Evaluating GPT Surprisal, Linguistic Distances, and Model Size for Predicting Cross-Language Intelligibility of Non-Compositional Expressions

Anonymous ACL submission

Abstract

Cross-language intelligibility is defined as the ability to understand related languages without prior study. This study investigates how and to what extent linguistic distances and surprisal values generated by GPT-based models predict cross-language intelligibility of microsyntactic units (MSUs), a type of non-compositional expression characterized by syntactic idiomaticity. We compare performance across two research questions: (1) How well do linguistic distances and surprisal values from GPT-based models predict intelligibility of non-compositional expressions? (2) Does model size impact prediction performance of GPT-based surprisal? The predictors were tested on two experimental conditions (spoken input vs. combined spoken-written input) and two tasks (free translation and multiple-choice) with native Russian participants translating MSUs across five Slavic languages: Belarusian, Bulgarian, Czech, Polish, and Ukrainian. Results revealed that although GPT-based surprisal is a significant predictor of MSU intelligibility, the most crucial predictor is linguistic distances, with variations based on experimental conditions and task types. Additionally, our analysis found no substantial performance gap between smaller and larger GPT models.

1 Introduction

Cross-language intelligibility refers to the ability of speakers to understand related languages without prior study (Doyé, 2005; Gooskens and van Heuven, 2021). It is influenced by phonological, lexical, and orthographic similarities, particularly among languages with close typological proximity (Gooskens and van Heuven, 2021; Stenger and Avgustinova, 2021). Speakers can recognize cognates, decipher grammar, and infer meanings, making comprehension or intelligibility across related languages achievable without any prior exposure to the language. Research on intelligibility has significant implications for language policy (e.g., designing language standards), education (e.g., transfer effects between languages), and human-machine interaction (e.g., multilingual NLP systems). Computational modeling of intelligibility helps us better understand and model cross-linguistic processing difficulty (Stenger et al., 2017b; Jágrová et al., 2018; Gooskens, 2024).

Cross-language intelligibility becomes significantly more challenging in case of *non-compositional expressions*, like microsyntactic units (Avgustinova and Iomdin, 2019). Non-compositional expressions have meanings that cannot be inferred from their individual components (Baldwin and Kim, 2010; Jackendoff, 2002; Kudera et al., 2023). Microsyntactic units, a specific type of non-compositional expression used as our experimental stimuli, are characterized by their *syntactic idiomaticity*, where the structure itself carries figurative meaning (Iomdin, 2015, 2016; Avgustinova and Iomdin, 2019). Examples of microsyntactic units in English are *'at the end of' 'to begin with'*[1].

Cross-language intelligibility of non-compositional expressions has been extensively explored in relation to various factors, but *linguistic distances* and *surprisal* are considered key indicators of how challenging an expression is to comprehend (Stenger et al., 2017a; Jágrová et al., 2018). Linguistic distance refers to the magnitude of differences between languages at the form level. It could capture similarities at various dimensions, including lexical, orthographic, phonetic, and phonological, among others. Surprisal quantifies how unexpected or informative a given linguistic element is to a perceiver. Formally, the surprisal of an event is defined as the negative logarithm of its probability (Demberg et al., 2012). Rooted in information theory and psycholinguistics, surprisal

[1] More examples of microsyntactic units in Slavic languages are given in Appendix A.

serves as a proxy for the difficulty of processing foreign expressions (Jágrová et al., 2018).

Among all linguistic distances available, we focus specifically on orthographic and phonological distances. This choice was motivated by the following considerations. First, the non-compositionality of the microsyntactic stimuli makes lexical distance measures less informative (Cutting and Bock, 1997; Wray, 2002). Second, phonetic distances, while relevant for language processing, are difficult to reliably measure in the context of unfamiliar languages (Best, 1995). Third, orthographic and phonological distances have been shown to be particularly relevant in studies of cross-language intelligibility (Vanhove and Berthele, 2015; Möller and Zeevaert, 2015; Gooskens and Swarte, 2017) and Slavic intercomprehension (Stenger et al., 2017a,b; Jágrová et al., 2018; Gooskens, 2024).

Regarding surprisal, its performance on predicting cross-language intelligibility could be influenced by the size of language models from which surprisal is derived. Recent advances in language modeling include large-scale transformer models like GPT (Radford and Narasimhan, 2018). While these large models excel in generating contextually rich sequences, it is often suggested that surprisal from smaller models predict human cognitive processes better (Oh and Schuler, 2023a,b; Vafa et al., 2024). Yet, previous findings on this topic investigated human reading time and native language comprehension. How model size relates to cross-language intelligibility, a comprehension across language instead of native comprehension, remains unknown. Therefore, in this study, we investigate how surprisal estimates from two monolingual Russian (RU) GPT models of different sizes (ruGPT-3-small and ruGPT-3-large) explain human performance when interpreting non-compositional expressions, in particular microsyntactic units, across foreign, but closely-related languages.

Lastly, although these factors have been previously shown to correlate with the intelligibility of non-compositional expressions (Zaitova et al., 2024a,b), it remains underexplored how these factors vary across inputs, e.g., spoken vs. written. To sum up, this study adresses the following research questions (RQs):

- **RQ1:** How well do linguistic distances and GPT-based surprisal predict cross-language intelligibility of non-compositional expressions in relation to different types of input?
- **RQ2:** Is the small variant of GPT model more effective than the large one in predicting intelligibility outcomes?

We conducted two experiments to evaluate the intelligibility of non-compositional expressions with spoken input only (Experiment 1), and with written input alongside spoken input (Experiment 2). Both experiments contain two tasks: free translation and multiple-choice question (MCQ). In the free translation task, participants need to listen or read a foreign expression presented in a sentential context and write a RU translation. In the MCQ task, participants need to select between a correct non-compositional translation and a literal, incorrect translation of the expression in the foreign language. RU native speakers were recruited as participants to translate the expressions from five Slavic languages, i.e., Belarusian (BE), Bulgarian (BG), Czech (CS), Polish (PL), and Ukrainian (UK). Slavic languages are traditionally divided into three branches: East Slavic (RU, BE, UK, etc.), West Slavic (PL, CS, Slovak, etc.), and South Slavic (BG, Croatian, Serbian, etc.) (Sussex and Cubberley, 2006). We assess cross-language intelligibility via a binary correctness measure (correct vs. incorrect) based on the nature of our stimuli: non-compositional microsyntactic units whose meanings cannot be inferred from their individual components. According to the prior work on intelligibility, we treat it as a all-or-nothing phenomenon: either participants grasp the idiomatic meaning, or they do not (Stenger et al., 2017a,b; Gooskens and van Heuven, 2021).

By combining linguistic distances and surprisal values as predictive factors, we aim to provide a comprehensive view of the interplay between structural similarity and cognitive difficulty. Our study contributes insights into psycholinguistic modeling and the role of model scale in predicting cross-language intelligibility, offering both theoretical and practical implications.

2 Methodology

2.1 Stimuli preparation

2.1.1 Written data

To prepare our non-compositional expression stimuli, we selected 60 most frequent microsyntactic units per target language from an existing dataset of RU microsyntactic units and their translational equivalents in BE, BG, CS, PL, and UK (Zaitova et al., 2023). Some examples of microsyntactic units are given in Appendix A. The dataset provides the unit in RU, its translational equivalent in the target Slavic language, and a contextual sentence in both languages with average lengths varying between 11 and 15 words. Limiting the stimuli to 60 units per language was aimed at minimizing the risk of participant fatigue on data quality.

2.1.2 Spoken data

We recorded the context sentences containing the target units using native speakers (one per target language) in self-paced reading sessions. All recordings were made in a controlled acoustic environment to ensure consistency across the samples. A 44.1 kHz sampling rate in an uncompressed format was used. Audio lengths averaged about 5–7 seconds. The speakers for BG, CS, and UK were female, while those for BE and PL were male due to difficulties finding female native speakers. The speakers' ages ranged from 21 to 29 (*mean*=25).

2.1.3 Literal translation options for the multiple-choice task

The MCQ task mentioned in Section 1 requires participants to choose between two options: a correct translation and a literal counterpart. The correct translations are described in Section 2.1.1, while the literal translation mimics the form of the stimulus but provides an inaccurate, but viable compositional translation of the expression. To create the literal translations, native RU speakers manually found word-by-word translations sourced from Glosbe (https://glosbe.com) and Vasmer's dictionary (https://lexicography.online/etymology/vasmer/). Having literal conterpart challenges participants to distinguish between non-compositional (correct) and literal (incorrect) options. Although a binary choice is limited, it served as a baseline measure for distinguishing idiomatic meanings from surface-level compositional interpretations. The literal options simulate a common cognitive strategy in cross-linguistic comprehension: mapping form to meaning even when it leads to less natural semantics.

2.2 Experimental setup

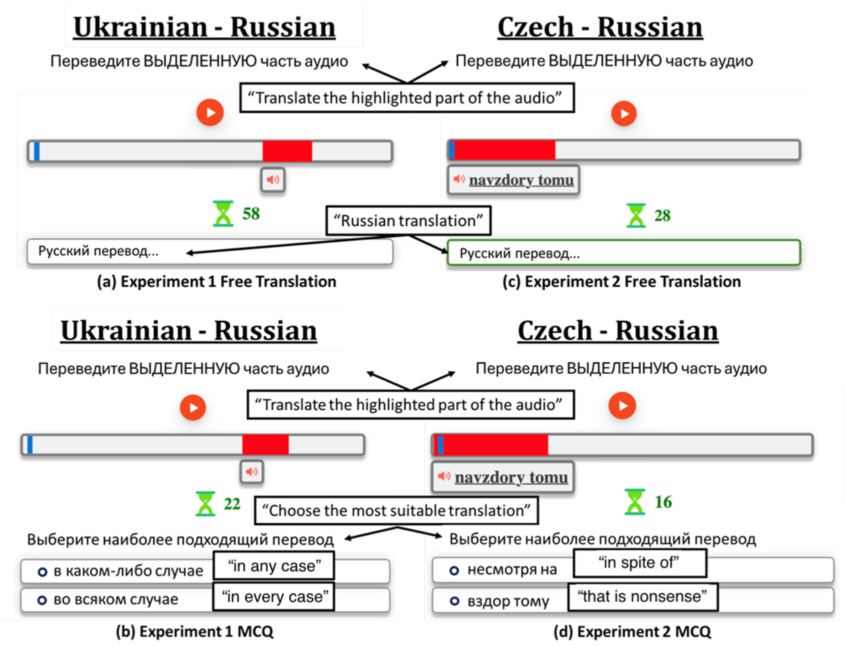
We conducted two web-based experiments with different types of input, namely Experiment 1 for spoken-only input and Experiment 2 for written input alongside spoken input, and with the two tasks (free translation and MCQ) mentioned in Section 1. The experiments were prepared via the website [thelinkisanonymized]. Before the experiments, participants first received instructions in RU detailing the procedure. After familiarizing themselves with the tasks, participants were required to register on the website and to complete a questionnaire in order to monitor their language background and to exclude those who had prior knowledge of the target languages, thereby maintaining the purity of the experiment's conditions.

An illustration of the two experiments and the two tasks is shown in Fig. 1. The only difference between the two experiments is whether participants were additionally presented with the written form of the test units, comparing Fig. 1 (a) and (b) for Experiment 1 (left panel) to Fig. 1 (c) and (d) for Experiment 2 (right panel). Note that participants were informed which language the test expression belonged to but were not told if their response was correct.

Further, in both experiments, participants were first presented with an audio clip containing the expression (highlighted in red bar) presented in its contextual sentence together with the free translation task, as shown in Fig. 1 (a) and (c), i.e., the upper panel. The time to enter the translation was based on a formula of 10 seconds per word in the test unit plus an additional 3 seconds per word in its context. Participants were allowed to replay each audio fragment of the whole contextual sentence and of the test unit up to three times, simulating real-life scenarios where listeners can ask speakers to repeat themselves.

After the free translation task, participants received the MCQ for the same expression. This ensured that participants attempted a genuine interpretation before choosing between correct and literal translations. MCQ is illustrated in Fig. 1 (b) and (d), i.e., the bottom panel. It asked participants to choose from two options in RU that they believed to be correct: (i) the correct non-compositional equivalent translation and (ii) an alternative word-by-word literal translation as explained in Section 2.1.3. The MCQ task aimed to assess participants' preference for the non-compositional (correct) translation over the literal (incorrect) one.

In total, each participant received 60 test units, each presented in a separate trial, together with their sentential context (in audio form). These 60 test units were evenly distributed across the five target languages. This means that each participant received 12 test units per target language, which is


Ukrainian - Russian
Czech - Russian
Переведите ВЫДЕЛЕННУЮ часть аудио
Переведите ВЫДЕЛЕННУЮ часть аудио
"Translate the highlighted part of the audio"
navzdory tomu
58
28
"Russian translation"
Русский перевод...
Русский перевод...
Ukrainian - Russian
Czech - Russian
Переведите ВЫДЕЛЕННУЮ часть аудио
Переведите ВЫДЕЛЕННУЮ часть аудио
"Translate the highlighted part of the audio"
navzdory tomu
22
16
"Choose the most suitable translation"
Выберите наиболее подходящий перевод
Выберите наиболее подходящий перевод
○ в каком-либо случае
"in any case"
○ во всяком случае
"in every case"
○ несмотря на
"in spite of"
○ вздор тому
"that is nonsense"

(a) Experiment 1 Free Translation

(c) Experiment 2 Free Translation

(b) Experiment 1 MCQ

(d) Experiment 2 MCQ

Figure 1: Task interface for the free translation and MCQ tasks received by Russian participants. The Czech test expression with written form in (c) and (d) is 'in spite of'. The green hourglass shows how many seconds are left for the participant to give their answer.

a random subset of the five subsets per language.

2.3 Participants

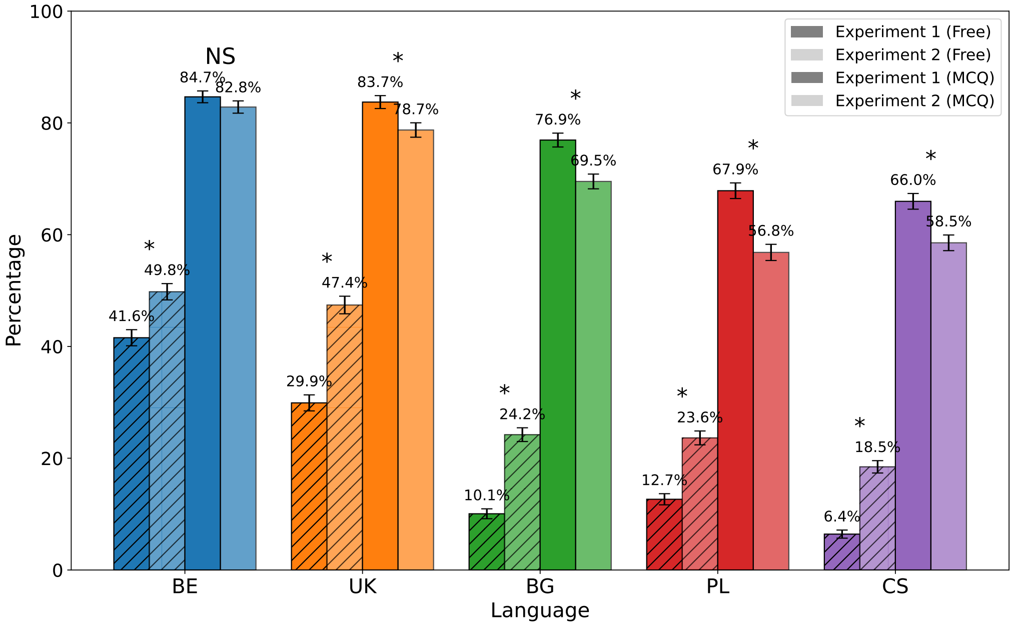
We recruited native RU speakers as our participants via Prolific (https://prolific.com), an online platform for research participant recruitment. Familiarity with the Latin script, which is used by CS and PL languages, was expected due to the English-language interface of Prolific. All of our participants provided informed consent and were assured to be anonymized in any published data. Participants with any prior knowledge of the target languages were excluded. For Experiment 1 (spoken-only input), we recruited 88 participants (26 males, 60 females, 2 identifying as other genders; age range 21-78 years, mean age 35). For Experiment 2 (spoken and written input), we recruited 118 participants (41 males, 76 females, 1 identifying as another gender; age range 18-59 years, mean age 32). There was also no overlap of participants in the two experiments. Having these large numbers of participants also aims to compensate the limited stimuli subset size.

2.4 Intelligibility scores, linguistic distances, and surprisal

The correctness of responses was considered as the intelligibility. For the free translation task, the responses were automatically considered correct if they matched allowed alternative answers in a predefined list. For instance, we allowed RU equivalents или что, что ли, или как as possible translations of UK чи що. The responses were further manually checked by a native RU speaker to include correct responses that could have been missed because of typos.

2.4.1 Linguistic distances

To address RQ1 regarding which distances are related to intelligibility, we extracted the orthographic and phonological distances (explained in Section 1) for each test unit in two different ways depending on the task. For the free translation task, we used the distance between the original expression and its correct non-compositional translation to RU. For the MCQ task, we measured how much closer the original expression was to its correct non-compositional translation in RU compared to its literal, word-by-word translation. A larger difference indicates that the true meaning is very different from the literal interpretation. For instance, in the UK expression все же (literally "everything or", but actually meaning "nonetheless"), we would expect a large difference since the true meaning differs substantially from the literal translation. The boxplots of linguistic distances, calculated independently from intelligibility scores, are presented per language in Appendix B.



Significance levels: * = p < .05, ** = p < .01, *** = p < .001, NS = Non-significant

Figure 2: Intelligibility scores of free translation and MCQ Responses in the two experiments (i.e., Experiment 1 referring to spoken-only inputs while Experiment 2 referring to combined spoken and written inputs). The languages are arranged in descending order by intelligibility scores.

Orthographic distance. We employed Levenshtein Distance which counts the minimum number of single-character operations (i.e., insertions, deletions, and substitutions) needed to transform one word into another (Levenshtein, 1966). It is worth noting that evaluating orthographic distance among Slavic languages is challenging due to their use of two writing systems – Latin and Cyrillic. To address this, we performed ISO 9 transliteration for CS and PL stimuli to convert them to Cyrillic, which is used by the other three target languages and RU. Levenshtein distance has been shown potential in analyzing intelligibility. For instance, (Stenger, 2019) found that Levenshtein distance of cognates is a reliable predictor of orthographic intelligibility of Slavic languages that use Cyrillic script. Also, as we mentioned in Section 2.3, our RU participants were not expected to know the correct orthographic pronunciation rules of the target languages, as they had not previously studied these languages. They might use their knowledge of Cyrillic and Latin scripts from exposure to RU and English (Prolific's interface language) to approximate the pronunciation of words written in Latin script.

Phonological distance. We employed Phonologically Weighted Levenshtein Distance (PWLD) which quantifies the distance between different phonemic sequences or word forms (Fontan et al., 2016). This distance extends the string-based Levenshtein Distance by considering the cost of each phoneme substitution based on their phonetic features like voicing and manner of articulation, making PWLD reflect fine-grained perceptual similarity. We employed the same adaptation of the original PWLD as the one proposed in Abdullah et al. (2021) which is based on PHOIBLE feature vectors (Moran and McCloy, 2019). For example, the pair of Czech and Bulgarian cognates: ucho /u x o/ and ухо /u x O/, the substitution cost would be lower than 0.5 (i.e., the maximum substitution). The phonemic transcriptions for all the data were obtained using CharsiuG2P, a transformer-based tool for grapheme-to-phoneme conversion (Zhu et al., 2022).

2.4.2 Surprisal values

In addition to linguistic distances, we extracted surprisal values to address RQ1. We developed a cascaded system that combines automatic speech recognition (ASR) and language modeling. This

Table 1: Mixed-Effects Model with GPT-Large Results

(a) Free translation Task (GPT-Large)

Predictor	Est.	SE	*z*	*p*
Main				
Intercept	−1.85	0.32	−5.79	< .001
PWLD	−0.58	0.13	−4.4	< .001
Levenshtein	−0.41	0.14	−2.9	0.004
GPT L	−0.54	0.2	−2.63	0.008
Written	1.23	0.33	3.76	< .001
South	−3.07	0.42	−7.23	< .001
West	−2.48	0.37	−6.79	< .001
2-way				
GL×Wr	0.46	0.14	3.27	.001
GL×S	0.55	0.49	1.12	0.26
GL×W	0.92	0.34	2.7	.007
Wr×S	1.1	0.32	3.47	< .001
Wr×W	0.59	0.26	2.28	.023
3-way				
GL×Wr×S	−1.19	0.37	−3.2	.001
GL×Wr×W	−0.4	0.26	−1.56	0.119

Note. GL = GPT Large, Wr = spoken+written input, S = South, W = West. Random effects variances: Source = 3.25, User = 3.88.

(b) MCQ Task (GPT-Large)

Predictor	Est.	SE	*z*	*p*
Main				
Intercept	2.07	0.18	11.45	< .001
PWLD	−0.43	0.08	−5.62	< .001
Levenshtein	−0.33	0.08	−4.15	< .001
GPT L	−0.31	0.13	−2.32	.021
Written	−0.29	0.19	−1.51	.131
South	−0.42	0.23	−1.84	.066
West	−1.05	0.19	−5.41	< .001
2-way				
GL×Wr	−0.03	0.12	−0.25	.805
GL×S	0.58	0.26	2.22	.027
GL×W	0.37	0.19	1.99	.047
Wr×S	−0.3	0.19	−1.57	0.12
Wr×W	−0.2	0.16	−1.25	0.21
3-way				
GL×Wr×S	−0.33	0.22	−1.51	.130
GL×Wr×W	0.09	0.16	0.57	.568

Note. GL = GPT Large, Wr = spoken+written input, S = South, W = West. Random effects variances: Source = 0.96, User = 1.03.

Table 2: Mixed-Effects Model with GPT-Small Results

(a) Free translation Task (GPT-Small)

Predictor	Est.	SE	*z*	*p*
Main				
Intercept	−1.81	0.32	−5.7	< .001
PWLD	−0.58	0.13	−4.43	< .001
Levenshtein	−0.41	0.14	−2.9	0.004
GPT S	−0.56	0.20	−2.73	0.006
Written	1.19	0.33	3.67	< .001
South	−3.11	0.42	−7.34	< .001
West	−2.53	0.36	−6.93	< .001
2-way				
GS×Wr	0.46	0.14	3.35	< .001
GS×S	0.57	0.46	1.23	.022
GS×W	1.05	0.35	3	0.003
Wr×S	1.1	0.32	3.49	< .001
Wr×W	0.61	0.26	2.37	0.02
3-way				
GS×Wr×S	−1.07	0.35	−3.06	0.002
GS×Wr×W	−0.37	0.26	−1.4	0.16

Note. GS = GPT Small, Wr = spoken+written input, S = South, W = West. Random effects variances: Source = 3.23, User = 3.88.

(b) MCQ Task (GPT-Small)

Predictor	Est.	SE	*z*	*p*
Main				
Intercept	2.10	0.18	11.60	< .001
PWLD	−0.43	0.08	−5.67	< .001
Levenshtein	−0.32	0.08	−4.12	< .001
GPT S	−0.31	0.13	−2.31	.021
Written	−0.29	0.19	−1.50	.135
South	−0.43	0.23	−1.86	.063
West	−1.07	0.19	−5.54	< .001
2-way				
GS×Wr	−0.01	0.12	−0.09	.927
GS×S	0.54	0.24	2.22	.027
GS×W	0.38	0.19	2.01	.044
Wr×S	−0.32	0.19	−1.71	0.08
Wr×W	−0.2	0.16	−1.24	0.21
3-way				
GS×Wr×S	−0.37	0.20	−1.82	.068
GS×Wr×W	0.05	0.16	0.34	.733

Note. GS = GPT Small, Wr = spoken+written input, S = South, W = West. Random effects variances: Source = 0.96, User = 1.03.

design is motivated by psycholinguistic models of spoken language comprehension, which assumes a layered process involving initial decoding of acoustic–phonological information followed by lexical integration (Marslen-Wilson, 1987; Cutler and Clifton, 1999; Friederici, 2002). By capturing both stages, our system approximates the cognitive processes RU speakers engage in when interpreting unfamiliar Slavic expressions. The cascaded system operates as a two-stage pipeline as described below:

1) Speech-to-Text: First, the ASR component enables us to simulate phonological decoding by transcribing foreign language speech into RU. Notably, ASR is only applied to the foreign language input and not to RU, ensuring the model mimics how RU listeners perceive foreign speech input. While ASR inevitably introduces errors, especially when processing unfamiliar languages, prior research demonstrates that such errors can approximate human comprehension difficulty in L2 contexts (Mirzaei et al., 2016). In this sense, ASR output does not merely add noise but reflects real-world variability in auditory processing. We used the Wav2Vec2-Large-Ru-Golos-With-LM model (Bondarenko, 2022) to convert speech input from foreign, target Slavic languages into RU text. This ASR component was specifically fine-tuned on the large-scale RU speech Sberdevices Golos dataset (Karpov et al., 2021), making it suited for emulating a native RU listener. Detailed model performance on RU can be found in Appendix D.

2) Surprisal Calculation: The second stage of the pipeline computes surprisal from the transcribed RU text described above. Surprisal, operationalized as the negative log-probability of a token given its preceding context, serves as a proxy for cognitive processing load: the less expected a word is in context, the higher its surprisal value. We compute *normalized sentence-level surprisal* for each expression. This decision was motivated by the assumption that sentence-level surprisal better captures the integrative processing effort required in real-time comprehension of non-compositional expressions. In particular, the RU text output from the Speech-to-Text stage is fed into two autoregressive models, ruGPT-3-small (125M parameters) and ruGPT-3-large (760M parameters) (Zmitrovich et al., 2024), in order to address RQ2. The ruGPT-3-small and ruGPT-3-large were chosen to represent different model capacities while maintaining domain consistency. Both models were trained on RU text, making them suitable for modeling native RU speakers' processing.

Note that since these language models generate output based solely on left-to-right context, they estimate probabilities for each word in a sequence by conditioning only on prior tokens.The models assign probability scores to each word in the transcribed sequences, and we converted these scores into surprisal values. We normalize surprisal scores for each stimulus sentence by summing up the scores for all tokens of the sentence given their preceding context, and then divide by the number of the tokens in the sentence to get *normalized sentence-level surprisal*.

2.5 Statistical Analysis

We analyzed the binary response data, i.e., correct vs. incorrect (baseline), using generalized linear mixed-effects models (GLMMs) with a binomial logit link by using *glmer* function in the *lmer* package (Bates, 2016) of R (Team et al., 2013). While more graded scoring methods of intelligibility exist, we opted for a binary approach to ensure comparability across task formats and to reduce subjectivity in judgment, especially when it comes to non-compositional expressions.

For both the free translation and MCQ tasks, the fixed effects were: (1) Linguistic Distances: PWLD and Levenshtein distance, (2) GPT-based Surprisal: Extracted from both large and small GPT models, and (3) Experimental Factors: Experiment input, i.e., spoken-only (Experiment 1) vs. spoken+written (Experiment 2) with spoken-only as the baseline, and Language group (East, South, West; East as the baseline), including relevant interaction terms.

All continuous predictors (i.e., linguistic distances and surprisal values) were centred to their mean values to reduce collinearity, of which more detailed explanation can be found in Appendix C. The experimental factors were dummy-coded. Random effects comprised intercepts for participants (`user_id`) and source texts (`source_text_to_be_translated`), with random slopes for Experiment input when justified by the data. Models were optimized using the bobyqa optimizer (maxfun = 200,000) with Laplace approximation. Model fit was assessed using AIC, and predictor significance was evaluated via z-values and corresponding p-values (with degrees of freedom estimated by Satterthwaite's method where applicable).

3 Results and Discussion

3.1 Intelligibility Scores

Figure 4 shows that intelligibility scores varied both by task and input type. In general, free translation scores were lower than those from the MCQ task, as expected given the greater cognitive demands. The additional written input in Experiment 2 improved free translation performance, but affected MCQ responses adversely, suggesting that orthographic cues aid deeper semantic processing but may interfere with rapid recognition-based decisions. Regarding language groups, East Slavic languages (BE and UK) demonstrated the highest intelligibility scores, South Slavic (BG) – intermediate scores, and West Slavic languages (PL and CS) – the lowest. This gradient reflects typological proximity of Slavic languages, and is consistent with prior findings on Slavic intercomprehension (Gooskens and van Heuven, 2021; Stenger and Avgustinova, 2021).

3.2 RQ1: Predictive Power of Linguistic Distances and GPT-based Surprisal

Our analysis reveals distinct patterns in how linguistic distances and GPT-based surprisal predict cross-language intelligibility across tasks. As the results of the free translation task show in Table 1a, both metrics were significant predictors: The higher the Levenshtein distances (Est. = -0.41, p = 0.004) and higher the surprisal values (Est. = -0.54, p = 0.008), the lower the log odds of having a correct response (reflecting lower intelligibility). The MCQ task shows a different pattern (Table 1b). While linguistic distances emerges as the primary predictor (PWLD: Est. = -0.43, p < .001; Levenshtein: Est. = -0.33, p < .001), GPT-based surprisal had a weaker effect on performance (Est. = -0.31, p = .021). The results with small GPT models in Table 2a and 2b demonstrate the same tendency.

Experiment input and language group also contributed to explaining the intelligibility. As evident in Table 1a, written input improved free translation performance (Est. = 1.23, p < .001) but showed no significant contribution for MCQ responses (Table 1b: Est. = -0.29, p = .131). Additionally, compared to East Slavic languages (the baseline level), both South Slavic (Est. = -3.07, p < .001) and West Slavic languages (Est. = -2.48, p < .001) showed significantly lower intelligibility in the free translation task. Whereas in the MCQ, only West Slavic languages (Est. = -1.05, p < .001) stood out.

3.3 RQ2: Comparison of GPT Model Sizes

The results in Tables 1a and 2a for free translation, and in Tables 1b and 2b for MCQ, revealed similar performance patterns of GPT-based surprisal across model sizes in both free translation (Large: Est. = -0.54, p = 0.008; Small: Est. = -0.56, p = 0.006) and MCQ tasks (Large: Est. = -0.31, p = .021; Small: Est. = -0.31, p = .021). These results contradict previous findings claiming that larger model capacity lead to a worse prediction of human performance (Oh and Schuler, 2023a,b). However, the previous studies considered reading times and monolingual experiments. Our results indicate that the role of surprisal in cross-language intelligibility should be treated differently than that in monolingual experiments. On the other hand, the difference in the results could also rise from the fact that we used RU ASR models to generate the input for language model surprisal, which could add more noise to the data.

4 Conclusion

This study investigated (1) how linguistic distances and surprisal derived from GPT models predict cross-language intelligibility of non-compositional expressions (2) and whether GPT-based model size matters for prediction power. The study used free translation and multiple-choice question tasks in speech-only or speech+written setups. Our results showed that linguistic distances (orthographic and phonological) emerged as the strongest predictors of intelligibility in both tasks. GPT-based surprisal was a significant predictor only in the free translation task, highlighting that such a task is more sensitive to contextual predictability. Additionally, minimal differences in surprisal's performance between large and small variants suggest that a larger GPT model can predict cross-language comprehension outcomes as effectively as a small one.

These findings underscore the complex interplay between typological proximity, orthographic and phonological similarities, and task demands in shaping cross-language intelligibility. The differential impact of written input across tasks further highlights that while orthography can support deeper semantic processing, it may confound recognition-based tasks. Future research should explore other language families and consider other language models for predicting cross-language intelligibility.

Ethical statement

Before taking part in the experiments, all the participants gave their consent that their anonymized responses would be used for research purposes. Participants were compensated for their work in standard rate suggested by Prolific.

Limitations

While our study provides valuable insights into the cognitive mechanisms underlying cross-language intelligibility, it is based on native Russian speakers and specific ASR and language models for Russian. Further work is needed to generalize these findings to other language groups and other ASR and language models. Additionally, the gender imbalance among recorded speakers may have influenced results and should be addressed in future studies. Also, our binary correctness metric may not be able to capture partial comprehension where participants may derive partial meanings through analogy or contextual inference. Incorporating alternative metrics, such as participant confidence ratings or graded correctness scales, could provide a more nuanced view of intelligibility in such cases.

References

Badr M. 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*, volume 11755 of Lecture Notes in Computer Science. Springer, Cham., pages 15–30.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing, Second Edition*, pages 267–292. Chapman and Hall/CRC.

Douglas Bates. 2016. lme4: Linear mixed-effects models using eigen and s4. *R package version*, 1:1.

Catherine Best. 1995. *A direct realist view of crosslanugage speech perception*, pages 171–204.

Ivan Bondarenko. 2022. Xlsr wav2vec2 russian with 2-gram language model by ivan bondarenko. `https://huggingface.co/bond005/wav2vec2-large-ru-golos-with-lm`.

Anne Cutler and Charles Clifton. 1999. Comprehending spoken language: A blueprint of the listener.

J. C. Cutting and K. Bock. 1997. That's the way the cookie bounces: Syntactic and semantic components of experimentally elicited idiom blends. *Memory & Cognition*, 25(1):57–71.

Vera Demberg, Asad Sayeed, Philip Gorinski, and Nikolaos Engonopoulos. 2012. Syntactic surprisal affects spoken word duration in conversational contexts. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 356–367, Jeju Island, Korea. Association for Computational Linguistics.

Pierre Doyé. 2005. *Intercomprehension. Guide for the Development of Language Education Policies in Europe: From Linguistic Diversity to Plurilingual Education*. Reference study. Council of Europe, Strasbourg.

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.

Angela Friederici. 2002. Friederici, a. d. towards a neural basis of auditory sentence processing. trends cogn. sci. 6, 78-84. *Trends in cognitive sciences*, 6:78–84.

Charlotte Gooskens. 2024. *Mutual intelligibility between closely related languages*. Walter de Gruyter GmbH Co KG.

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

Charlotte Gooskens and Vincent van Heuven. 2021. *Mutual Intelligibility*, pages 51–95. Studies in Natural Language Processing. Cambridge University Press.

Leonid Iomdin. 2015. Microsyntactic constructions formed by the Russian word raz. *SLAVIA c ˇasopis pro slovanskou filologii*, 84(3).

Leonid Iomdin. 2016. Microsyntactic phenomena as a computational linguistics issue. In *Proceedings of the Workshop on Grammar and Lexicon: interactions and interfaces (GramLex)*, pages 8–17, Osaka, Japan. The COLING 2016 Organizing Committee.

Ray Jackendoff. 2002. *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford University Press UK.

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

Nikolay Karpov, Alexander Denisenko, and Fedor Minkin. 2021. Golos: Russian dataset for speech research. *arXiv preprint*.

Jacek Kudera, Irina Stenger, Philip Georgis, Bernd Möbius, Tania Avgustinova, and Dietrich Klakow. 2023. Cross-linguistic intelligibility of idiomatic phrases in polish-russian translation tasks. In Jean-Pierre Colson, editor, *Phraseology, Constructions and Translation: Corpus-based, Computational and Cultural Aspects*, pages 237–249. Presses Universitaires de Louvain.

V. I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710.

William Marslen-Wilson. 1987. Functional parallelism in spoken word-recognition. *Cognition*, 25:71–102.

Maryam Sadat Mirzaei, Kourosh Meshgi, and Tatsuya Kawahara. 2016. Automatic speech recognition errors as a predictor of L2 listening difficulties. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 192–201, Osaka, Japan. The COLING 2016 Organizing Committee.

Steven Moran and Daniel McCloy, editors. 2019. *PHOIBLE 2.0*. Max Planck Institute for the Science of Human History, Jena.

Robert Möller and Ludger Zeevaert. 2015. Investigating word recognition in intercomprehension: Methods and findings. *Linguistics*, 53.

Byung-Doh Oh and William Schuler. 2023a. Transformer-based language model surprisal predicts human reading times best with about two billion training tokens. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1915–1921, Singapore. Association for Computational Linguistics.

Byung-Doh Oh and William Schuler. 2023b. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, 11:336–350.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

Irina Stenger. 2019. *Doctoral Dissertation: Zur Rolle der Orthographie in der slavischen Interkomprehension mit besonderem Fokus auf die kyrillische Schrift*. Ph.D. thesis, Saarbrücken: universaar.

Irina Stenger and Tania Avgustinova. 2021. On Slavic cognate recognition in context. In *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference 'Dialogue'*, volume 20, pages 660–668, Moscow, Russia.

Irina Stenger, Tania Avgustinova, and Roland Marti. 2017a. Levenshtein distance and word adaptation surprisal as methods of measuring mutual intelligibility in reading comprehension of slavic languages. In *Computational Linguistics and Intellectual Technologies: International Conference "Dialogue 2017"*, pages 304–317.

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

Roland Sussex and Paul Cubberley. 2006. *The Slavic Languages*. Cambridge University Press, Cambridge.

R Core Team et al. 2013. R: A language and environment for statistical computing. *Foundation for Statistical Computing, Vienna, Austria*.

Keyon Vafa, Ashesh Rambachan, and Sendhil Mullainathan. 2024. Do large language models perform the way people expect? measuring the human generalization function. *Preprint*, arXiv:2406.01382.

Jan Vanhove and Raphael Berthele. 2015. Item-related determinants of cognate guessing in multilinguals. *Crosslinguistic Influence and Crosslinguistic Interaction in Multilingual Language Learning*, 95:118.

Alison Wray. 2002. *Formulaic Language and the Lexicon*. Cambridge University Press.

Iuliia Zaitova, Irina Stenger, and Tania Avgustinova. 2023. Microsyntactic unit detection using word embedding models: Experiments on slavic languages. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pages 1265–1273. INCOMA Ltd.

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, Torino, Italia. ELRA and ICCL.

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 *Proceedings of Interspeech 2024*, Saarbrücken, Germany.

Jian Zhu, Cong Zhang, and David Jurgens. 2022. Byt5 model for massively multilingual grapheme-to-phoneme conversion.

Dmitry Zmitrovich, Alexander Abramov, Andrey Kalmykov, Maria Tikhonova, Ekaterina Taktasheva, Danil Astafurov, Mark Baushenko, Artem Snegirev, Vitalii Kadulin, Sergey Markov, Tatiana Shavrina, Vladislav Mikhailov, and Alena Fenogenova. 2024. A family of pretrained transformer language models for russian. *Preprint*, arXiv:2309.10931.

A Microsyntactic Units in six Slavic languages used as Experimental Stimuli

Type	**BE**	**UK**	**BG**	**CS**	**PL**	**RU**
Prep	ў канцы	у кінці	в края на	na konec	w końcu	в конце
Eng. trans.	*at the end of*	*at the end of*	*at the end of*	*at the end of*	*at the end of*	*at the end of*
Adv & Pred	не раз	не раз	не веднъж	ne jednou	niejednokrotnie	не раз
Eng. trans.	*not once*	*not once*	*not once*	*not once*	*not once*	*not once*
Parenth	такім чынам	таким чином	такъв начин	tímto způsobem	w taki oto sposób	таким образом
Eng. trans.	*in this way*	*in this way*	*in this way*	*in this way*	*in this way*	*in this way*
Conj	хіба толькі	хіба що	освен да	snad jen	chyba że	разве что
Eng. trans.	*except (only) that*	*except (only) that*	*except (only) that*	*except (only) that*	*except (only) that*	*except (only) that*
Part	усе ж	все же	все пак	asi spíš	więc jednak	все же
Eng. trans.	*nonetheless*	*nonetheless*	*nonetheless*	*nonetheless*	*nonetheless*	*nonetheless*

Note: We use ISO 639-1 codes for the languages: Belarusian (BE), Ukrainian (UK), Bulgarian (BG), Czech (CS), Polish (PL), Russian (RU).

Table 3: Microsyntactic units in six Slavic languages.

B Linguistic distances

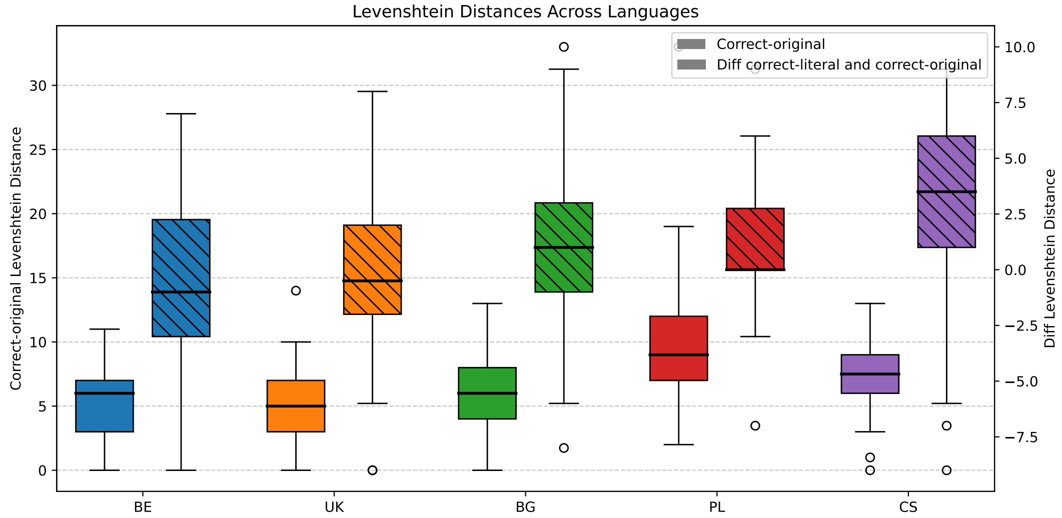


Figure 3: Ranking of orthographic Levenshtein Distance by language

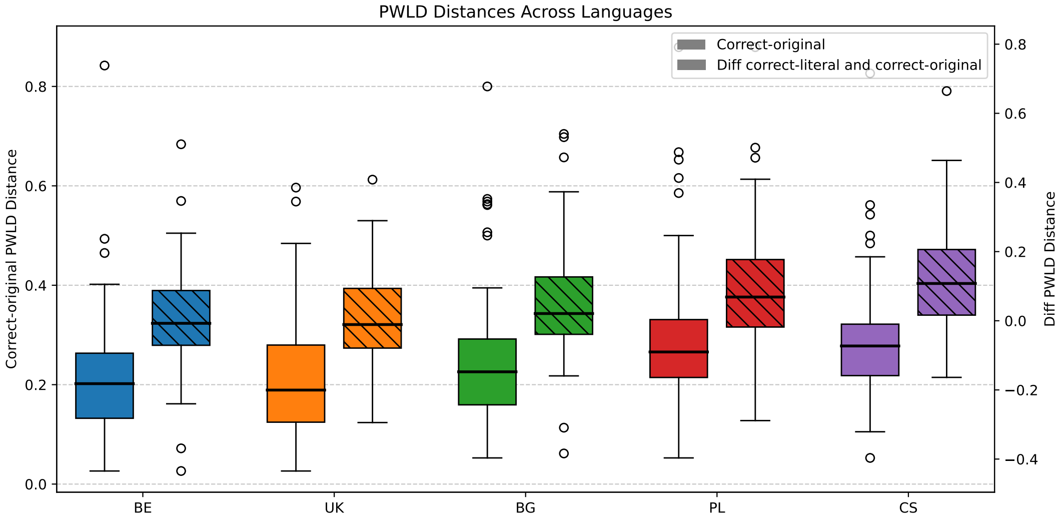


Figure 4: Ranking of phonological Distance by language

C Colinearity concern

In order to avoid the impact of multicollinearity issue on our GLMM models, we first checked the simple Pearson correlations between PWLD and Levenshtein distance used for the models of free translation and MCQ tasks, and they are 0.44 and 0.50, respectively. Their correlations with GPT-based surprisal values are 0.08 (*SD* = 0.037).

However, collinearity is more than correlation between two predictors. What else is important is how the predictor interacts within the full set of variables in the model. Thus, in order to have the full picture of how the predictors contribute to explaining the variance of our dependent variable, we further checked the scaled Generalized Variance Inflation Factor (sGVIF) of all predictors in our GLMM models. VIF in general quantifies how much multicollinearity exists in a regression model, and sGVIF can be applied to categorical variables (e.g., language group) and takes into account predictors' degree of freedom. An sGVIF =1 indicates no multicollinearity, and a value below 2 is generally considered acceptable.

The mean (SD) of sGVIF values for all predictors in the four models reported in Tables 1 and 2 are: 1.379952 (0.2320059) for free translation with GPT large, 1.46668 (0.345058) for MCQ with GPT large, 1.384291 (0.2342397) for free translation with GPT small, and 1.465348 (0.3460249) for MCQ with GPT small. As can be seen, these sGVIF values are closer to 1, indicating that multicollinearity is not a big concern in our GLMM models.

D Performance of Automatic Speech Recognition on Russian dataset

Table 4: ASR performance of wav2vec2-large-ru-golos-with-lm on Russian datasets (WER and CER)[2]

Dataset	WER (%)	CER (%)
Sberdevices Golos (crowd)	6.88	1.64
Common Voice RU	12.12	2.98
Russian Librispeech	15.74	3.57

Be aware that we do not report the performance (e.g., WER) of ASR on foreign speech inputs as it is just not applicable. Thinking about when people listen to an unknown language, they do not know the ground truth of perceiving the speech, and just simply map it to their known language(s). There can be various mappings.

[2] Model: `bond005/wav2vec2-large-ru-golos-with-lm`, available at https://huggingface.co/bond005/wav2vec2-large-ru-golos-with-lm