

On Slavic cognate recognition in context

Irina Stenger

Saarland University / Campus C5.3,
66123 Saarbrücken, Germany
ira.stenger@mx.uni-
saarland.de

Tania Avgustinova

Saarland University / Campus C7.2,
66123 Saarbrücken, Germany
avgustinova@coli.uni-
saarland.de

Abstract

This study contributes to a better understanding of reading intercomprehension as manifested in the intelligibility of East and South Slavic languages to Russian native speakers in contextualized cognate recognition experiments using Belarusian, Ukrainian, and Bulgarian stimuli. While the results mostly confirm the expected mutual intelligibility effects, we also register apparent processing difficulties in some of the cases. In search of an explanation, we examine the correlation of the experimentally obtained intercomprehension scores with various linguistic factors, which contribute to cognate intelligibility in a context, considering common predictors of intercomprehension associated with (i) morphology and orthography, (ii) lexis, and (iii) syntax.

Keywords: Slavic intercomprehension, online experiments, cognate recognition, linguistic context

DOI: 10.28995/2075-7182-2021-20-660-668

О распознавании славянских слов-когнатов в контексте

Ирина Штенгер

Университет земли Саар / Кампус
C5.3, 66123 Саарбрюккен, Германия
ira.stenger@mx.uni-
saarland.de

Таня Августинова

Университет земли Саар / Кампус
C7.2, 66123 Саарбрюккен, Германия
avgustinova@coli.uni-
saarland.de

Аннотация

Данное исследование способствует лучшему пониманию межъязыковой понятности славянских языков в ситуации, когда читателю необходимо извлечь информацию из текста при чтении незнакомого, но (близко)родственного языка. В частности, мы изучаем понятность восточнославянских и южнославянских языков носителями русского языка. В настоящей статье основное внимание уделяется вопросам спонтанного понимания белорусских, украинских и болгарских слов-когнатов в контексте. Результаты, полученные в ходе проведенных онлайн-экспериментов, свидетельствуют в целом о высокой степени распознавания славянских стимулов носителями русского языка, а также о некоторых трудностях в процессе понимания. В поисках объяснения полученных результатов мы рассматриваем лингвистические факторы, которые могут предсказать степень понятности славянских слов-когнатов в контексте. Результаты онлайн-экспериментов сравниваются со следующими потенциальными параметрами, оценивающими понятность незнакомого, но (близко)родственного языка при чтении: (i) морфологические и орфографические факторы, (ii) лексические факторы и (iii) синтаксические факторы.

Ключевые слова: славянская межъязыковая понятность, онлайн-эксперименты, распознавание когнатов, лингвистический контекст

1 Introduction

Multiple studies have evidenced that reading is a complex and structured process that involves not only familiarity with linguistic elements, but also the entire knowledge of the reader [Frost 2012], [Lutjeharms 2004]. In an intercomprehension scenario, readers have the potential to understand a message encoded in an unknown (stimulus) language if they are speakers of or have in their linguistic repertoire some genetically related language(s). The previous research on reading intercomprehension indicates a remarkably good performance of Russian speaking subjects in spontaneous out-of-context guessing of isolated cognate words from other Cyrillic-script Slavic languages [Stenger 2019]. Context-free cognate recognition is, of course, quite far from the real-world situation, inasmuch as inferences based on contextual assumptions represent a central technique in intercomprehension. On the other hand, for a context to be useful it needs to be understandable, too. We shall focus on the contextualized cross-lingual cognate recognition, looking for predictors of the human performance in intelligibility tests.

The assumption that cognate words across genetically related languages are better recognizable in a context than as isolated items is quite intuitive and may sound trivial. To model and predict this phenomenon, however, we need to differentiate the linguistic factors that facilitate in-context recognition from those that do not. The shape of cognate words may change in the course of time to the extent of being no longer transparent to a reader with no historical-linguistic background. This poses the question of morphological and orthographic similarities regarding their potential to predict and explain the results of intelligibility tests. On the other hand, lexical and syntactic (dis)similarities appear to influence more directly the use of context in intercomprehension. We shall consider all these linguistic factors to find out to what extent a particular context facilitates the understanding of target words, and why this is not always the case.

After presenting the experimental setup and results, we consider potential predictors of cross-lingual contextualized cognate intelligibility, analyze the influence of individual linguistic factors on human performance, and draw conclusions.

2 Intercomprehension experiment

The stimuli material represents a collection of parallel sentences from the parallel text corpus of the Russian National Corpus¹, which contains the words we are interested in here and which we have previously tested as stimuli² for online single-word free translation tasks. In such a way, we can obtain intercomprehension scores for native speakers of Russian (RU). To reveal the inherent intercomprehension, we include in the analysis only people who do not know the stimulus language. To avoid possible learning effects, we consider the results of the participant's initial experiment, and exclude subjects³ who have already completed other experiments at the intercomprehension website.⁴ Thus, the number of participants is 87, aged between 17 and 49 years (i.e. average age 22) with 79 women and 8 men. After completing a background questionnaire, they were introduced to 60 Belarusian (BE), 60 Ukrainian (UK) and 120 Bulgarian (BG) sentences⁵ in 16 sessions (15 sentences per session). The number of subjects for each stimulus sentence ranges from 35 to 46 (i.e. in average per session: 40 participants). In a session on contextualized cognate recognition, the subjects see a randomized stimulus sentence on their screen (see Figure 1) and have 10 seconds to translate the marked target word.⁶

¹ <https://ruscorpora.ru/new/>

² The stimuli items come from parallel lists consisting of internationalisms, Pan-Slavic vocabulary, and cognates from Slavic Swadesh lists – for more details cf. [Stenger 2019].

³ In this case, 4 out of 91 participants were excluded from the analysis.

⁴ For our online experiments, we use the INCOMSLAV platform. The website includes a large number of different online experiments in 11 Slavic languages (as well as in German and English) carried out as challenges in a linguistic game (<https://intercomprehension.coli.uni-saarland.de/en/>).

⁵ The mean length of BE sentences: 9 tokens, UK sentences: 8 tokens, BG sentences: 8 tokens.

⁶ In addition, 3 seconds per item in a sentence are given for reading the whole stimulus.

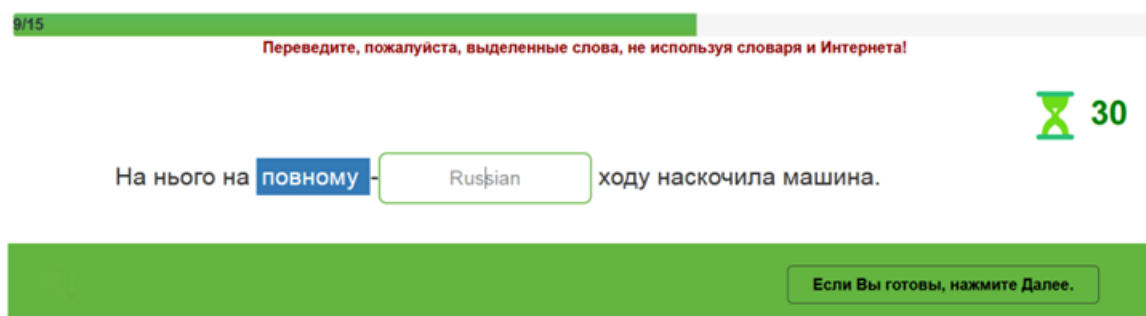


Figure 1: Contextualized cognate translation from UK into RU of ‘He was run over by a car at full speed’. Instruction on top: “Please translate the marked words without a dictionary or the internet!”

The target words occupy different sentence positions equally distributed in the respective stimuli sets: BE – 22% initial, 40% middle, 38% end; UK – 20% initial, 47% middle, 33% end; BG – 23% initial, 41% middle, 36% end. After the first session (25 subjects started with UK stimuli, 26 with BE, and 36 with BG), each participant may continue the experiment by completing the task for the remaining stimulus languages offered in a random order. Their responses are categorized automatically via pattern matching with pre-defined correct answers and acceptable alternatives, and checked manually in the final analysis. The mean percentage of successfully translated items constitutes the achieved intercomprehension score for a given constellation (Table 1). The overall cognate intelligibility order attested here for the in-context condition resembles what [Stenger 2019] has observed for the out-of-context condition: UK > BE > BG. While the differences in the comprehension of UK and BG are in favor of the out-of-context condition, the role of context appears more prominent for BE.

stimuli–subject	in-context (this study)	out-of-context [Stenger 2019]
BE–RU group	81.32%	72.56%
UK–RU group	84.93%	85.61%
BG–RU group	69.78%	71.33%

Table 1: Cross-lingual intelligibility of cognates for RU speaking subjects

Regarding the number of word pairs with higher intercomprehension scores in the respective condition, observe that more cognates are successfully translated in-context than out-of-context, e.g., BE 48 > 11 (1)⁷, UK 25 > 23 (12), BG 70 > 46 (4). The results are correspondingly visualized as a percentage of correctly produced translations for BE (Figure 2), UK (Figure 3), and BG (Figure 4). In these representations, cognate pairs better recognized in-context are presented to the left, those more successfully guessed out-of-context – to the right, and stimuli yielding identical scores for both conditions – in the middle.

As human performance depends not only on context (un)availability, but also on the individual forms, here are some examples. The UK stimulus *dim* (dim) ‘house’ is correctly understood to 51% in-context and only to 11% without context. However, the UK stimulus *lito* (lito) ‘summer’ is better understandable to RU subjects as a single word (95%) rather than in context: *влітку* (vlitku) ‘in summer’ (17%). The BE stimulus *дзіця* (dzicja) ‘child’ is to 95% correctly translated in-context and only to 35% without context. Yet, the BE stimulus *бераг* (berah) ‘bank’ is to 71% successfully guessed as a single word in comparison to its only 8% recognition in context: *на беразе* (na beraze) ‘on the bank’. The BG stimulus *пет* (pet) ‘five’ is much better understood with context than without (76% intelligibly vs. 15%), while *славей* (slavej) ‘nightingale’ causes less difficulties as a single word (90%) than in context (5%) *славоят* (slavejat) ‘the nightingale’.

⁷ The number of cognate pairs yielding identical scores for both conditions is given in brackets.

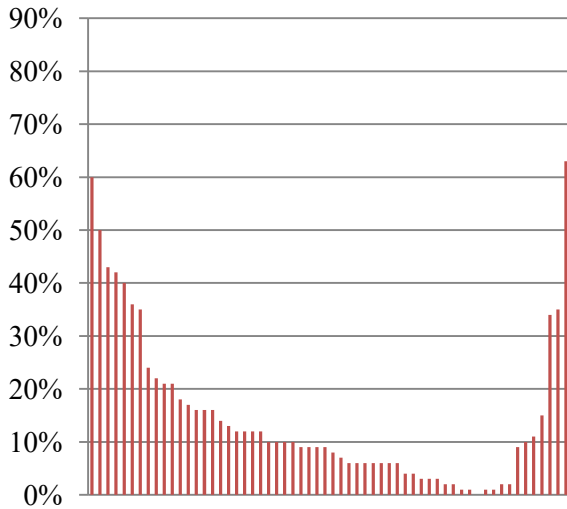


Figure 2: Number of successfully guessed cognates in context (left) vs. without context (right): BE

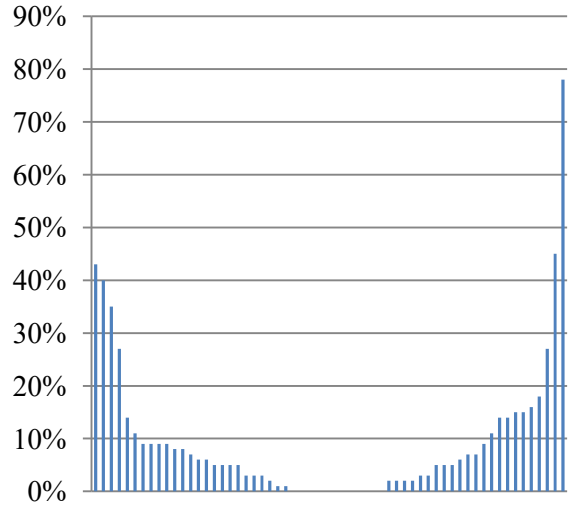


Figure 3: Number of successfully guessed cognates in context (left) vs. without context (right): UK

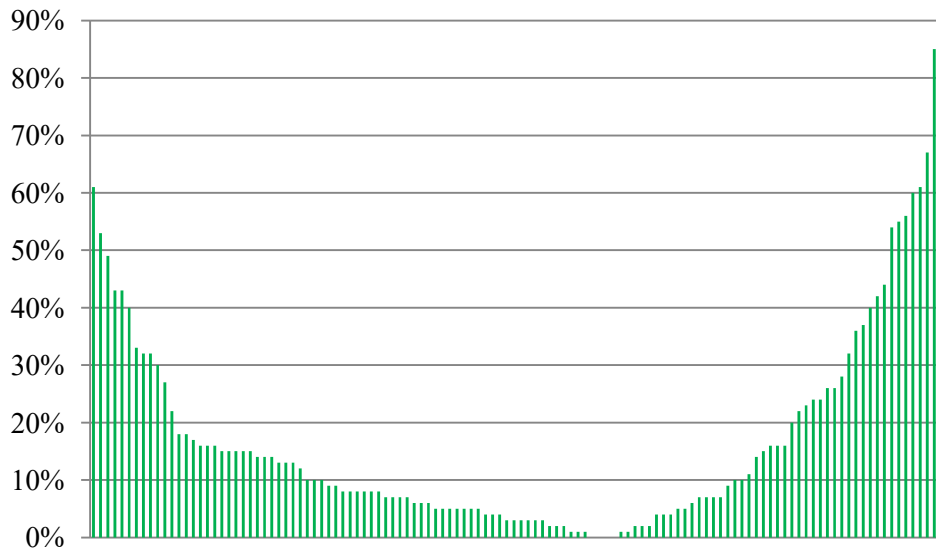


Figure 4: Number of successfully guessed cognates in context (left) vs. without context (right): BG

3 Measuring contextualized cognate intelligibility

Morphological and orthographic factors: A string similarity measure [Levenshtein 1965] known as Levenshtein distance (LD) approximates synchronically observable morphological and orthographic characteristics of cross-linguistic correspondences. We perform objective LD computations between word-forms automatically, using the *incom.py* tool [Mosbach et al. 2019]. In a previous study [Stenger and Avgustinova 2020] have successfully used the normalized LD to measure orthographic and phonetic distances between related languages. Here, we assume that the larger the distance between a stimulus and its correspondence in subjects' native language is, the more difficult intercomprehension would be. Figure 5 shows a sample LD calculation between contextualized word-forms for 'head' (in 'head downwards') – cf. the BG stimulus *с главата надолу* (*s glavata nadolu*,) and its corresponding RU translation *вниз головой* (*vniz golovoj*). To normalize, the sum of all costs for the character-by-character transformation of one string into another ($1+1+1+1+1=5$) we divide the number of alignment slots (8), which results in a normalized LD (nLD) of 0.63 or 63%. Applying this method, we obtain the mean nLD (%) between stimulus word-forms and their target cognates in the subjects' language (Table 2).

	1	2	3	4	5	6	7	8
BG	г		л	а	в	а	т	а
RU	г	о	л	о	в	о	й	
Cost	0	1	0	1	0	1	1	1

Figure 5: LD between in-context cognates

We further apply the information-theoretical notion of surprisal [Shannon 1948] to model the predictability of a particular cross-lingual correspondence for a given pair of source and target languages. In particular, we assume that higher word adaptation surprisal (WAS) values cause more intercomprehension difficulties, so that it would be harder to recognize the actual cognate stimulus (BE, UK, BG) from the viewpoint of subjects' language (RU). WAS corresponds to the sum of the character adaptation surprisal (CAS) values, and is calculated in bits according to the character transformation probabilities⁸. Again, we use the *incom.py* tool [Mosbach et al. 2019] for calculating the mean normalized WAS (nWAS) values (in bits) for the stimulus word-forms w.r.t. their target cognates in the subjects' language (Table 2).

in-context word-forms	mean nLD to RU cognates	mean nWAS w.r.t. RU
BE	34%	0.64
UK	29%	0.59
BG	37%	1.06

Table 2: Mean normalized Levenshtein distance and mean normalized word adaptation surprisal between BE, UK, BG and RU

Lexical factors: As non-cognates, i.e. historically non-related words, tend to be unintelligible to readers with no prior knowledge of the stimulus language, we expect their large proportion to impede intercomprehension [Gooskens 2019]. As a rule, the percentage of non-cognates indicatively determines the lexical distance between closely related languages. The so-called *false friends* need special attention here, since they may cause even larger difficulties than non-cognates. For marking cross-lingual lexical differences, we give points to cognate word-form pairs of aligned sentences, namely: a non-cognate (incl. false friends) obtains one point, a partial cognate⁹ obtains half a point, and a cognate (i.e. with a common root and similar meaning) obtains zero points. In some cases the cognate correspondence may consist of non-cognates, for example, BG *око* (oko) 'eye' translates into RU *глаз* (glaz), forming a pair of words that are non-cognate. Nevertheless, such a word pair obtains zero points too, as there is a synonym *око* (oko) in RU, which makes the written BG word *око* (oko) understandable. Our hypothesis is that with an increasing percentage of non-cognates, partial cognates and false friends, the subjects perceive less similarity, which makes it more difficult for them to understand an unknown even though related language. In Table 3 we present for each language pair the mean lexical distances.¹⁰

⁸ $CAS(L1 = c1|L2 = c2) = -\log_2 P(L1 = c1|L2 = c2)$

L1 – native language, c1 – character of the native language

L2 – stimulus language, c2 – character of the stimulus language

⁹ Partial cognates are words with a similar meaning, for example, BG *жена* (žena) 'woman' and RU *жена* (žena) 'wife' both point the reader towards the correct category of adult female (cf. [Golubović 2016: 244]).

¹⁰ This calculation is based on lexical words (i.e. nouns, adjectives, verbs, adverbs, and numerals), as they are more important for intelligibility than function words (cf. [van Bezooijen and Gooskens 2005], [Gooskens 2006]).

language pair	mean lexical distances
BE–RU	9%
UK–RU	12%
BG–RU	13%

Table 3: Lexical distance between BE, UK, BG and RU

Syntactic factors: When processing a sentence in a (closely) related language, the readers may experience that some items are displaced, missing or superfluous. To approximate this situation, we use the respective syntactically relevant measures of movement, insertion and deletion [Heeringa et al. 2017]. As an illustration, let us take the BG–RU alignment in Figure 6.

	1	2	3	4	5	6	7	8	9	10	11
BG		Искам		и	моят	мъж	да	бъде	свободен	като	вятъра
RU	Я	хочу	чтобы	и	мой	муж		был	волен	как	ветер

Figure 6: BG–RU alignment for the sentence ‘I want my husband to wander as free as the wind.’

Our assumption is that the more positions a word moves and the more words need to be added or deleted in order to achieve a perfect cross-lingual correspondence, the more negative the effect on intercomprehension tends to be. Here, the RU *чтобы* (čtoby ‘in order to’: position 3) corresponds to the BG *да* (da: position 7), even if they occur in different alignment slots and are not cognates of each other, so the number of movements here is $7-3=4$. With the InDel measure, we register insertions and deletions in the alignment slots. From the perspective of RU subjects, we need one insertion, at the first position, which results in InDel value 1. Normalizing the InDel distance by the number of the alignment slots (11), we obtain nInDel: 0.09 or 9% (cf. [Gooskens and Swarte 2017]). Table 4 summarizes the mean nInDel and movement distances from the stimuli sentence alignments to their RU targets.

stimulus language	subjects’ language: RU	
	mean nInDel distance	mean movement distance
BE	11% (1.15)	1.38 (0.33)
UK	7% (0.87)	1.67 (0.35)
BG	16% (1.50)	2.04 (0.58)

Table 4: Syntactic distances between BE, UK, BG and RU

The average number of words to be inserted or deleted per sentence in the stimulus from a RU perspective are presented (in brackets) after the respective nInDel value. Similarly, the average number of words to be moved to another sentence position from a RU perspective are given (in brackets) after the respective movement value.

4 Potential predictors of human performance

We can now examine to what extent the presented linguistic factors (cf. Section 3) can predict and explain the experimental results (cf. Section 2) on cross-lingual in-context recognition of cognates. Regarding the impact of morphological and orthographic factors on human performance in intelligibility tests, we correlate the obtained intelligibility scores with the calculated nLD and nWAS between their stimulus word-forms and their target cognates in RU¹¹. To understand the role of the context in cognate recognition, we correlate the obtained intercomprehension scores with the

¹¹ The correlations with the nLD and nWAS for the language group BG–RU are based on 118 BG–RU cognate pairs, since two word pairs consist of non-cognates, compare BG–RU: *киселец-щавель* (kiselece-ščavel’) ‘dock’ and *барут-порох* (barut-poroch) ‘gun powder’.

calculated lexical and syntactic distances. Table 6 presents the correlation coefficients (Pearson's r) and the p -values. Negative correlations can be expected for all linguistic factors.

The correlation of the intercomprehension scores with the nLD is higher than with the nWAS, and it is significant for all groups. This reveals a clear relationship between morphological and orthographic similarities and successful in-context cognate recognition. Yet, the nWAS is likely to be the predictor of intelligibility only for UK and BG stimuli, but it shows no significant correlation with the intelligibility of BE stimuli.

The effect of lexical distances may be more difficult to predict. The correlation of the intelligibility scores for BG with the lexical distance is negative but very small and not significant. For BE and UK cognates the correlations are not negative, but insignificant.

As far as the syntactic distances are concerned, the highest negative correlation is found for the nInDel distance in BE cognate recognition. For UK and BG cognates, the correlations are negative but smaller. However, none of them is significant. Furthermore, the movement distance seems not to play any role for RU readers. For all three languages, the correlations are not negative as assumed, but insignificant.

in-context cognate recognition	linguistic factors				
	nLD	nWAS	lex. distance	nInDel	movement
BE–RU group	$r = -.342$ $p < .01$	$r = -.178$ $p = .17$	$r = .036$ $p = .78$	$r = -.210$ $p = .11$	$r = .169$ $p = .20$
UK–RU group	$r = -.536$ $p = 1e-05$	$r = -.447$ $p < .0005$	$r = .055$ $p = .67$	$r = -.075$ $p = .57$	$r = .011$ $p = .93$
BG–RU group	$r = -.413$ $p = 3.40e-06$	$r = -.191$ $p < .05$	$r = -.171$ $p = .06$	$r = -.113$ $p = .22$	$r = .041$ $p = .66$

Table 6: Correlations between linguistic factors and intercomprehension scores of RU subjects in three stimuli groups

5 Discussion and conclusion

The observed human performance validates the intuition that Russian-speaking subjects understand better East Slavic, to which RU belongs together with UK and BE, than South Slavic. Comparing “out-of-context guessing” with “in-context recognition”, we see that the latter is much better only for BE–RU, while the former shows slightly higher intercomprehension scores for UK–RU and BG–RU. However, a closer look at individual cognate pairs reveals that in all three stimuli-subject combinations, more cognates are successfully recognized in a context than without a context. The slightly better performance in single word recognition in UK–RU and BG–RU is due to the higher intelligibility score of some correctly translated single words in the two groups.

<i>stimuli–subject</i>	“free translation” out-of-context guessing	“contextualized translation” in-context recognition
<i>UK–RU group</i>	slightly higher	
<i>BE–RU group</i>		much better
<i>BG–RU group</i>	slightly higher	
<i>overall intelligibility</i>		more successful
<i>subject performance</i>	UK>BE>BG	UK>BE>BG

The present investigation has shown that morphological and orthographic factors play an important role in correct cognate recognition with context. The nLD is a significant predictor of intelligibility of in-context cognates in all three Slavic language groups. In addition to morphological and orthographic similarities, the predictability of correspondences by means of nWAS seems to influence the successful human performance too, but the significance of this variable holds only for UK and BG cognates in context.

The relationship between lexical distances and intelligibility of cognates in context is less clear. Investigating linguistic predictors of inter-Scandinavian intelligibility, [Gooskens 2006: 111] points

out that the impact of lexical differences depends on the nature of the lexical deviances and in some cases one single deviant word can be very disturbing for the comprehensibility of the whole context while in other cases a number of non-cognates is hardly disturbing, because they are not important concepts. Examining the mutual intelligibility of some West and South Slavic languages, [Golubović 2016: 123] found out that lexical distance (on the basis of all words) is a good predictor of mutual intelligibility in individual word translation tasks and cloze tests. In our experiments, the quantitative effect of lexical differences seems not to play so important role. The readers seem to pay more attention to morphological and orthographic similarities and differences of the particular cognate pair in order to succeed in intelligibility tests. We assume that lexical distance may be a better predictor in non-cognate recognition (cf. the role of predictive context in [Jágrová and Avgustinova 2019]).

We have also compared the experimental results to the two presented syntactic distances and found negative correlations between the nInDel distance and the intelligibility scores for all stimulus languages. However, the correlations are low and none of the correlations is significant. According to [Heeringa et al. 2017] the InDel distance highly depends on the way sentences are translated. Indeed, in a reading intercomprehension scenario a reader is trying to match words in a stimulus sentence to the words in his or her native language. According to our results, it seems that the number of words that should be added or deleted in comparison to the closest possible sentence which the reader would have used him/herself, can be seen only as a tendency that has a negative effect on correct cognate recognition in context.

Additionally we found that the movement distances do not correlate negatively with the intelligibility scores, as assumed, but all of them were insignificant. This means that the assumption that the further a word is moved the more negatively it will affect intelligibility cannot be confirmed in this study. [Heeringa et al. 2017] pointed that [Swarte 2016] measured mutual intelligibility between five Germanic languages by means of a spoken and a written cloze test and correlated the intelligibility scores of the experiments with the movement distance. She found a significant correlation between the movement measure and written and spoken intelligibility at the 0.05 level. The reason that the movement distance does not explain the intelligibility of cognates in our experiments might be that the word order is not so rigid in Slavic languages, as, for example, in Germanic languages.

Linguistic similarity is a multidimensional phenomenon [van Heuven 2008]. In this investigation, we focused on linguistic factors that predict inherent intelligibility between related languages. New testing methods have been established in the last decade [Gooskens 2018] to define certain breakdown points at which language varieties become unintelligible. The goal is to provide a more solid and experimentally grounded foundation for the classical and traditional claims made by linguists about genealogical relatedness among languages. In future research, we shall extend our approach to include extra-linguistic factors with obvious impact on mutual intelligibility across languages [Gooskens and van Heuven 2020] in order to adequately model what is known as receptive multilingualism.

Acknowledgements

We thank Marius Mosbach for the implementation of our data in the *incom.py* tool, Philip Georgis for assistance in calculating of lexical and syntactic distances, and Hasan Alam for support with the implementation of online experiments. This work has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102.

References

- [1] van Bezooijen Renée, Gooskens Charlotte. How easy is it for speakers of Dutch to understand spoken and written Frisian and Afrikaans, and why? — *Linguistics in the Netherlands* 22. — John Benjamins, Amsterdam, 2005. — P. 13–24.
- [2] Frost Ram. Towards a universal model of reading. — *Behavioral and Brain Sciences* 35(5). — Cambridge University Press, 2012. — P. 263–329.
- [3] Golubović Jelena. Mutual intelligibility in the Slavic language area. — University of Groningen, PhD thesis, 2016.
- [4] Gooskens Charlotte. Linguistic and extra-linguistic predictors of Inter-Scandinavian intelligibility. — *Linguistics in the Netherlands* 23. — John Benjamins, Amsterdam, 2006. — P. 101–113.

- [5] Gooskens Charlotte. Dialect intelligibility. — *Handbook of dialectology*. — Wiley-Blackwell, Oxford, 2018. — P. 204–218.
- [6] Gooskens Charlotte. Receptive multilingualism. — *Multidisciplinary perspectives on multilingualism: The fundamentals*. — De Gruyter Mouton, Berlin, 2019. — P. 149–174.
- [7] Gooskens Charlotte, van Heuven Vincent J. How well can intelligibility of closely related languages in Europe be predicted by linguistic and non-linguistic variables? — *Linguistic Approaches to Bilingualism* 10(3). — John Benjamins, Amsterdam, 2020. — P. 351–379.
- [8] Gooskens Charlotte, Swarte Femke. Linguistic and extra-linguistic predictors of mutual intelligibility between Germanic languages. — *Nordic Journal of Linguistics* 40(2). — Cambridge University Press, 2017. — P. 123–147.
- [9] Heeringa Wilbert, Swarte Femke, Schüppert Anja, Gooskens Charlotte. Measuring syntactical variation in Germanic texts. — *Digital Scholarship in the Humanities* 33(2). — Oxford University Press, 2017. — P. 279–296.
- [10] van Heuven Vincent J. Making sense of strange sounds: (Mutual) intelligibility of related language varieties. A review. — *International Journal of Humanities and Arts Computing* 2(1-2). — Edinburgh University Press, 2008. — P. 39–62.
- [11] Jágrová Klára, Avgustinova Tania. Intelligibility of highly predictable Polish target words in sentences presented to Czech readers. — *CICLing 2019, Springer's Lecture Notes in Computer Science: preprint*, 2019.
- [12] Levenshtein Vladimir I. Binary codes capable of correcting deletions, insertions, and reversals. — *Doklady of the Soviet Academy* 163(4), 1965. — P. 845–848.
- [13] Lutjeharms Madeline. Verarbeitungsebenen beim Lesen in Fremdsprachen. — *Neuere Forschungen zur Europäischen Interkomprehension*. — Shaker Verlag, Aachen, 2004. — P. 67–82.
- [14] Mosbach Marius, Stenger Irina, Avgustinova Tania, Klakow Dietrich. *incom.py* – A Toolbox for Calculating Linguistic Distances and Asymmetries between Related Languages // *Proceedings of the International Conference Recent Advances in Natural Languages Processing 2019*. — Varna, Bulgaria, 2019. — P. 811–819.
- [15] Shannon Claude E. A mathematical theory of communication. — *Bell System Technical Journal* 27(3), 1948. — P. 623–656.
- [16] Stenger Irina. Zur Rolle der Orthographie in der slavischen Interkomprehension mit besonderem Fokus auf die kyrillische Schrift. — *Universaar, Saarbrücken, PhD thesis*, 2019.
- [17] Stenger Irina, Avgustinova Tania. Visual vs. auditory perception of Bulgarian stimuli by Russian native speakers // *Proceedings of the Annual International Conference ‘Dialogue’ 2020*. — Moscow, Russia, 2020. — P. 623–656.
- [18] Swarte Femke. Predicting the Mutual Intelligibility of Germanic languages from linguistic and extra-linguistic factors. — *University of Groningen, PhD thesis*, 2016.