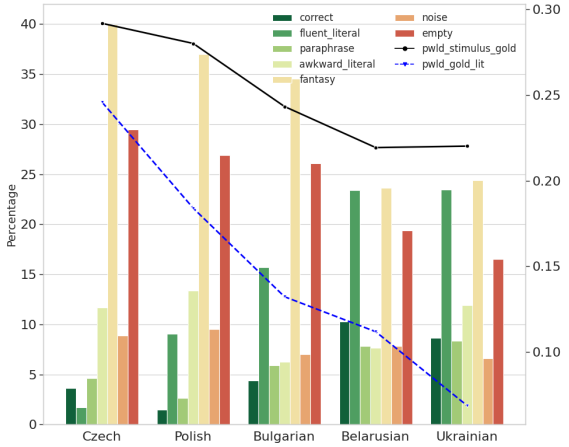


Figure 1: Bars for translation solutions and line plots for mean PWLD between original and gold MSUs (pwld_stim_gold), and between gold and literal variants (pwld_gold_lit), with PWLD values on the left y-axis. The greener end of the spectrum marks more successful translation task completion.

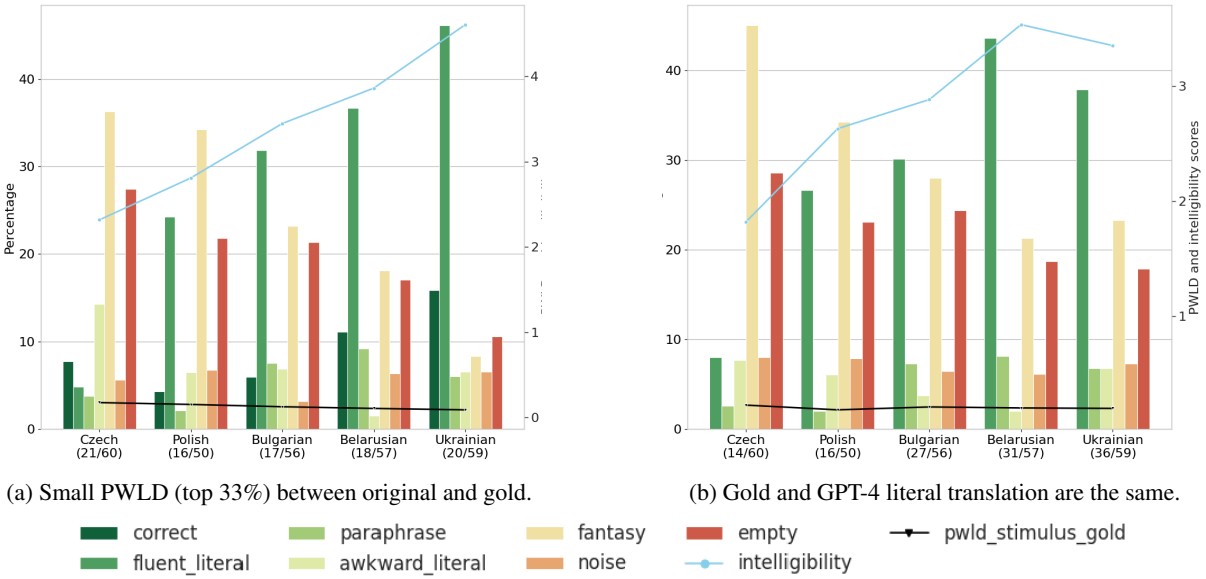
Figure 1 shows the distribution of translation solutions for each source language. The translation solutions are colour-coded and ordered based on the declining degree of intelligibility from the green end of the spectrum towards red. It can be seen that the percentage of correct translations (height of greener bars) increases from left to right across the languages. The intelligibility of the Slavic languages for speakers of Russian (as one can see from the bar charts) increases from left to right, i.e., from Czech to Ukrainian.

The lines in Figure 1 represent the PWLD values (i) between the original MSUs in Slavic languages and their gold translations attested in parallel corpora (pwld_stim_gold; solid black line), and (ii) between gold and literal translations (pwlf_gold_lit; dashed blue line). The lower the PWLD values, the more similar the items are. As shown in the figure, both lines have a clear left-to-right downward pattern confirming the intuitively expected relation between the cross-lingual formal similarity and intelligibility captured by the distribution of translation solutions. That is, when stimulus items have smaller distances to gold translations (and between gold and literal translations), the participants are more likely to return a higher proportion of acceptable translation solutions (correct, paraphrase or literal) and there are fewer fantasy, noise and empty responses.

The difference in slopes of the two lines can be interpreted as reflecting the properties of the automatically generated literal translations. GPT-4 generated literal translations that were closer (lower PWLD) to the gold translations for Ukrainian than for Belarusian. The analysis of distances between stimulus MSUs and literal translations for these languages shows that GPT-4 variants in Russian for Ukrainian items were more distant from the stimulus than the Russian translations for Belarusian items. This might reflect the relations between East-Slavic languages, where for the Ukrainian items it was difficult to find more literal Russian variants than gold translations.

To further explore the literal translation as an intercomprehension strategy, we used two approaches to identify stimulus MSUs that might be more suitable for literal cross-comprehension strategy: (a) items with small PWLD between stimulus and gold translation, and (b) items, where GPT-4 returned translations identical to the professional gold translations. Figure 2 shows which types of translations were offered for the Slavic MSUs extracted by each sampling method. The complementary line plots show the average intelligibility scores and the stimulus-to-gold PWLD values across each MSU sample. The sample in Figure 2a is based on the top 33% of original MSUs (the cut-off is selected arbitrary) that

[9] https://scikit-learn.org/1.5/modules/generated/sklearn.svm.SVR.html



(a) Small PWLD (top 33%) between original and gold.

(b) Gold and GPT-4 literal translation are the same.

Figure 2: Two approaches to define suitable conditions for literal translation. Translation solution bars and mean intelligibility scores across the stimulus MSUs in each sample. Flat black line indicates that stimulus-to-gold PWLD is about the same level across languages on average. Brackets have the number of sampled stimuli to their total for each sampling method.

have the smallest distance to the gold translations. For these items, the intercomprehension pattern is clear: MSUs with the same cross-linguistic distance are more successfully processed in East-Slavic than in West-Slavic languages. Although the Czech data in our experiment offered as many opportunities (21 MSUs) for literal comprehension as Ukrainian (20 MSUs), the participants failed to recognise these similarities. We can hypothesise that the Latin script can introduce some of the confusion. In Figure 2b, the literal-translation-friendly sample includes MSUs, for which GPT-4 returned the same Russian variants as used in gold translations. This plot highlights the differences between Belarusian and Ukrainian as processed by GPT-4 and by the participants (compare lighter-green bars of fluent_literal translations for these languages). The participants did not see the fluent Russian correspondences for Ukrainian items picked by GPT-4 and returned fewer fluent translations and more mistranslations (light-yellow phantasy bars) than for Belarusian in this sample. For other languages, the distance between gold and literal established by GPT-4 was proportional to the participants' success in the translation task.

Figure 3 shows the distribution of intelligibility scores for each source language. The mean score across all MSUs (red diamonds) increases

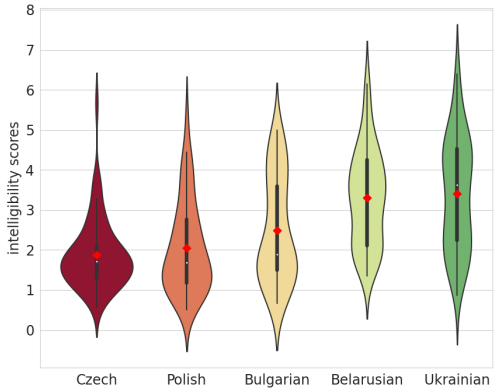


Figure 3: Distributions of the intelligibility scores. Red diamonds are means; the dark stripes with a white dot inside violins represent 25th, 50th, and 75th percentiles.

from left to right (from Czech to Ukrainian), which confirms previous findings and is intuitively expected. The scores are more homogeneous and centred around the low mean value for the less cross-intelligible West-Slavic languages, especially Czech. The distribution of intelligibility scores for the Ukrainian MSUs is more spread, with a bimodal tendency. It suggests that some Ukrainian MSUs are very intelligible, while others trigger intercomprehension difficulties and mis-

language	Pearson	MAE	nobs
Czech	0.21±0.43	0.63±0.21	60
Polish	0.23±0.50	0.90±0.30	50
Bulgarian	0.50±0.35	0.85±0.26	56
Belarusian	0.34±0.53	0.87±0.16	57
Ukrainian	0.62±0.31	0.98±0.34	59

Table 2: Regression results on intelligibility score for the top five language-specific predictors.

language	Pearson	MAE	nobs
Czech	0.23±0.36	0.40±0.10	60
Polish	0.19±0.55	0.51±0.17	50
Bulgarian	0.32±0.38	0.58±0.13	56
Belarusian	0.36±0.51	0.52±0.20	57
Ukrainian	0.65±0.37	0.52±0.21	59

Table 3: Regression results on entropy for the top five language-specific predictors.

translations.

4.2 Predicting Intelligibility via SVR

Table 2 shows correlations using the five features which returned the highest results for each language described. The intelligibility scores for the MSUs in the Cyrillic-based South- and East-Slavic languages are not only consistently higher than in the West-Slavic languages (see Figure 3) but also more predictable. Bulgarian and Ukrainian have the Pearson correlation coefficients 0.50 and 0.62, while the values of Pearson r for Polish and Czech do not exceed 0.23. For Belarusian (as well as for Ukrainian and Bulgarian) adding more features (up to a certain level) yield higher results. However, for West-Slavic languages the performance is unstable, and new features often introduce noise. The correlations on all features are considerably lower, especially for West-Slavic languages.

The regression results on MSU translation entropy as the learning target are 2% higher for Czech, Belarusian and Ukrainian but much lower for Polish and Bulgarian (see Table 3).

The variation in performance on the two variables describing the participants' translation choices (intelligibility scores and MSU translation entropy) is due to the lack of consistency in their relations across the Slavic languages. The Pearson correlation coefficient (r) between the entropy of translation variants and intelligibility score ranges from -0.799 (Ukrainian) to -0.325 (Czech) at $p < 0.05$. Figure 4 shows the regression lines fitted for each language separately and in combination. The entropy values are on average higher for Czech and Polish (2.70 and 2.47) than for Belarusian and Ukrainian (2.19 and 2.12) but for Czech and Polish they are less associated with intelligibility judging by the slopes and univariate r. It means that the participants' responses were less more varied across functionally similar MSUs in the West-Slavic languages than in Ukrainian.

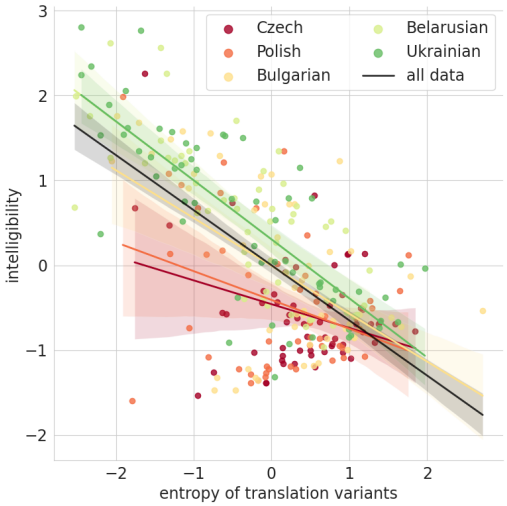


Figure 4: Relation between entropy of translation variants and intelligibility score for Slavic MSUs.

Low entropy scores characterise cases where the participants returned only a few unique responses and the probability distribution of these responses is skewed towards one type of translation solution. In other words, participants largely agreed on a Russian rendition for a given MSU.

(1) For example, a Ukrainian conjunction чим більше *(the more;* чем больше*)* is formally very similar to the gold Russian variant (PWLD=0.085) and has a low entropy of 0.569 based on the three types of solutions: correct (чем больше), fantasy (больше) and *empty*. The probability of the first variant is 0.9, and the intelligibility has a maximum value of 6.4 across all MSUs.

For East-Slavic languages, this consensus often meant successful task completion, i.e., high intelligibility. The ratio of MSUs with the lower-than-average entropy and higher-than-average in-

telligibility was 36.8% and 42.4% for Belarusian and Ukrainian, respectively. For West-Slavic languages it does not exceed 24%. However, low entropy can also signal lack of comprehension for West-Slavic languages: in another 24% of cases (more for Czech) lower-than-average entropy was linked to lower-than-average intelligibility.

(2) The Czech discourse marker *lze rici (it can be said,* можно сказать*)* had a low entropy (D=1.849) and below average intelligibility of 0.941. Despite the formal distance for this MSU was below the Czech average (0.260 vs. 0.287), only one response was correct, and 65% of participants did not come up with any solution within the given time.

The next most important predictor of a different type is the formal distance between original MSUs and their gold translations (pwld_stim_gold). It is reasonable to expect that smaller original-to-gold PWLD would be negatively correlated with intelligibility. While this general trend is observed in our data (see row 1 in Table 4b in Appendix A), it is less expressed for West-Slavic languages. Figure 2a shows that the same level of PWLD results in lower intelligibility for them. Formal similarities between West-Slavic languages and Russian are more often false friends or prompt awkward solutions. Hence, PWLD is a less reliable predictor for intelligibility of West-Slavic MSUs.

(3) The Czech particle *nejen ze (not only from,* не то что*)* has a low PWLD=0.161 (Czech average 0.286) and relatively low intelligibility (1.083 vs. average 1.872). For this item participants returned a variety of false literal solutions (e.g. неужели, нужен ли, не один же).

Another factor that correlates with the intelligibility of MSUs both within and across Slavic languages is the context sentence complexity. This property of the translation task is captured by the surprisal of the source or translated sentence (surprisal_stim_sent and surprisal_gold_sent). These features do not return significant correlations with intelligibility in univariate analysis for all languages but they are seen among the most informative features (except Belarusian).

Other features either are not consistently selected among the strong predictors and/or do not demonstrate a significant correlation with intelligibility in univariate analysis.

Thus, the analysis of the combinations of strong predictors (Table 5, Appendix A) and the correlation analysis outcomes suggest the following conditions for the intelligibility of Slavic MSUs for Russian speakers. We have seen that the most important role is played by the participants' perception of the similarities, their ability to recognise and interpret them, captured by the entropy of translation variants. Then, the scale of these similarities between the languages matters. It is reflected by the point-wise PWLD distance between original MSU and its gold translation. Finally, average context sentence surprisal in either source or target language is an important intelligibility factor for all languages. Although West-Slavic and East-Slavic languages demonstrate some group similarities, each language seems to have a unique set of MSU intelligibility conditions. For Ukrainian, for example, the stimulus-to-gold PWLD is strongly positively correlated with entropy ($r = 0.555$) and with the number of participants' variants ($r = 0.636$). That is, the smaller the PWLD, the fewer variants are generated by the participants, the lower the entropy of translation variants and the higher the intelligibility of original MSU. This pattern is not seen in any other Slavic language so clearly.

5 Conclusion

This study explored the intelligibility of microsyntactic units (MSUs) in Slavic languages. We conducted a free translation experiment where Russian-speaking participants were asked to translate MSUs from Czech, Polish, Bulgarian, Belarusian, and Ukrainian into Russian. The aim of the study was to measure intercomprehension levels manifested in participants' responses and to explore the factors related to intelligibility between similar languages.

As expected, the MSUs in East-Slavic languages (Belarusian and Ukrainian) were most intelligible, followed by the South-Slavic Bulgarian. West-Slavic languages (Czech, Polish) presented a greater challenge for our participants. We demonstrated that the level of intercomprehension was related to the ability of the participants to identify and interpret the cross-lingual similarities. Generally, fewer translation variants for an original MSU indicated higher intelligibility.

Lower phonological distance between the MSUs in the source and target languages was another well-correlated and typical predictor of intercomprehension. Intra-linguistically, MSUs that were offered in easier contexts returned higher intelligibility scores.

6 Limitations

The data is limited to one direction of intercomprehension. Our approach is highly contingent on how the formal distance between original and gold items is calculated and what is accepted as a literal translation from the Slavic languages into Russian. The context sentences were not controlled for complexity or topic across stimulus languages. The phonological distance calculations rely heavily on automated grapheme-to-phoneme conversion.

Acknowledgments

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102 and by Saarland University (UdS-Internationalisierungsfonds).

References

Badr Abdullah, Marius Mosbach, Iuliia Zaitova, Bernd Möbius, and Dietrich Klakow. 2021. Do Acoustic Word Embeddings Capture Phonological Similarity? An Empirical Study. In *Proceedings of Interspeech 2021*. pages 4194–4198.

Tania Avgustinova and Leonid Iomdin. 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*. Springer, pages 15–30.

Michael Carl and Moritz Jonas Schaeffer. 2017. Why translation is diffcult: A corpus-based study of non-literality in post-editing and from-scratch translation. *Hermes (Denmark)* 56:43–57.

Lionel Fontan, Isabelle Ferrané, Jérôme Farinas, Julien Pinquier, and Xavier Aumont. 2016. Using phonologically weighted levenshtein distances for the prediction of microscopic intelligibility. In *Annual conference Interspeech (INTERSPEECH 2016)*. page 650.

Charlotte Gooskens and Femke Swarte. 2017. Linguistic and extra-linguistic predictors of mutual intelligibility between germanic languages. *Nordic Journal of Linguistics* 40:123–147.

Klara Jagrova, Tania Avgustinova, Irina Stenger, and Andrea Fischer. 2018. Language models, surprisal and fantasy in slavic intercomprehension. *Computer Speech & Language* 53.

Machteld Meulleman and Alice Fiorentino. 2018. What is intercomprehension and what is it good for? In François Grin, Manuel Célio Conceição, Peter A. Kraus, László Marácz, Žaneta Ozoliņa, Nike K. Pokorn, and Anthony Pym, editors, *The MIME vademecum: Mobility and inclusion in multilingual Europe*, Artgraphic Cavin SA, pages 146–147. Hal-02497697.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), pages 578–585.

Irina Stenger, Klára Jágrová, Andrea Fischer, Tania Avgustinova, Dietrich Klakow, and Roland Marti. 2017. Modeling the impact of orthographic coding on czech–polish and bulgarian–russian reading intercomprehension. *Nordic Journal of Linguistics* 40(2):175–199.

Peter Trudgill. 2003. *Mutual intelligibility*, Edinburgh University Press, Edinburgh, page 91.

Yuxiang Wei. 2022. Entropy as a measurement of cognitive load in translation. In *AMTA 2022 - 15th Conference of the Association for Machine Translation in the Americas, Proceedings - Workshop on Empirical Translation Process Research*. volume 1, pages 75–86.

Iuliia Zaitova, Irina Stenger, Muhammad Umer Butt, and Tania Avgustinova. 2024a. Cross-linguistic processing of non-compositional expressions in slavic languages. In *Proceedings of the Workshop on Cognitive Aspects of the Lexicon@ LREC-COLING 2024*. pages 86–97.

Iuliia Zaitova, Irina Stenger, Wei Xue, Tania Avgustinova, Bernd Möbius, and Dietrich Klakow. 2024b. Cross-linguistic intelligibility of non-compositional expressions in spoken context. In *Proc. Interspeech 2024*. pages 4189–4193.

Jian Zhu, Cong Zhang, and David Jurgens. 2022. Byt5 model for massively multilingual grapheme-to-phoneme conversion. In *Annual conference Interspeech (INTERSPEECH 2022)*.

Dmitry Zmitrovich, Aleksandr Abramov, Andrey Kalmykov, Vitaly Kadulin, Maria Tikhonova, Ekaterina Taktasheva, Danil Astafurov, Mark Baushenko, Artem Snegirev, Tatiana Shavrina, et al. 2024. 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)*. pages 507–524.

Appendix A. List of predictors with correlation analysis outcome

Table 4: Association between translation task features and response variables. Asterisks indicate statistically significant results at the confidence level of 0.05. The tables are sorted to have the features with significant results across more languages on top. Features with the same number of significant results are sorted alphabetically.

(a) Pearson correlation coefficient between predictors and entropy of translation variants.

#	feature	Czech	Polish	Bulgarian	Belarusian	Ukrainian
1	pwld_stim_gold	0.314*	0.08	0.331*	0.333*	0.556*
2	surprisal_lit	-0.001	-0.088	0.043	0.385*	0.328*
3	pwld_stim_lit	0.144	0.042	0.238	0.3*	0.481*
4	cosine_stim_lit	0.067	0.02	-0.316*	-0.16	-0.327*
5	cosine_stim_gold	0.001	-0.094	-0.382*	-0.139	-0.377*
6	eval_lit	0.014	-0.046	-0.114	-0.284*	-0.099
7	surprisal_stim_sent	0.045	0.104	0.298*	0.258	0.076
8	surprisal_gold	0.123	-0.076	-0.08	0.177	0.393*
9	surprisal_stim	0.008	0.141	0.347*	0.101	0.18
10	surprisal_lit_sent	-0.197	-0.095	0.172	0.229	0.012
11	qe_lit	-0.126	-0.035	0.176	-0.107	-0.088
12	pwld_gold_lit	0.051	0.0	0.236	0.079	0.137
13	qe_gold	-0.091	-0.058	0.198	0.044	-0.008
14	surprisal_gold_sent	-0.066	-0.081	0.11	0.019	-0.019

(b) Pearson correlation coefficient between predictors and intelligibility scores.

#	feature	Czech	Polish	Bulgarian	Belarusian	Ukrainian
1	pwld_stim_gold	-0.305*	-0.375*	-0.432*	-0.384*	-0.594*
2	pwld_stim_lit	-0.304*	-0.29*	-0.469*	-0.211	-0.418*
3	cosine_stim_gold	0.008	0.233	0.438*	0.272*	0.438*
4	surprisal_stim_sent	-0.263*	-0.366*	-0.258	-0.426*	-0.083
5	eval_lit	0.16	0.276	0.268*	0.439*	-0.003
6	pwld_gold_lit	-0.153	-0.293*	-0.396*	-0.178	-0.099
7	surprisal_lit	-0.223	-0.09	-0.318*	-0.415*	-0.186
8	cosine_stim_lit	-0.099	0.129	0.224	0.213	0.388*
9	qe_lit	0.171	0.153	-0.069	0.304*	0.024
10	surprisal_gold	-0.039	-0.026	-0.004	-0.226	-0.308*
11	surprisal_lit_sent	-0.183	-0.135	-0.188	-0.37*	0.072
12	qe_gold	0.151	0.093	-0.097	0.111	-0.027
13	surprisal_gold_sent	-0.072	-0.001	-0.032	-0.137	0.034
14	surprisal_stim	0.2	-0.174	-0.236	-0.251	-0.187

Table 5: Language-specific selections of best intelligibility predictors (by RFE, N=5)

	feature names
Czech	surprisal_lit, surprisal_gold, surprisal_stim_sent, pwld_stim_lit, qe_gold
Polish	surprisal_stim_sent, surprisal_lit_sent, surprisal_gold_sent, cosine_stim_gold, pwld_stim_gold
Bulgarian	surprisal_stim, surprisal_lit, cosine_stim_gold, pwld_stim_lit, pwld_stim_gold
Belarusian	surprisal_stim, surprisal_gold, surprisal_lit_sent, surprisal_gold_sent, pwld_stim_gold
Ukrainian	surprisal_lit_sent, surprisal_gold_sent, pwld_stim_gold, qe_gold, qe_lit