Cross-Linguistic Processing of Non-Compositional Expressions in Slavic Languages

Iuliia Zaitova, Irina Stenger, Muhammad Umer Butt and Tania Avgustinova
Department of Language Science and Technology, Saarland University, Germany
izaitova@lsv.uni-saarland.de, ira.stenger@mx.uni-saarland.de,
mubu00001@stud.uni-saarland.de, avgustinova@coli.uni-saarland.de

Abstract

This study focuses on evaluating and predicting the intelligibility of non-compositional expressions within the context of five closely related Slavic languages: Belarusian, Bulgarian, Czech, Polish, and Ukrainian, as perceived by native speakers of Russian. Our investigation employs a web-based experiment where native Russian respondents take part in free-response and multiple-choice translation tasks. Based on the previous studies in mutual intelligibility and non-compositionality, we propose two predictive factors for reading comprehension of unknown but closely related languages: 1) linguistic distances, which include orthographic and phonological distances; 2) surprisal scores obtained from monolingual Language Models (LMs). Our primary objective is to explore the relationship of these two factors with the intelligibility scores and response times of our web-based experiment. Our findings reveal that, while intelligibility scores from the experimental tasks exhibit a stronger correlation with phonological distances, LM surprisal scores appear to be better predictors of the time participants invest in completing the translation tasks.

Keywords: non-compositionality, closely related languages, language models, surprisal, linguistic distances

1. Introduction

The principle of compositionality in linguistics states that the meaning of a complex expression is determined by the meanings of its constituent parts (Partee, 2008). However, non-compositional expressions deviate from this principle. Non-compositional expressions are linguistic constructs where the overall meaning cannot be straightforwardly inferred from the meanings of their individual components (Baldwin and Kim, 2010). The meaning of non-compositional expressions often relies on cultural, contextual, or conventional associations, making them an aspect of language that requires specialized analysis beyond the scope of compositional interpretation (Jackendoff, 2002). Examples of non-compositional expressions include idioms (*e.g., English: "to kick the bucket" meaning: to die*), metaphors (*Czech: "Život je cesta", meaning "Life is a journey"*), and microsyntactic units (*Bulgarian:* "не веднъж" *transliterated as "ne vedn"ž"*[1]*, meaning "not once"; Russian:* "в конце" *transliterated as "v konce", meaning "at the end of"*).

While the mechanisms underlying the comprehension and processing of non-compositional expressions within a single language have been investigated extensively (Cacciari and Tabossi, 1988; Conklin and Schmitt, 2008; Titone et al., 2019), the dynamics of cognitive processing of written non-compositional expressions across languages – especially within closely related language groups like Slavic languages, remain a subject for exploration.

In light of this, the current study addresses the following research questions:

- **RQ1:** How well can native Russian speakers spontaneously understand non-compositional expressions from unfamiliar Slavic languages, namely Belarusian (BE), Bulgarian (BG), Czech (CS), Polish (PL), and Ukrainian (UK) in written context?
- **RQ2:** To what extent do algorithmic factors, namely surprisal from Language Models and linguistic distances, predict the cross-lingual intelligibility of non-compositional expressions?

The paper is structured as follows: we start by providing information on previous research in non-compositionality and language intercomprehension (Section 2) and stating our hypotheses (Section 3); then we describe our web-based experiment (Section 4) and algorithmic predictors (Section 5) to finally present (Section 6) and discuss the results in Section 7. The code for this paper is available at the following link: https://github.com/IuliiaZaitova/non-compositional-expressions-slavic.

2. Related Work

Spontaneous comprehension of unknown but related languages is detectable by means of differently designed experiments, e.g., cloze tests, multiple-choice questions, or translation tasks. For example, testing the Cyrillic script intelligibility by Russian native speakers in a context-free

[1] Here and further, we used ISO 9:1995 transliteration from Cyrillic.

The Workshop on Cognitive Aspects of the Lexicon, pages 86–97
20 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

word translation task, Stenger, 2019 reveals that Ukrainian and Belarusian are more understandable by the participants than Bulgarian, Macedonian and Serbian. The observed human performance in contextualized cross-lingual cognate recognition, as reported by Stenger and Avgustinova, 2021, also validates the intuition that Russian readers spontaneously understand stimuli in Ukrainian and Belarusian better than in Bulgarian.

When it comes to factors explaining the intercomprehension of related languages, researchers generally assume that the more similarities two languages share, the higher their degree of mutual intelligibility is (Gooskens and Swarte, 2017). As shown by Stenger and Avgustinova, 2021 linguistic distances are highly significant for correct in-context recognition of cognates from closely related languages. When looking at the intelligibility of Polish words to Czech readers, Jágrová et al., 2021 also confirms the role of linguistic similarity in predicting cross-lingual comprehension and finds that context-aware Language Models (LMs) perform better than 3-gram Language Models when predicting intercomprehension.

The exploration of different kinds of non-compositional expressions is fortified by a body of research consistently showing that these linguistic units exhibit increased processing facilitation (Cacciari and Tabossi, 1988; Conklin and Schmitt, 2008; Vespignani et al., 2009; Siyanova-Chanturia et al., 2011; Titone et al., 2019).

A relevant work by Kudera et al., 2023 investigates the auditory comprehension of idiomatic phrases, which is also a type of non-compositional expressions, in two closely related Slavic languages, Polish and Russian. The study builds on information-theoretic measures of word adaptation surprisal, coupled with syntactic distances between non-compositional expressions, to predict lay translators' preferences. Kudera et al.'s work serves as a foundational reference for our work; however, our approach diverges in several aspects: 1) we employ a reading comprehension scenario; 2) we test the comprehension in context; 3) we use five different target languages and compare the comprehension of non-compositional expressions across them.

A noteworthy study of the correlation between non-compositional expression intelligibility and LM performance is presented by Rambelli et al., 2023. Their work particularly focuses on idiomatic and high-frequency compositional expressions. The study indicates that humans process idioms with non-compositional meaning and high-frequency compositional phrases much faster than low-frequency compositional phrases. In parallel, LMs assign to idioms significantly lower surprisal values. In the context of our work, their findings underscore the potential of LM surprisal as a robust metric for predicting the processing of non-compositional expressions.

3. Hypotheses

RQ1: Our intention is to critically examine the alignment of our intelligibility tests with genealogic taxonomies established by comparative linguistics (Sussex and Cubberley, 2006), similarly to what is demonstrated in Charlotte Gooskens and Voigt, 2018. We hypothesize that native Russian speakers exhibit a higher comprehension level when exposed to non-compositional expressions in languages of the same East Slavic group (Belarusian and Ukrainian), and a lower comprehension level for languages in different groups (West Slavic and South Slavic). Moreover, we anticipate longer response times for languages more distant from Russian.

RQ2: Drawing upon previous studies in mutual intelligibility and non-compositionality, mainly Stenger and Avgustinova, 2021, Jágrová et al., 2021, Kudera et al., 2023, and Rambelli et al., 2023, we propose a dual-factor framework for predicting percentage of correct responses (intelligibility scores) and response times within our experimental context.

Factor 1: Linguistic Distances – we anticipate that more distant linguistic units will be more challenging for participants to recognize. Taking into account both orthographic and phonological distances, we predict a negative correlation between both types of linguistic distances and intelligibility scores.

Factor 2: Surprisal Scores from Language Models (LMs) – additionally, we incorporate surprisal scores from monolingual LMs trained on Russian. We analyze LM surprisal for 1) non-compositional Russian expressions in Russian context; 2) literal Russian expressions in Russian context; 3) non-compositional foreign expressions in foreign language context. We hypothesize a positive correlation between surprisal scores and user task completion time, with lower surprisal indicating processing facilitation. Additionally, we expect that surprisal scores of 1) and 2) correlate with results of multiple-choice question task since the low surprisal of the option in a particular context, which might be partially intelligible to the reader, can trigger the choice of that option (either literal or non-compositional). We also predict that 3) correlates with the outcomes of both tasks.

4. Human Translation of Unfamiliar Non-Compositional Expressions

In order to measure the intelligibility of non-compositional expressions we designed a two-task experiment that includes a free translation task and a multiple-choice task, each serving to probe different aspects of the participants' comprehension skills.

4.1. Stimuli

In our study, we utilize an existing dataset, initially crafted for the analysis of microsyntactic units, which are defined as non-compositional expressions with inherent syntactic idiomaticity. Such units include all the syntactic units that have very specific syntactic properties and do not fit into the standard syntax (Iomdin, 2015). The dataset consists of 227 Russian microsyntactic units, each accompanied by translational correlates and two parallel bilingual context sentences across six Slavic languages, as it is thoroughly described in Zaitova et al., 2023. The dataset was created using the Russian National Corpus (RNC) and its parallel sub-corpora as the primary linguistic resource (https://ruscorpora.ru). The microsyntactic dictionary provided by the RNC served as the pivot database. It includes various syntactic categories such as prepositions, adverbials, conjunctions, etc. The researchers selected the most frequent microsyntactic units for further analysis, totaling 227 units in Russian. Translational correlates were extracted from the RNC's parallel sub-corpora and the Czech National Corpus (Machálek, 2020), resulting in six parallel sets for each Slavic language under analysis. The dataset is open-sourced and available for use (https://huggingface.co/datasets/izaitova/slavic_fixed_expressions). While it was developed with a focus on microsyntactic units, in the current study we categorize these units as non-compositional expressions since, in line with the definition presented in Section 1, their meaning cannot be readily derived from their individual components. For each Slavic language in the dataset, we have selected a total of 60 expressions. The average sentence length in tokens per sentence is as follows: BE: 15.3, BG: 14.9, CS: 11.3, PL: 13.6, and UK: 14.8.

4.2. Word-by-Word Translation Options for Multiple-Choice Questions

A multiple-choice question format is employed in the experiment design as one the methods to assess participants' comprehension of presented non-compositional expressions. For each stimulus, participants are provided with a choice between two options: a correct translation and an literal translation counterpart, with the latter being crafted as a plausible yet inaccurate compositional translation of the respective expression, mirroring the stimulus in form. The goal is to challenge participants to discern between non-compositional (correct) and literal (incorrect) options. In the preparation of the assumed incorrect translations, we have utilized word-by-word translations provided by the online bilingual Glosbe Dictionary (https://glosbe.com). Additionally, for the identification of cognates, we use the etymological online dictionary of the Russian language by Max Vasmer (https://lexicography.online/etymology/vasmer/). The inclusion of literal translations as incorrect options aims at providing insights into participants' ability to move beyond surface-level comprehension and engage with the deeper (non-compositional) meanings of the investigated expressions.

4.3. Experimental Procedure

Cross-lingual intelligibility of non-compositional expressions to native Russian speakers has been assessed using a custom-built application available online at https://intercomprehension.coli.uni-saarland.de, as described by Stenger et al., 2020. The subjects received instructions in Russian about the tasks and procedures to follow. After familiarizing themselves with the task, participants registered on the website hosting our web application and completed a questionnaire about their background and language skills. During the experiment, participants saw five sets of 12 contextualized non-compositional expressions from one of the stimulus languages – Belarussian (BE), Bulgarian (BG), Czech (CS), Polish (PL), Ukrainian (UK). Each time, a set of 12 stimuli was randomly selected from all available sets per language, totalling 60 sentences per participant. Repetitions were avoided by ensuring that each stimuli set is presented to each participant only once. Stimuli sentences were presented one by one, and participants were first asked to type a free translation of the highlighted non-compositional expression (see Figure 1). Next, participants were presented with the multiple-choice question task (MCQ) task (see Figure 2) for the same stimulus. They were provided with two possible solutions for the translation of a foreign non-compositional expression into RU: (i) non-compositional translation; (ii) an alternative word-by-word translation, which is an inaccurate translation of the expression.

This combination of tasks was designed to be concatenated, with the addition of time limits to discourage lengthy reflection. While there may be some priming effect within the same stimulus, the difference between the two tasks (outlined in Sec-

tion 4.5) does not appear to be primarily attributable to priming. Participants are presented more information in the multiple-choice options, leading to an expected increase in accuracy compared to the free translation task.

The time allocated for translating the highlighted non-compositional expression is based on a formula of 10 seconds per token plus an additional 3 seconds per sentence. For the second task, we add 3 more seconds plus 10 seconds per token in both translation options. Such timing is based on the experience with contextualized cognate guessing tasks and aligns with related studies, e.g., Stenger and Avgustinova, 2021. The timing of response for each stimulus starts when the question is shown to the user, and ends when the user proceeds to the next stimulus, either by providing a response or pressing the Skip button. For free translation task, we considered alternative semantically equivalent translations and typographical errors as correct responses. Accuracy in both tasks is defined as the percentage of correct responses out of total responses.

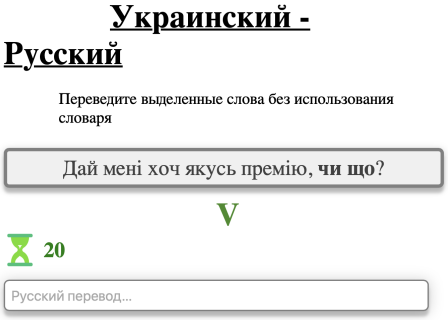


Figure 1: Experimental screen of the free translation task as seen by Russian respondents. The instruction reads: 'Ukrainian - Russian. Translate the highlighted words without using a dictionary'. The Ukrainian sentence is: 'Give me at least some kind of bonus *or something*?' The translation is to be written in the white box

4.4. Participants

In total, 135 native Russian participants took part in the study, aged between 20 and 78 years old (i.e. average age 35), comprising 92 females, 41 males, and 2 individuals who identified as another gender. The subjects were untrained in translation and were recruited for participation in the experiment through Prolific (https://prolific.com), an online platform specializing in participant recruitment for research purposes. To reveal the inherent intercomprehension, we excluded 12 participants because they had some knowledge of the stimulus language. Since the Prolific platform is in English,

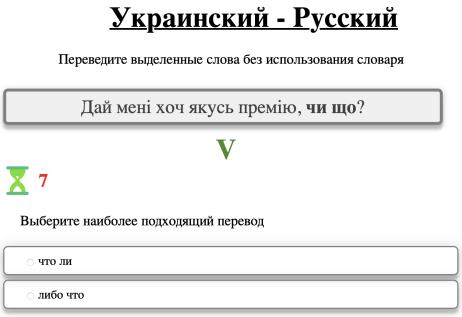


Figure 2: Experimental screen of the MCQ task as seen by Russian respondents. The instruction reads: 'Ukrainian - Russian. Translate the highlighted words without using a dictionary'. The Ukrainian sentence is: 'Give me at least some kind of bonus *or something*?' Below is the prompt: 'Choose the most suitable translation.'

we expect that the speakers are familiar with the Latin script used by CS and PL languages. The number of subjects for each stimulus ranges from 17 to 55 with an average of 24 participants per stimulus. After each block, each participant may continue the experiment by completing the task for the remaining stimulus sentences offered in a random order.

4.5. Results

In Figure 3, the left plot illustrates the accuracy for both multiple-choice questions and free translation tasks, represented as the percentage of correct responses out of total responses. The right plot displays the response time for both tasks, organized by stimulus language. In both tasks, the highest accuracy is observed in translations from BE and UK. Since BE and UK belong to the same branch of Slavic languages as RU, such results are in line with the previous studies on Slavic language intercomprehension (Stenger and Avgustinova, 2021). Translations from BG also exhibited a relatively high accuracy. However, the accuracy dropped significantly for CS and PL. Generally, the participants' performance is much lower in the free translation task, which is expected given that the task requires more open-ended and expressive language production.

As for time measurements, we can observe the opposite tendency: participants generally required more time when translating from BG, CS, and PL compared to BE and UK. This difference in time may reflect the additional effort and processing demands involved in comprehending and generating translations for languages that are less closely related or have greater linguistic differences.

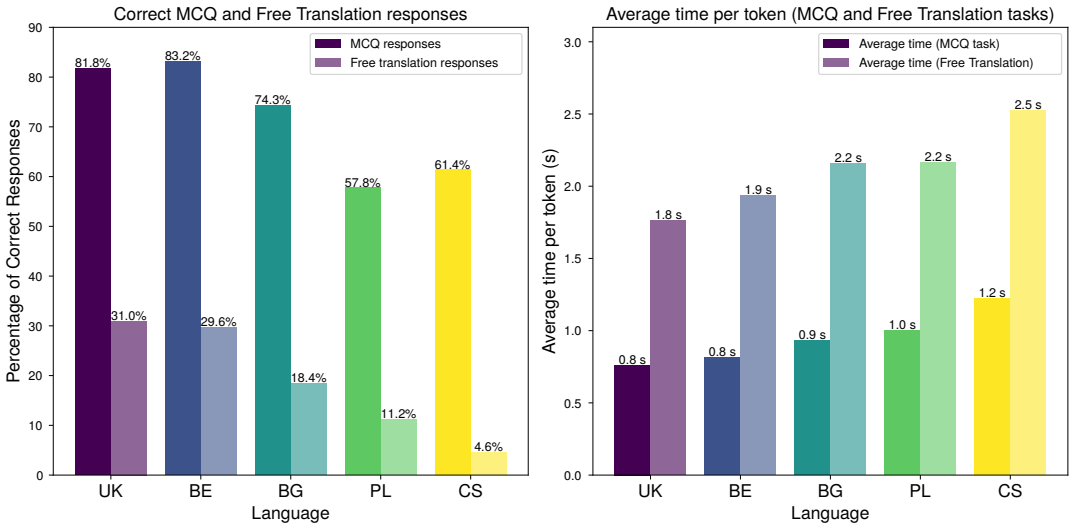


Figure 3: Left: Percentage of correct responses out of total responses (accuracy). Right: Average time per token in seconds. Both plotted by stimulus language

5. Predictors of Non-Compositional Expression Intelligibility

In this section, we describe the factors that we identified as predictors of our experiment metrics: linguistic distances and surprisal from LMs. In the section, we describe the two types of linguistic distances that we utilized and provide their comparative scores that further demonstrate their potential. We aim to investigate to what extent they can serve as a reliable proxy for cross-lingual intelligibility of non-compositional expressions in closely related languages.

5.1. Linguistic Distance

As outlined in Section 2, previous studies on intercomprehension provide strong support for using orthographic and phonological distances as a predictor of cross-lingual intelligibility (Vanhove and Berthele, 2015; Möller and Zeevaert, 2015; Gooskens and Swarte, 2017). However, measuring the distance between modern Slavic languages could be challenging due to the use of two writing scripts – Latin and Cyrillic. To accommodate for this, we employed two measures of phonological and orthographic distances that are adapted to deal with different scripts and were used before in Slavic intercomprehension studies specifically (Zaitova et al., 2023; Stenger et al., 2022; Mosbach et al., 2019).

5.1.1. Orthographic Distance

To measure the orthographic distance, we used **normalized Word Adaptation Surprisal (nWAS)**, which quantifies the degree of unexpectedness of a word form given a possibly related word form and set of transformation probabilities (Stenger et al., 2017). To use nWAS, orthographic character alignment costs are necessary. Based on these costs, words are aligned with the Needleman-Wunsch algorithm (Needleman and Wunsch, 1970). For our analysis, we adapted the code and the orthographic alignment costs previously computed for Slavic languages in Stenger et al., 2022. Here, identical characters have zero alignment cost, while characters differing only in diacritics (e.g., ⟨á⟩ and ⟨a⟩) were assigned a cost of 0.5. Unrelated vowel-vowel or consonant-consonant character pairs (e.g., ⟨a⟩ and ⟨i⟩, or ⟨k⟩ and ⟨v⟩) were assigned alignment costs of 1. Cyrillic hard and soft signs (⟨ь, ъ, ’⟩) were also assigned alignment costs of 1 to each other. For all other character pairs (e.g., consonant-vowel pairs), a cost of 4.5 was assigned. Cyrillic words were aligned by converting Cyrillic characters to ISO 9 Latin characters and then applying the alignment costs specified above.

5.1.2. Phonological Distance

Phonologically Weighted Levenshtein Distance (PWLD) is a measure of phonological similarity between different phonemic sequences or word forms (Fontan et al., 2016). The PWLD metric is an extension of the string-based Levenshtein distance that also takes into account the cost of each phoneme substitution based on phoneme features. These features are based on the PHOIBLE (Moran and McCloy, 2019) feature set. The substitution cost between phonemes is computed as the Hamming distance between their feature vector representations. We suppose that PWLD is more suitable for cross-lingual analysis than Levenshtein Distance

since it is capable of catching less apparent phonological similarities. For example in the pair of Czech and Bulgarian 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. We use the same adaption of the original PWLD proposed in Abdullah et al. (2021) to make it suitable for our analysis. To obtain the phonetic transcription of all stimuli and MCQ task options, we used CharsiuG2P, which is a transformer based tool for grapheme-to-phoneme conversion (Zhu et al., 2022).

It might seem counterintuitive that we consider phonological distance for written data. After all, native RU participants are not expected to know the correct pronunciation of the stimuli since they never learnt stimulus languages before. However, they can try to read stimulus aloud, i.e. try to understand unfamiliar languages using their inner speech (Alderson-Day and Fernyhough, 2015). Additionally, previous research has shown that a pronunciation-based distance is a better predictor of intelligibility than traditionally calculated orthographic distance (Jagrova, 2022).

5.1.3. Linguistic Distance Results

Table 1 presents the nWAS and PWLD scores, indicating the average distance from the correct non-compositional expression (NC) in RU to the source expression in the foreign language (L2). Additionally, it shows the distances from the inaccurate word-by-word translation (LIT) to foreign language (L2).

Language	Type	nWAS	PWLD
BG	LIT–L2	3.175	0.204
	NC–L2	**3.221**	**0.253**
BE	LIT–L2	3.236	0.213
	NC–L2	**3.249**	**0.220**
CS	LIT–L2	3.323	0.175
	NC–L2	**3.382**	**0.291**
PL	LIT–L2	3.332	0.208
	NC–L2	**3.389**	**0.298**
UK	LIT–L2	3.257	0.198
	NC–L2	**3.298**	**0.210**

Table 1: nWAS and PWLD scores

5.2. Surprisal from Language Models

Surprisal is a quantifiable measure of unpredictability, grounded in information theory (Crocker et al., 2016). Specifically, surprisal quantifies the negative log-likelihood of encountering a particular unit given its preceding context. The surprisal of a unit increases with decreasing probability, reflecting a higher degree of unexpectedness in a given linguistic context.

Surprisal from Language Models (LMs) serves as a proxy for the difficulty of cognitive processing of (foreign) non-compositional expressions in context. For sequential models like ruGPT3Large and ruGPT3Small, the probability of the expression given context is based solely on the left side, simulating reading from left to right. In contrast, for masked models like ruBERTa-large and ruBERTa-small, it considers both the left and right sides, simulating the utilization of the entire sentential context by the reader.

For example, let's take a sentence from the dataset that we used "_ _, что трассу полета можно менять только в интересах безопасности и защиты здоровья.." (transliteration: *"__, čto trassu polëta možno menât' tol'ko v interesah bezopasnosti i zaŝity zdorov'â..."*, translation: *"__ that the flight path can only be changed for safety and health protection."*) If the missing part is 'можно сказать' (transliteration: *"možno skazat'"*, translation: *"one cay say"*) and the surrounding context makes it highly expected, then the surprisal of the expression 'можно сказать' in this sentence would be low if one considers both left and right context (like masked models like ruBERTa-large and ruBERTa-small). If we consider only the nonexistent context left to the blank space, the model's surprisal would be higher as its uncertainty about the correct sequence of tokens increases.

The LM surprisal scores were obtained using the Python library minicons (Misra, 2022) for three scenarios:

- Surprisal of RU non-compositional expressions in RU context.
- Surprisal of RU literal expressions in RU context.
- Surprisal of foreign non-compositional expression in foreign context.

5.2.1. Language Models

We employ both large and small monolingual Russian LMs to compute surprisal values, using autoregressive models (ruGPT3Large and ruGPT3Small) and bidirectional models (ruBERTa-large and ruBERTa-small).

The LMs utilized in our experiments were developed by the SberDevices team[2] and are detailed as follows:

1. **ruBERTa-large (ruBL)** is an adaptation of the Roberta model (Liu et al., 2019), a masked model that was pre-trained on a substantial 250GB corpus of Russian text.

[2] https://sberdevices.ru

2. **ruGPT3Large (ruGPT3L)** is a large-scale sequential model based on the GPT-2 architecture (Radford et al., 2019).

3. **ruBERTa-small (ruBS)** is a smaller variant of the ruBERTa-large. While it maintains the robustness of its larger counterpart, ruBERTa-small offers a computationally less intensive alternative.

4. **ruGPT3Small (ruGPT3S)** is a scaled-down version of the ruGPT3Large model. The training process was designed to be more computationally efficient while pertaining the generation of linguistically rich and coherent text.

By employing models that utilize both sequential and masked prediction mechanisms, our experiments were designed to provide a full comparison and capture various aspects of language comprehension.

5.3. Surprisal Scores

Table 2 gives an overview of average surprisal scores of the RU non-compositional expressions in RU context (NC), literal RU expressions in RU context (LIT), and foreign non-compositional expression in foreign context (L2). In the last column, we can see the statistical significance of the difference between LIT and NC computed using the Wilcoxon signed-rank test. Additionally, Appendix A presents the boxplots for surprisal values from all stimuli. All the scores were derived from the models described above. We can see that the model ruBS does not detect any statistically significant difference between LIT and NC expressions. For that reason, we exclude this model from our predictors of intelligibility and response times.

	Model	NC	LIT	L2	LIT–NC
BG	ruGPT3S	3.916	7.597	9.333	***
	ruBS	14.549	14.540	14.918	NS
	ruGPT3L	3.646	7.662	9.540	***
	ruBL	1.013	7.496	10.511	***
BE	ruGPT3S	3.524	6.821	9.069	***
	ruBS	0.965	13.688	14.667	NS
	ruGPT3L	3.331	6.754	7.719	***
	ruBL	0.925	6.143	2.795	***
CS	ruGPT3S	3.758	7.258	14.388	***
	ruBS	14.660	15.305	23.778	NS
	ruGPT3L	3.695	7.329	13.743	***
	ruBL	1.140	7.004	10.783	***
PL	ruGPT3S	3.679	7.508	14.183	***
	ruBS	14.749	14.681	26.592	NS
	ruGPT3L	3.475	7.459	13.291	***
	ruBL	1.037	6.380	9.015	***
UK	ruGPT3S	3.570	7.270	8.599	***
	ruBS	14.411	14.424	14.886	NS
	ruGPT3L	3.394	7.284	7.412	***
	ruBL	0.850	6.510	1.946	***

*=p< .05, **=p< .01, ***=p< .001, NS=Not Significant

Table 2: LM surprisal + Wilcoxon signed-rank Test

6. Results

6.1. Correlation Results

We have computed the Pearson correlation of the percentage of correct responses and average response time in both tasks with orthographic and phonological distances, as well as with surprisal scores listed in Table 2. In Appendix B, you can find the tables with results for all Pearson correlations, along with corresponding p-values. For accuracy in free translation task, the strongest correlation is observed with phonological distance (PWLD) between Russian non-compositional expression and foreign non-compositional expression (BE: -0.405**, BG: -0.471***, CS: -0.361**, PL: -0.428***, UK: -0.606***).

For accuracy in MCQ task, there is also a strong correlation for PWLD for all languages except BE (BE: -0.229 NS, BG: -0.417**, CS: -0.283*, PL:, -0.385**, UK: -0.502***)[3]. Additionally, for BE and UK, we can observe a strong significant positive correlation of MCQ translation accuracy and PWLD between Russian literal expressions and foreign non-compositional expressions (BE: 0.429***, UK: 0.307*).

Figure 4 presents the correlation of free translation accuracy and PWLD between Russian non-compositional expressions and foreign non-compositional expressions for UK on the left, and the correlation of MCQ translation accuracy and PWLD between Russian literal expressions and foreign non-compositional expressions for BE on the right.

Average time measurements for both tasks have a stronger correlation with surprisal from LMs for foreign expression in foreign context in most languages, especially with that from the model ruBERTa-large (ruBL). For free translation time: BE: 0.443***, BG: 0.135 NS, CS: 0.547***, PL: 0.304*, UK: 0.217 NS. For MCQ time: BE: 0.457***, BG: 0.102 NS, CS: 0.452***, PL: 0.308*, UK: 0.215.

Figure 5 presents the correlation of free translation time and ruBL surprisal for foreign expression in foreign context for CS on the left, and the correlation of MCQ time and ruBL surprisal for foreign expression in foreign context for PL on the right.

It is worth noting that no statistically significant correlation was detected for the time measurements in the Ukrainian language.

[3]here and further: NS: Not Significant, *: $p < .05$, **: $p < .01$, ***: $p < .001$.

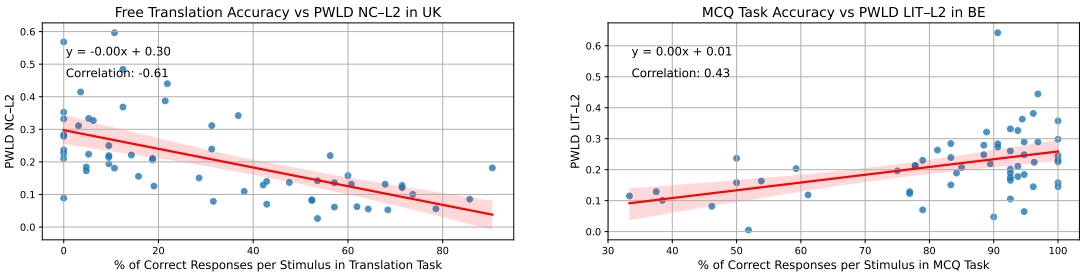


Figure 4: Relation of accuracy with phonological distances

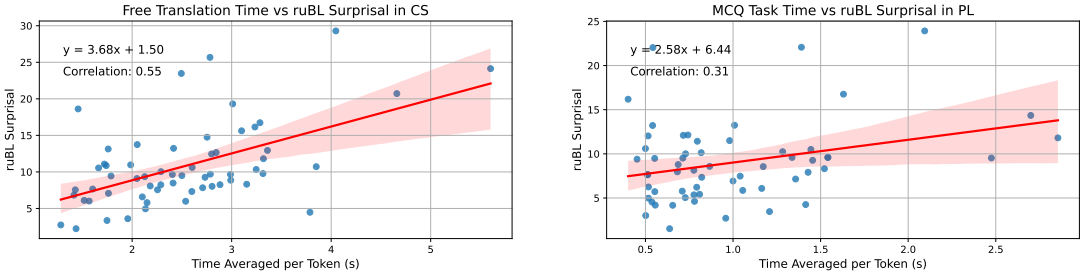


Figure 5: Relation of response time with LM surprisal

6.2. Multiple Regression Results

By adding all variables for surprisal and linguistic distances into a multiple linear regression model for predicting intelligibility across all stimuli from all source languages, we sought to identify the best-fitting models to predict intelligibility scores and average response times in our dataset. To achieve this, we performed a series of regression analyses using the Ordinary Least Squares (OLS) method.

We began by considering all potential predictor variables and identified the models that demonstrated the best fit for our data using stepwise regression. The summary of these results is presented in Table 3. Though overall regression scores are low when comparing the results for all language sets jointly, certain patterns could be observed across variables. For free translation (FT) task accuracy, the phonological distance between the correct RU translation and foreign stimulus (PWLD NC–L2), as well as surprisal scores for foreign stimulus in foreign context from models ruBL and ruGPT3L turned out to be the most significant predictors. For MCQ Task accuracy, phonological distances, namely PWLD NC–L2 and PWLD LIT–L2 (distances between the incorrect/literal translation option and foreign stimulus), again emerged as most impactful, followed by LIT surprisal by ruGPT3S and ruBL, and L2 surprisal by ruBL.

When considering response times, the best model for FT time includes ruBL L2, nWAS NC–L2, ruGPT3S L2, and ruBL LIT. Conversely, the MCQ Time model indicates that ruBL L2 and ruBL LIT are the most significant predictors, while nWAS

Dep. Variable	R^2	Adj. R^2	F	Variable	Coef
FT Accuracy	0.349	0.342	52.03	PWLD NC-L2	-72.4037
				ruBL L2	-0.8590
				ruGPT3L L2	-0.5645
MCQ Accuracy	0.310	0.298	25.94	PWLD NC-L2	-69.4396
				PWLD LIT-L2	50.6198
				ruGPT3S LIT	1.2018
				ruBL L2	-0.8810
FT Time	0.244	0.237	31.36	ruBL L2	0.0461
				nWAS NC-L2	0.2156
				ruGPT3S L2	0.0186
				ruBL LIT	0.9105
MCQ Time	0.182	0.177	32.53	ruBL L2	0.0335
				ruBL LIT	0.0176

Table 3: Multiple regression results

orthographic distance does not have any significant impact. Phonological distances do not have any significant effect on both response time variables.

7. Discussion and Conclusion

Addressing our first research question **(RQ1)**, the study reveals the following findings:

1. Non-compositional expression comprehension scores are highest for Belarusian and Ukrainian, languages within the same (East Slavic) group as Russian. The response times for these languages are the lowest.

2. Notably, there is minimal difference in the performance metrics between Belarusian and Ukrainian.

3. Bulgarian, the only representative of the South Slavic group, scored lower than East Slavic languages but higher than West Slavic languages

(Polish and Czech). This could be attributed to the use of Cyrillic script in Bulgarian, which likely facilitated intercomprehension by native Russian speakers.

4. Within the West Slavic group, participants exhibited significantly lower scores in the free translation task for Czech compared to Polish. However, only a slight difference was observed in the multiple-choice question task performance between Czech and Polish.

5. Overall, the observed pattern in scores aligns with the traditional linguistic classification of Slavic languages.

Regarding the second research question **(RQ2)** we demonstrate that:

1. The percentage of correct responses in both experimental tasks exhibits a strong and statistically significant correlation with phonological distance between foreign and Russian non-compositional expressions. For all target languages, this correlation is stronger than the correlation with orthographic distance. Although it may seem surprising, it is in line with previous research (e.g. Jagrova, 2022).

2. Accuracy in MCQ task additionally has a significant positive correlation with the phonological distance between foreign non-compositional and Russian literal expressions, but only for East Slavic languages. The positive correlation suggests that when making a choice between a non-compositional and literal Russian expressions, participants are likely to choose non-compositional expression if the distance between foreign non-compositional and Russian literal expression is large.

3. Response time in both tasks has a stronger relationship with LM surprisal (especially for masked model ruBERTa-large) for all languages except Ukrainian, which supports our initial hypothesis and suggests that advanced language models can reflect the difficulty in congitive processing. We do not observe a strong correlation of response time with any of the linguistic distance variables.

4. Response time for Ukrainian language, in contrast to all other target languages, does not show any significant correlation with LM surprisal. The absence of this correlation suggests a greater difference in the perception of non-compositional expressions between humans and language models in Ukrainian compared to other languages. Additionally, we hypothesize that additional factors, such as cultural influences or variations in participant demographics, may contribute to the observed results for Ukrainian. Further investigation into these potential factors is required to gain a better understanding of this phenomenon.

5. From the multiple regression analysis involving the data for all language sets jointly, we can additionally see the impact of both masked and autoregressive language models on accuracy in both tasks. This fact is significant, considering that the two types account for both the contextual information to the left and the entire sentential context, recognizing their joint importance in predicting the intelligibility scores.

In summary, this research contributes to our understanding of how non-compositional expressions are comprehended across languages, with implications for fields such as linguistics, cognitive science, and natural language processing. Future research could explore the differences of cross-lingual non-compositional comprehension intelligibility in written and spoken modality.

Acknowledgements

We would like to thank the anonymous reviewers for their constructive and insightful feedback on the paper. This research is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Project-ID 232722074 – SFB 1102 and by Saarland University (UdS-Internationalisierungsfonds).

Limitations

While this study offers valuable insights into the cross-lingual intelligibility of non-compositional expressions in context, it is essential to acknowledge certain limitations. Firstly, because we rely only on native speakers of Russian as study participants, the findings may not be fully generalizable to other language groups and even to other Slavic languages. Secondly, our analyses were conducted using Language Models specifically tailored for Russian, which means we need to be cautious when applying the results to other languages. Additionally, the predictive factors used in the study, including linguistic distances and surprisal scores, may not fully capture all the complexities of cross-linguistic intelligibility. Factors such as semantic similarity, syntactic structures, and cultural nuances could also play significant roles but were not included in our analysis. Acknowledging and addressing these limitations is crucial for a thorough understanding of the study's findings.

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.

Ben Alderson-Day and Charles Fernyhough. 2015. Inner speech: Development, cognitive functions, phenomenology, and neurobiology. *Psychological Bulletin*, 141(5):931–965.

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.

Cristina Cacciari and Patrizia Tabossi. 1988. The comprehension of idioms. *Journal of Memory and Language*, 27(6):668–683.

Jelena Golubović Anja Schüppert Femke Swarte Charlotte Gooskens, Vincent J. van Heuven and Stefanie Voigt. 2018. Mutual intelligibility between closely related languages in europe. *International Journal of Multilingualism*, 15(2):169–193.

Kathy Conklin and Norbert Schmitt. 2008. Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics*, 29:82–89.

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

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. 2013. Experimental methods for measuring intelligibility of closely related language varieties. In *The Oxford Handbook of Sociolinguistics*.

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. 2022. *Mutual intelligibility*, pages 51–95. Cambridge University Press.

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

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

Klara Jagrova. 2022. *Reading Polish with Czech Eyes: Distance and Surprisal in Quantitative, Qualitative, and Error Analyses of Intelligibility*. universaar.

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

Klára Jágrová, Michael Hedderich, Marius Mosbach, Tania Avgustinova, and Dietrich Klakow. 2021. On the correlation of context-aware language models with the intelligibility of polish target words to czech readers. *Frontiers in Psychology*, 12.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Tomáš Machálek. 2020. Kontext: Advanced and flexible corpus query interface. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7003–7008, Marseille, France. European Language Resources Association.

Kanishka Misra. 2022. minicons: Enabling flexible behavioral and representational analyses of transformer language models. *arXiv preprint arXiv:2203.13112*.

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

Marius Mosbach, Irina Stenger, Tania Avgustinova, and Dietrich Klakow. 2019. incom.py - a toolbox for calculating linguistic distances and asymmetries between related languages. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 810–818, Varna, Bulgaria. INCOMA Ltd.

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

S. B. Needleman and C. D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48:443–453.

B.H. Partee. 2008. *Compositionality in Formal Semantics: Selected Papers*. Explorations in Semantics. Wiley.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical report.

Giulia Rambelli, Emmanuele Chersoni, Marco S. G. Senaldi, Philippe Blache, and Alessandro Lenci. 2023. Are frequent phrases directly retrieved like idioms? an investigation with self-paced reading and language models. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 87–98, Dubrovnik, Croatia. Association for Computational Linguistics.

SberDevices. 2023. ruroberta-large. Hugging Face Model Hub.

Anja Schüppert, Johannes C. Ziegler, Holger Juul, and Charlotte Gooskens. 2022. On-line activation of l1 danish orthography enhances spoken word recognition of swedish. *Nordic Journal of Linguistics*, 45:80–98.

Anna Siyanova-Chanturia, Kathy Conklin, and Norbert Schmitt. 2011. Adding more fuel to the fire: An eye-tracking study of idiom processing by native and non-native speakers. *Second Language Research*, 27(2):251–272.

I. Stenger and T. 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. 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, Philip Georgis, Tania Avgustinova, Bernd Möbius, and Dietrich Klakow. 2022. Modeling the impact of syntactic distance and surprisal on cross-Slavic text comprehension. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7368–7376, Marseille, France. European Language Resources Association.

Irina Stenger, Klara Jagrova, and Tania Avgustinova. 2020. The INCOMSLAV platform: Experimental website with integrated methods for measuring linguistic distances and asymmetries in receptive multilingualism. In *Proceedings of the LREC 2020 Workshop on "Citizen Linguistics in Language Resource Development"*, pages 40–48, Marseille, France. European Language Resources Association.

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.

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

Debra Titone, Kyle Lovseth, Kristina Kasparian, and Mehrgol Tiv. 2019. Are figurative interpretations of idioms directly retrieved, compositionally built, or both? evidence from eye movement measures of reading. *PsyArXiv*.

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.

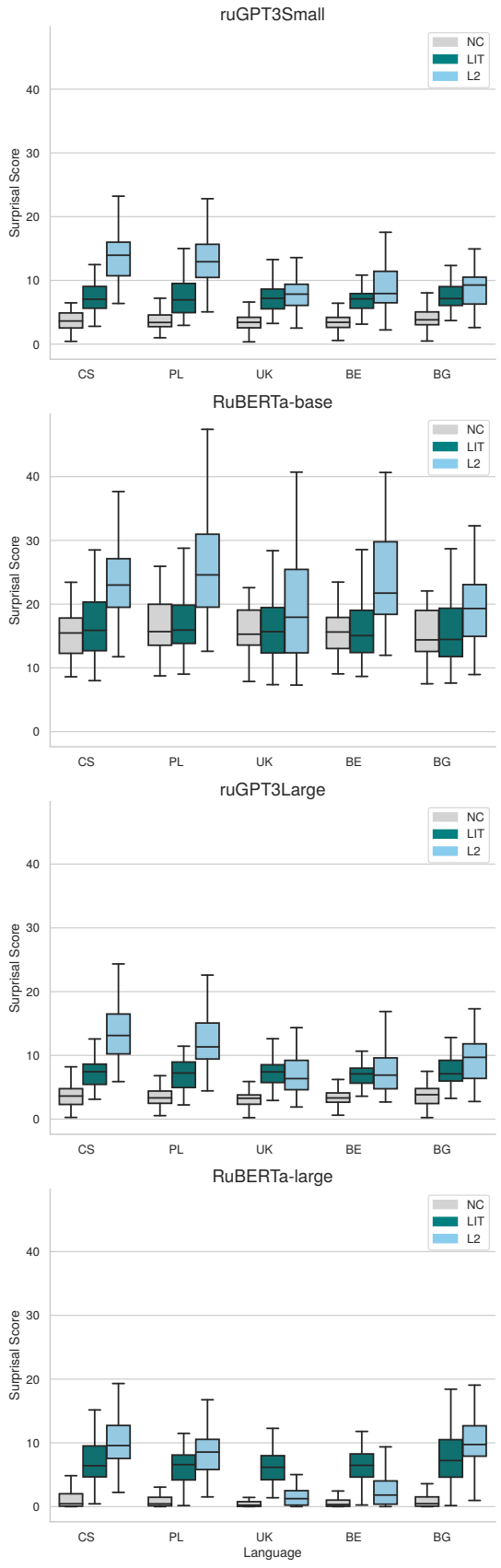
Francesco Vespignani, Paolo Canal, Nicola Molinaro, Sergio Fonda, and Cristina Cacciari. 2009. Predictive mechanisms in idiom comprehension. *Journal of cognitive neuroscience*, 22:1682–700.

Iuliia Zaitova, Badr Abdullah, and Dietrich Klakow. 2022. Mapping phonology to semantics: A computational model of cross-lingual spoken-word recognition. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 54–63, Gyeongju, Republic of Korea. Association for Computational Linguistics.

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.

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

A. Language Model Surprisal



Surprisal across models and languages for LIT, NC, and L2 phrases

B. Correlation Tables

Correlation of distances and surprisal with FT task accuracy

Metrics	BE	BG	CS	PL	UK
nWAS					
NC-L2	-0.206 (NS)	-0.287*	-0.025 (NS)	0.067 (NS)	-0.216 (NS)
PWLD					
NC-L2	**-0.405****	**-0.471*****	**-0.361****	**-0.428*****	**-0.606*****
ruGPT3S					
NC	-0.297*	0.035 (NS)	-0.082 (NS)	0.011 (NS)	-0.180 (NS)
L2	-0.296*	-0.232 (NS)	-0.236 (NS)	-0.162 (NS)	-0.131 (NS)
ruGPT3L					
NC	-0.303*	0.053 (NS)	-0.046 (NS)	0.050 (NS)	-0.151 (NS)
L2	-0.299*	-0.247 (NS)	-0.257*	-0.220 (NS)	-0.106 (NS)
ruBL					
NC	0.093 (NS)	0.046 (NS)	-0.097 (NS)	0.061 (NS)	-0.286*
L2	-0.062 (NS)	-0.173 (NS)	-0.233 (NS)	-0.209 (NS)	-0.255*

Correlation of distances and surprisal with MCQ task accuracy

Metrics	BE	BG	CS	PL	UK
nWAS					
NC-L2	-0.102 (NS)	-0.266*	-0.171 (NS)	-0.001 (NS)	-0.279*
LIT-L2	-0.044 (NS)	-0.018 (NS)	0.064 (NS)	-0.097 (NS)	0.064 (NS)
PWLD					
NC-L2	-0.229 (NS)	**-0.417****	**-0.283***	**-0.385****	**-0.502*****
LIT-L2	**0.429*****	-0.035 (NS)	0.063 (NS)	0.056 (NS)	0.307*
ruGPT3S					
NC	-0.216 (NS)	-0.147 (NS)	0.098 (NS)	0.084 (NS)	-0.030 (NS)
LIT	-0.015 (NS)	0.055 (NS)	0.143 (NS)	0.301*	0.254 (NS)
L2	-0.021 (NS)	-0.064 (NS)	0.210 (NS)	-0.019 (NS)	0.001 (NS)
ruGPT3L					
NC	-0.208 (NS)	-0.124 (NS)	0.115 (NS)	0.007 (NS)	0.047 (NS)
LIT	0.018 (NS)	0.039 (NS)	0.117 (NS)	0.260*	0.253 (NS)
L2	-0.030 (NS)	-0.025 (NS)	0.177 (NS)	-0.045 (NS)	0.076 (NS)
ruBL					
NC	0.041 (NS)	-0.025 (NS)	-0.013 (NS)	-0.063 (NS)	-0.169 (NS)
LIT	0.195 (NS)	0.188 (NS)	0.122 (NS)	0.338**	0.153 (NS)
L2	0.040 (NS)	-0.042 (NS)	0.189 (NS)	-0.010 (NS)	-0.129 (NS)

*=p< .05, **=p< .01, and ***=p< .001. Pearson correlation of intelligibility metrics

Correlation of distances and surprisal with for FT task time

Metrics	BE	BG	CS	PL	UK
nWAS					
NC-L2	0.145 (NS)	0.078 (NS)	0.312*	0.133 (NS)	-0.039 (NS)
PWLD					
NC-L2	0.060 (NS)	-0.047 (NS)	0.026 (NS)	0.221 (NS)	0.068 (NS)
ruGPT3S					
NC	0.225 (NS)	0.003 (NS)	0.278*	-0.028 (NS)	0.143 (NS)
L2	0.363**	**0.318***	0.501***	0.201 (NS)	0.177 (NS)
ruGPT3L					
NC	0.265*	0.009 (NS)	0.280*	0.020 (NS)	0.080 (NS)
L2	0.410**	0.277*	0.441***	0.182 (NS)	0.209 (NS)
ruBL					
NC	0.311*	-0.120 (NS)	0.223 (NS)	0.149 (NS)	**0.222** (NS)
L2	**0.443*****	0.135 (NS)	**0.547*****	**0.304***	0.217 (NS)

Correlation of distances and surprisal with MCQ task time

Metrics	BE	BG	CS	PL	UK
nWAS					
NC-L2	0.118 (NS)	0.078 (NS)	0.285*	0.023 (NS)	-0.101 (NS)
LIT-L2	-0.144 (NS)	**-0.322***	0.125 (NS)	0.007 (NS)	-0.180 (NS)
PWLD					
NC-L2	0.000 (NS)	-0.019 (NS)	0.040 (NS)	0.135 (NS)	0.062 (NS)
LIT-L2	0.007 (NS)	-0.220 (NS)	0.151 (NS)	0.057 (NS)	-0.013 (NS)
ruGPT3S					
NC	0.229 (NS)	0.015 (NS)	0.213 (NS)	-0.093 (NS)	0.143 (NS)
LIT	0.155 (NS)	-0.008 (NS)	0.167 (NS)	0.235 (NS)	0.013 (NS)
L2	0.311*	**0.322***	0.374**	0.089 (NS)	0.131 (NS)
ruGPT3L					
NC	0.278*	0.029 (NS)	0.201 (NS)	-0.047 (NS)	0.113 (NS)
LIT	0.183 (NS)	-0.032 (NS)	0.199 (NS)	0.254 (NS)	0.019 (NS)
L2	0.368**	0.289*	0.309*	0.079 (NS)	0.174 (NS)
ruBL					
NC	0.330*	-0.102 (NS)	0.191 (NS)	0.154 (NS)	0.083 (NS)
LIT	0.142 (NS)	0.200 (NS)	0.323 (NS)	0.288*	-0.080 (NS)
L2	**0.457*****	0.102 (NS)	**0.452*****	**0.308***	**0.215** (NS)

*=p< .05, **=p< .01, and ***=p< .001. Pearson correlation of time metrics