The INCOMSLAV Platform: Experimental Website with Integrated Methods for Measuring Linguistic Distances and Asymmetries in Receptive Multilingualism

Irina Stenger, Klára Jágrová, Tania Avgustinova
Saarland University, Collaborative Research Center (SFB) 1102: Information Density and Linguistic Encoding,
Campus A 2.2, 66123 Saarbrücken, Germany
Project C4: INCOMSLAV – Mutual Intelligibility and Surprisal in Slavic Intercomprehension
ira.stenger@mx.uni-saarland.de, {kjagrova, avgustinova}@coli.uni-saarland.de

Abstract

We report on a web-based resource for conducting intercomprehension experiments with native speakers of Slavic languages and present our methods for measuring linguistic distances and asymmetries in receptive multilingualism. Through a website which serves as a platform for online testing, a large number of participants with different linguistic backgrounds can be targeted. A statistical language model is used to measure information density and to gauge how language users master various degrees of (un)intelligibilty. The key idea is that intercomprehension should be better when the model adapted for understanding the unknown language exhibits relatively low average distance and surprisal. All obtained intelligibility scores together with distance and asymmetry measures for the different language pairs and processing directions are made available as an integrated online resource in the form of a Slavic intercomprehension matrix (SlavMatrix).

Keywords: Slavic languages, intercomprehension, linguistic distance, asymmetric intelligibility, surprisal-based modelling

1. Introduction

1.1 Background

The terms "intercomprehension" (Doyé, 2005), "receptive multilingualism" (Braunmüller and Zeevaert, 2001) or "semi-communication" (Haugen, 1966) all refer, on the one hand, to a communicative practice of understanding an unknown foreign language based on already acquired linguistic repertoire, and on the other hand to a field of study which exploits linguistic similarities to model this special mode of language use. Its success relies on various types of information: linguistic, communicative, contextual, socio-demographic, etc. In the last decade, researchers focused mostly on uncovering the variables that influence intercomprehension between related languages (Gooskens and Swarte, 2017), with the assumption that the more linguistic similarities two languages share, the higher their degree of mutual intelligibility. This is quite apparent for modern Slavic languages as descendants of a single ancestor – Proto- or Common Slavic – that can be reconstructed by comparing diachronically and synchronically attested language varieties (Carlton, 1991; Comrie and Corbett, 1993). In general, linguistic phenomena may be unique to a language, shared between two languages, or common to many languages from a given family. In addition, Ringbom (2007: 11) distinguishes cross-linguistically between *objective similarities* (established as symmetrical) and *perceived similarities* (not necessarily symmetrical). Asymmetric intelligibility can be of linguistic nature, e.g., if language A has more complicated rules and/or irregular developments than language B, this results in structural asymmetry (Berruto, 2004). It can also be due to extra-linguistic and socio-demographic factors like attitude, language exposure, age, level of education, linguistic repertoire etc.

1.2 This Paper

In the INCOMSLAV project, we employ language modelling and information-theoretic concepts to investigate various intercomprehension scenarios with Slavic languages. We report on a website for conducting intercomprehension experiments as a resource. Besides the experiments, the site contains an integrated overview of the experimental results (intelligibility scores) together with the respective linguistic distances and surprisal as predictors for the intelligibility. We present our methods for measuring linguistic distances and asymmetries between related languages. A statistical model of linguistic distance and surprisal is used to measure information density and to gauge how language users master various degrees of distance and surprisal in view of partial incomprehensibility. The key idea here is that comprehension of an unknown but related language should be better, when the language model adapted for understanding the unknown language exhibits relatively low average distance and surprisal. Thus, our approach is based on three pillars: (i) linguistic resources, (ii) language technologies, (iii) experimental study of intercomprehension. This article is organized as follows. Section 2 gives an overview of the INCOMSLAV experiment platform and the conducted tests. Section 3 presents our methods for measuring linguistic distances and asymmetries among related languages. In Section 4 we analyze so far the obtained results that are made available in the Slavic intercomprehension matrix. Finally, some general conclusions are drawn and future work is outlined.

2. The INCOMSLAV platform

We test the mutual intelligibility of Slavic languages by means of the following tests: (i) intelligibility at the word level (individual words in spoken and written modality); (ii) intelligibility at the phrasal level (adjective-noun sequences in NPs); (iii) intelligibility at the sentence level (target words in predictive context). All experiments are available at http://intercomprehension.coli.uni-saarland.de with an interface in 11 Slavic languages, English and German. The participants have been recruited through universities, Prolific Academic, and social media. The respondents are continuously encouraged to participate in the challenges through the gamified character of the experiment website. They obtain a language medal for every completed experiment, can view their medal collection

Proceedings of the LREC 2020 Workshop "Citizen Linguistics in Language Resource Development", , pages 40–48
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

and select experiments with other languages to participate in. A short statistic overview of the automatically classified correct answers together with the average response time is displayed at the end of each experiment. The participants have the opportunity to see their performance in different challenges in a visualization of their achievements on a timeline showing the individual completed experiments. They get an immediate feedback in which unknown but related language they have achieved better results. These intercomprehension scores reveal what is known as *inherent* intelligibility, i.e. based on structural linguistic similarities (Gooskens, 2019). What's more, our website can be used as an e-learning component of intercomprehension courses on Slavic languages offered at universities or elsewhere. To this effect, we provide an additional try-again functionality for already completed experiments. Thus, the students have the opportunity to repeat completed tasks once again towards the end of a course and to compare the initial results (inherent intelligibility) with the intercomprehension scores achieved after a focused teaching intervention, with the latter results revealing the so-called *acquired* intelligibility. An acquired *lingua receptiva* can apply to less related or unrelated languages, too (Muikku-Werner, 2014). And *mediated* receptive multilingualism (Branets et al., 2019) utilizing a bridge language can ease the understanding even between typologically distant languages, for example, when German participants with some training in Russian (RU) try to understand Bulgarian (BG) through RU in our experiments. In the following sections, we present only results of the inherent intelligibility for Slavic native speakers in an intercomprehension scenario. With regard to socio-demographic data, the participants are asked to specify their age, sex, level of education, linguistic repertoire, learning duration, assumed proficiency of (non)-native languages in written and spoken modality, place/country of residence, linguistic surroundings, etc. This information can be used for further analyses concerning the influence of extra-linguistic and socio-demographic factors on receptive multilingualism. After having completed the registration process, including the questionnaire, the participants are introduced to the challenge.

2.1 Intelligibility at the word level

This challenge is designed as a cognate guessing task. The participants are asked to translate randomized written and spoken stimuli into their native language. In the written condition, participants see the stimuli on their screen, one by one, and have 10 seconds to translate each stimulus. In the spoken condition, participants listen to the stimuli one by one with the task to provide a written translation within the same duration (10 seconds). In the spoken translation task, each word is played twice. The time limit is chosen based on the experience from other intercomprehension experiments, including, among others, a pilot study by Golubović (2016). The allocated time is supposed to be sufficient for typing even the longest words, but not long enough for using a dictionary or an online translation tool. It is possible to finish before the 10 seconds are over by either clicking on the 'Next' button or pressing 'Enter' on the keyboard. After 10 seconds, the participants hear or see the next stimulus on their screen. The order of stimuli presentation is randomized. The system saves everything that is entered, regardless of whether a participant confirms the translation by pressing the return key (or clicking 'Next') or not. The results are automatically categorized as 'correct' or 'wrong' via pattern matching with predefined correct answers and acceptable alternatives. An immediate feedback is given in the shape of an emoticon on the left at the bottom of the page – a thumb up for a successful translation or a sad face for a wrong or missing translation. There is a tolerance for lower/upper case and diacritical signs, i.e. if translations were entered without diacritics, but are otherwise correct, the participants get a positive feedback. The responses can then be checked manually for typographical errors in the final analysis.

2.2 Intelligibility at the phrasal level

This challenge is designed as a translation of noun and adjective sequences, with the adjective occurring pre- or post-nominally. For each stimulus phrase, the participants have 20 seconds for entering a translation into their language. The individual target words, together with the words directly preceding them, are extracted from the sentence stimuli in order to be also tested in their base forms (if applicable) at the word level.

2.3 Intelligibility at the sentence level

This challenge is designed as a cloze (fill-in-the-gap) translation task. The respondents see initially only the first word of the sentence. They are prompted to click on the word so that the next word in the sentence appears. After they have clicked through and consequently read the entire stimulus sentence in that way, a box appears at the position of the last word, which should be translated. This method ensures that participants read each sentence word by word. There are two separate time limits: one for clicking and reading through the sentence and one for entering the translation of the target word. The latter is automatically set by the system to 20-30 seconds, depending on the length of the sentence. The time limit for clicking and reading through the whole sentence is set to a maximum value of 300 seconds.

3. Methods for measuring intelligibility

In the INCOMSLAV framework, we developed measuring methods of immediate relevance to the concept of receptive multilingualism. Similarities between Slavic orthographies were captured by (modifications of) the Levenshtein metric (Levenshtein, 1966). Being frequently used as a predictor of phonetic and orthographic similarity (Beijering, Gooskens, and Heeringa, 2008; Gooskens, 2007; Vanhove, 2014), this mathematical distance is, however, completely symmetric. In order to account for the asymmetries of intercomprehension, additional measures of conditional entropy and surprisal (Shannon, 1948) were applied. Conditional character adaptation entropy and word adaptation surprisal (Mosbach et al., 2019; Stenger, 2019; Stenger et al., 2017) quantify the difficulties humans encounter when mapping one orthographic system on another and reveal the asymmetries in language pairs. Consider, for example, the language pairs Czech (CS) - Polish (PL) (West Slavic with Latin script) and BG-RU (South and East Slavic with Cyrillic script). While having similar lexical distances (share of non-cognates) of 10-15% depending on the direction, CS and PL are orthographically more distant from each other than BG and RU (for more details see Jágrová et al., 2017).

Our measures suggest that Czech readers should have more difficulties reading PL than vice versa, and that the asymmetry between BG and RU is very small with a minimal predicted advantage for Russian readers (Stenger et al., 2017). Furthermore, the word-length normalized adaptation surprisal appears to be a better predictor than the aggregated Levenshtein distance when the same stimuli sets in different language pairs are compared (Stenger, Avgustinova, and Marti, 2017). Previous research shows that additional factors such as word length, neighborhood density and word frequency play a significant role in spoken word recognition without context (Kürschner, van Bezooijen, and Gooskens, 2008). We also found (Stenger, 2019) that word length as an explanatory variable is essential in the recognition of written South Slavic (BG, Macedonian (MK), and Serbian (SR)) stimuli by Russian readers, since the South Slavic words are generally shorter than their RU and East Slavic (Ukrainian (UK) and Belarusian (BE)) cognates. Neighbors are linguistically defined as word forms that are very similar to the stimulus word and may therefore serve as competing responses (ibid.), for example the BG word *цел* (*cel*) 'target' with the correct RU translation *цель* (*cel'*) has two RU neighbors: *мел* (*mel*) 'chalk' and *цех* (*cech*) 'workshop', while the BG word *автомобил* (*avtomobil*) 'car' has no neighbors. BG and SR written intelligibility to Russian native speakers shows that the higher the neighborhood density, the lower is the number of successful translations, although this is not the case for UK, BE, and MK stimuli when presented to Russian readers. According to our experimental results, the frequency of cognates is not a reliable predictor for Russian readers. In reality, the orthographic and phonetic correspondences (their nature, position, and frequency) can considerably influence intercomprehension. Investigating Cyrillic script intelligibility to Russian readers, we saw that (i) identical orthographic correspondences increase intelligibility, while non-identical correspondences yield a barrier, and (ii) cognates are generally easier to understand if the beginning of the word is identical (ibid.). Until recently, the role of context in intercomprehension has been addressed in relatively few studies. In a monolingual situation, statistical language models (LMs) provide information about the predictability of words in context. Levy (2008) showed that n-gram LMs, specifically trigrams, performed well at predicting the processing effort measured by the reading times of variably difficult texts. In information theory, a commonly used unpredictability measure is surprisal. It can be thought of as a measure for the information conveyed by a linguistic unit and scales the cognitive effort required to process this information (Crocker, Demberg, and Teich 2016). The lower the surprisal, the more predictable a word is in a sentence, given its preceding words. Whenever there is a drop in surprisal after a word, the word with the lower surprisal should be highly predictable after its preceding word. We investigated the intelligibility of highly predictable target words in PL sentences presented to Czech readers (Jágrová et al., 2018), and saw that predictions based on surprisal scores do not always agree with the actually observed intercomprehension difficulty by humans. In order to study the role of predictive context and its correlation with intelligibility in the intercomprehension scenario quantitatively, we presented 149 PL target words both in highly predictive sentential context (cloze probability ≥90%, Block and Baldwin, 2010) and without context to Czech readers (Jágrová and Avgustinova, 2019). We found that surprisal had a significant correlation with target words that were non-cognates or false friends (there were 65.1% cognates, 11.4% non-cognates, and 23.5% false friends). During the disambiguation of these, readers did rely on context rather than on word similarity (ibid.).

4. Intercomprehension resources

Currently, we provide 162 online experiments (spoken and written individual word translation (40-60 words per spoken and written challenge), phrasal translation (30-35 phrases per challenge), and word translation in predictive context (10-20 sentences per challenge) for native speakers of 11 Slavic languages (BE, BG, CS, Croatian (HR), MK, PL, RU, SR, Slovak (SK), Slovenian, UK) as well as German and English. The designed experimental sets stem from a collection of parallel lists of internationalisms, Panslavic vocabulary, cognates from Swadesh lists[1], frequency lists of the respective languages (e.g. Křen (2010) for CS, Ljaševskaja and Šarov (2009) for RU) and resources from available corpora (InterCorp, Czech National Corpus, Russian National Corpus etc.).

About 2000 native speakers[2] participated in the challenges. The online available Slavic intercomprehension matrix (SlavMatrix)[3] contains currently obtained intelligibility scores and measures of linguistic distances and asymmetries for different language pairs and processing directions. Table 1 gives a high-level overview of the SlavMatrix.

Level	Sublevel
Intelligibility	Individual words: Automatic Panslavic vocabulary Top 100 Verbs
	Phrases (adjective-noun combinations)
	Words in predictive contexts
Predictors	Linguistic distances: Orthographic Lexical Phonetic Morphological Syntactic
	Conditional entropy
	Word adaptation surprisal (WAS)
Correlations	Intelligibility with Levenshtein distance
	Intelligibility with lexical distance
	Intelligibility with conditional entropy
	Intelligibility with word adaptation surprisal

Table 1: High-level overview of the SlavMatrix.

[1] Refer to Angelov (2004), Likomanova (2004), and Swadesh lists for Slavic languages, accessed on 2015-04-22.

[2] Status of 2020-03-02.

[3] http://intercomprehension.coli.uni-saarland.de/en/SlavMatrix/Results/

In Section 4.1 we discuss the level of intelligibility of individual words, in Section 4.2 we analyze the level of predictors, and in Section 4.3 we address the level of correlations.

4.1 SlavMatrix: individual words

The sublevel of individual words contains the following data: (i) automatically calculated experimental results, (ii) experimental results for the Panslavic vocabulary, (iii) experimental results for the 100 most frequent nouns (Top 100), and (iv) experimental results for verbs. The automatically calculated results cover all individual word translation tasks. Since reading and listening are different cognitive activities, we differentiate between the written and the spoken version of the tests and consider in the following the reading intelligibility only. Intelligibility scores are calculated for each of the above mentioned sublevels. The scores are converted to percentages by dividing the number of correct responses by the number of items in the test (and multiplying the result by 100). According to the automatically calculated experimental results, the highest scores were observed for Slovak participants reading CS (84.1%[4]), and for Croatian subjects reading SK (84.0%). As expected, Czech readers also understand SK at a high level (77.8%). Slovak readers understand HR at 68.0%. Here we have an asymmetry of 16.0% in favor of Croatian readers. The smallest intelligibility scores were observed for Slovak subjects reading UK (4.0%). This can be explained by the fact that SK is written with the Latin script and UK with the Cyrillic script. Thus, UK can generally only be understood by readers who know the Cyrillic script. Across the West Slavic languages with Latin script (PL, CS, and SK) and East Slavic languages with the Cyrillic script (BE, RU, and UK) the comprehensibility values are at a high level in both sub-groups, e.g. participants of East Slavic languages managed to translate more than 74% of the words correctly and readers of West Slavic languages reached almost 68%. All these percentages are intelligibility scores based on answers that were automatically classified as correct by the website.

For more precise and representative data, we have considered the sublevel of experimental results for Panslavic vocabulary that has been checked manually in the final analysis. The stimuli are cognates (etymologically related words) containing historical cross-lingual orthographic correspondences, e.g. BG–RU: *б*:*бл*, *ж*:*жд*, *ла*:*оло*, *я*:*е* etc. (for more details see Fischer et al., 2015; Fischer et al. 2016). The initial hypothesis was that correct cognate recognition would be the key to successful intercomprehension. The experimental results show in particular that among the East Slavic languages UK is more understandable to Russian readers than BE. The average comprehensibility values for UK and BE stimuli are relatively high – almost 86% and 73% respectively. Among the three South Slavic languages, BG is the most understandable one for Russian readers, with an average comprehensibility value of approx. 71%, followed by MK with 62% and SR with almost 59%. Thus, we can state for Russian readers[5] that, on average, a successful cross-lingual recognition of individual East and South Slavic cognates is generally registered here. Concerning the language pair BG and RU, the results show that there is virtually no asymmetry in written intelligibility between these languages: the Bulgarian participants understand a slightly larger number of the 120 RU words (74.67%) than the Russian participants understand the 120 BG words they are presented with (71.33%)[6]. This can be explained by the fact that there are only slight differences between the two languages on the graphic-orthographical level (for more details see Stenger et al., 2017).

4.2 SlavMatrix: predictors

Two measurement methods provide predictions of mutual intelligibility between (closely) related languages: Levenshtein distance (LD, here as orthographic string edit distance) and word adaptation surprisal (WAS) (see Table 1). LD is, in its basic implementation, a symmetric similarity measure between two strings, in our case between written words. It quantifies the number of operations in order to transform one word into another. When computing LD for a pair of words, three different character transformations are considered: deletion, insertion, and substitution. These operations are assigned weights. In the simplest form of the algorithm, all operations have the same cost. We use 0 for the cost of mapping a character to itself, e.g. *a*:*a*, and a cost of 1 to align it to a character of the same kind (vowel characters vs. consonant characters), e.g. *a*:*o*. All vowel-to-consonant combinations are given a weight of 4.5 (most expensive) in the algorithm. Thus, we obtain distances which are based on linguistically motivated alignments. In more sensitive versions, a base and a diacritic may be distinguished. For example, the base of *ë* is *e*, and the diacritic is the diaeresis. Even though it is not exactly clear what weight should be attributed to each of the components, it is generally assumed that differences in the base will usually confuse the reader to a much greater extent than diacritical differences. If two characters have the same base but differ in diacritics, we assign them a substitution cost of 0.5 (for more details s. Mosbach et al., 2019). In our analysis we consider normalized LD (nLD) in accordance with the assumption that a segmental difference in a word of, e.g., two segments has a stronger impact on intelligibility than a segmental difference in a word of, e.g. ten segments (Beijering, Gooskens, and Heeringa, 2008). The nLD of BG–RU: *език–язык* (*ezik–jazyk*) 'tongue/language' is 2/4=0.5 or 50%. Measuring the orthographic distance on the basis of the Levenshtein

[4] This value cannot be compared to the intelligibility scores for cognate lists in the other language pairs, since the stimuli sets for CS-SK included non-cognates. The intelligibility score for CS-SK cognates might in fact be higher.

[5] 119 Russian native speakers took part in the experiments with 340 East and South Slavic stimuli, the mean age of the participants was 34 years, ¾ women and ¼ men. We only analyzed answers from participants who indicated that they did not know the stimulus language and only of the initial challenge for each participant in order to avoid any learning effects (for more details see Stenger, 2019).

[6] The analysis of the collected material is based on the answers of 37 native speakers of BG (31 women and 6 men, mean age 27 years) and 40 native speakers of RU (32 women and 8 men, mean age 33 years) of the initial challenge. All participants have indicated that they did not know the stimulus language (for more details see Mosbach et al., 2019).

algorithm allows us to model the mutual intelligibility based on the following hypothesis: The larger the distance, the more difficult it is to comprehend an unknown language. Displaying a more generalized view of modelling mutual intelligibility among Slavic languages, the nLD matrix (Table 2) shows aggregated orthographic distances (in percentages) between East and South Slavic languages on 190 cognate pairs of Common Slavic vocabulary, published in (Carlton, 1991) (for more details on the used material see Stenger, 2019).

stimulus language	native language					
	BE	BG	MK	RU	SR	UK
BE	0	40.66	41.11	27.23	41.98	36.56
BG	40.66	0	17.04	32.05	24.89	35.52
MK	41.11	17.04	0	32.19	19.37	36.37
RU	27.23	32.05	32.19	0	32.09	22.77
SR	41.98	24.89	19.37	32.09	0	33.03
UK	36.56	35.52	36.37	22.77	33.03	0

Table 2: Aggregated nLD as predictor of mutual intelligibility among BE, BG, MK, RU, SR, and UK.

In general, the average symmetrical Levenshtein distance values of the 15 analyzed East and South Slavic language pairs are below 42%, which indicates a relatively high orthographic similarity between these languages (all using Cyrillic) and, hence, mutual intelligibility on the orthographic level. According to the nLD matrix, mean normalized orthographic distances between South Slavic languages are smaller than between East Slavic languages, which leads to the assumption that readers of a South Slavic language may be better able to understand cognates in written texts of in another South Slavic language than East Slavic readers who are confronted with a written text in another East Slavic language. Furthermore BG and MK are the closest language pair in the South Slavic subgroup, since they get the smallest symmetric orthographic distance (17.04%). As already pointed out, a disadvantage of this string-edit method is that the LD cannot show any asymmetries depending on the processing direction in a given language pair. Given two aligned words, we can also compute for them the word adaptation surprisal (WAS), which, intuitively, measures how confused a reader would be trying to map a character of the stimulus word to a character of the target word. In order to define WAS we introduce the notation of character adaptation surprisal (CAS) which is defined as follows:

$$\mathrm{CAS}(L1 = c1 | L2 = c2) = -\log_2 P(L1 = c1 | L2 = c2)$$

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

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

Now, WAS between two words is computed by summing up the CAS values of the contained characters in the aligned word pair (for more details see Mosbach et al., 2019; Stenger 2019). Note that in contrast to LD, CAS and WAS are not symmetric. Moreover, the WAS highly depends on the number of available word pairs. Computing CAS (and therefore also WAS) depends on the conditional probability P, which is based on corpus statistics of the aligned word pairs by means of the Levenshtein algorithm. For example, the RU character *а* (which occurs 175 times) corresponds exclusively to the BG character *а* (which occurs 194 times). The BG character *а* may correspond to the RU character *а* (175 times), *о* (15 times) or *я* (4 times) (these examples are based on the 291 cognate pairs, for more details see Stenger et al., 2020). Thus, for our example above, we would get P(BG = а | RU = а) = 175/175 = 1.0, while P(RU=а | BG = а) = 175/194 ≈ 0.9, P(RU = о | BG = а) = 15/194 ≈ 0.07, and P(RU = я | BG = а) = 4/194 ≈ 0.02. In such a case, we can expect a Russian reader to have more difficulties to correctly guess which characters in RU correspond to the BG one he/she is confronted with. As in the case with the LD, we normalized the WAS and calculated the average value of the normalized WAS (nWAS) for 190 cognate pairs of the Common Slavic vocabulary (Carlton, 1991). The nWAS matrix (Table 3) displays the mean nWAS (in bits) between selected languages reflecting the asymmetry and complexity of the mapping of one orthographic system on another, based on the following assumption: The higher the mean nWAS, the more difficult it is to comprehend the unknown language. According to the nWAS matrix, BG and MK are not only the closest language pair in the South Slavic sub-group, but there is an orthographic asymmetry between BG and MK in favor of MK. The mean nWAS gives us the following values: 0.66 bits for Bulgarian readers of MK and 0.49 bits for Macedonian readers of BG, thus predicting that a Bulgarian reader may have more difficulties reading MK than vice versa.

stimulus language	native language					
	BE	BG	MK	RU	SR	UK
BE	0	1.18	1.12	0.69	1.09	0.80
BG	1.39	0	0.49	1.18	0.82	1.36
MK	1.50	0.64	0	1.28	0.82	1.46
RU	0.72	0.98	0.90	0	0.87	0.68
SR	1.36	0.87	0.72	1.13	0	1.23
UK	0.79	1.16	1.09	0.66	0.99	0

Table 3: Mean nWAS as predictor of mutual intelligibility among BE, BG, MK, RU, SR, and UK.

4.3 SlavMatrix: correlations

Normalized LDs were calculated for all word pairs of the respective experimental tasks in order to correlate the orthographic distance with the human intelligibility scores. For example, in the Cyrillic script intelligibility tests for Russian native speakers, mentioned in Section 4.1, the negative correlations were statistically significant for all analyzed language pairs: BE–RU (r = –0.509, p = 3.17e-05), BG–RU (r = –0.566, p = 1.47e-11), MK–RU (r = –0.305, $p < 0.05$), SR–RU (r = –0.659, p = 1.87e-07), UK–RU (r = –0.456, $p < 0.0005$), although they could be classified as low to medium. The highest negative correlation is characteristic for the SR–RU language pair. In other words, the initial hypothesis that small orthographic distances between two cognates correlate with high intelligibility values – and large orthographic distances with low intelligibility values – can be considered confirmed. In addition, we also calculated the nWAS for each cognate pair of the above mentioned tests. The significant negative correlation was recorded only for the UK–RU language pair (r = –0.491, p = 6.67e-05), suggesting that the complexity of a mapping between two cognates measured by the nWAS method plays the most important role in the recognition of individual cognates for the UK–RU language pair.

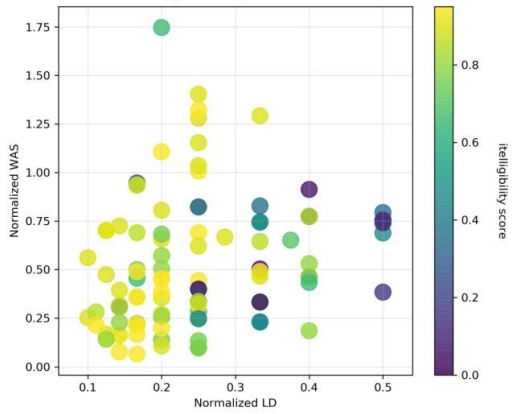

Figure 1: Intelligibility score depending on normalized LD and normalized WAS, BG for Russian readers

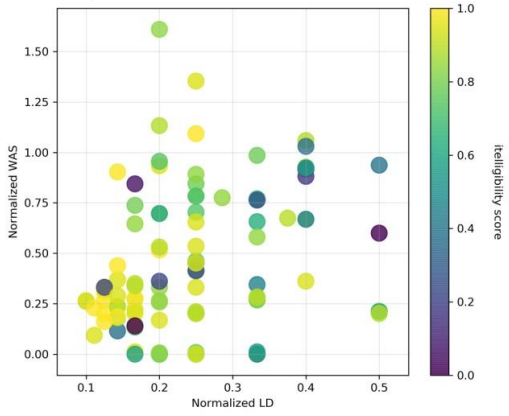

Figure 2: Intelligibility score depending on normalized LD and normalized WAS, RU for Bulgarian readers

For the other three language pairs the negative correlations were not significant: BG–RU (r = –0.135, p = 0.142), MK–RU (r = –0.131, p = 0.364), and SR–RU (r = –0.270, p = 0.058). For the fifth language pair BE–RU, the calculated correlation was even slightly positive (r = 0.196, not significant p = 0.134), which speaks against the initial hypothesis (for more details see Stenger, 2019). The question is why the correlation at the cognate level is so low and insignificant for three language pairs (with the BE–RU language pair representing an outlier with regard to the formulated hypothesis). Intuitively, it seems plausible that a stimulus word is easier to understand if it is more similar to a cognate in the target language. So, a possible explanation could be that identical characters can have a CAS value on the basis of the nWAS method, which automatically increases the total nWAS value. A modified nWAS method (described in Mosbach et al., 2019 and in Stenger, 2019) allows us to consider CAS values for all identical characters with costs of 0 in a manual post-processing step. After the modification of the nWAS method, a negative correlation between the modified nWAS and the test results was found for all language pairs: BE–RU (r = –0.035), BG–RU (r = –0.210), MK–RU (r = –0.155), SR–RU (r = –0.396), UK–RU (r = –0.555). However, the examination of the statistical results for their significance showed that the negative correlations were only for three language pairs at a significant level: BG–RU (p < 0.05), SR–RU (p < 0.005), and UK–RU (p = 4.156e-06) (for more details see Stenger, 2019). As already mentioned in Section 1.2, the intercomprehension should be better, when the language model adapted for understanding the unknown language exhibits relatively low average distance and surprisal. Concerning the mutual intelligibility between BG and RU (described in Section 4.1) the nLD and nWAS account for 32% (R^2 = 0.32) of the variance in the intelligibility scores for Russian readers and for only 14% (R^2 = 0.14) of the variance in the intelligibility scores for Bulgarian readers, which leaves the majority of variance unexplained (see Figures 1 and 2). Note that the calculated mean nLD and nWAS data are based here on a small experimental corpus. There are a number of arguments why distance measurements should be calculated not on the basis of the experimental material, but on the basis of larger amounts of data. In particular, distance measurements become more stable and correlate better with mutual intelligibility when calculated on larger data (van Heuven, Gooskens, and van Bezooijen, 2015). This relationship may be different if the distance measurements are specifically based on the experimental material used in the intelligibility test (ibid.). The CAS values are different and depend on the respective cognate lists. If the scope of the cognate list is extended with further pairs, the CAS values may change, which would lead to a change in the nWAS values, too. In the web-based experiments, subjects are confronted with a limited amount of data. Therefore, the regularity of one or the other correspondence from the cognate lists of the experimental material does not necessarily correspond to the one observed in the respective correspondences from a larger corpus. We measured nLD and nWAS values on the experimental material and correlated them with the intelligibility values from the web-based experiments, namely, the intelligibility scores based on the initial challenge for each participant in order to avoid any learning effects (see Section 4.1). The WAS values between language A and language B are not necessarily the same as between language B and language A, which indicates an advantage of the surprisal-based method compared to LD in modelling asymmetry. We calculated the mean nWAS for BG and RU using a cognate word list from the intelligibility tests (see Section 4.1). For the BG–RU language pair the difference in the mean nWAS is very small: 0.46 bits for the RU to BG transformation and 0.50 bits for the BG to RU transformation, with a very small amount of asymmetry of 0.04 bits. These results predict that speakers of RU reading BG words are more uncertain than speakers of BG reading RU words. This is in accordance with the experimental results where the language combination with the slightly higher mean nWAS (speakers of RU reading BG words) had a slightly lower intelligibility score (see Section 4.1).

5. Discussion and Future Work

In this paper we presented the INCOMSLAV platform as a web-based resource for conducting intercomprehension experiments with native speakers of Slavic languages, and illustrated our methods for measuring linguistic distances and asymmetries in receptive multilingualism. All ob-

tained intelligibility scores as well as distance and asymmetry measures are made available as an integrated online resource in the form of a Slavic intercomprehension matrix (SlavMatrix), which will be maintained and further completed as new data and correlations become available.

Among presented intelligibility tests we discussed here automatically calculated experimental results of individual words as well as manually checked experimental results for a Panslavic vocabulary. Even though it may seem artificial to test individual words without context, since the latter may provide helpful information, our underlying assumption is that the cognate recognition is a precondition of success in reading intercomprehension. If the reader correctly recognizes a minimal proportion of words, he or she will be able to piece the written message together. An important practical criterion for choosing a test is the ease with which it can be developed, administered and analyzed. If more languages should be tested, extensive time and effort would be needed to collect a large number of participants. Since we have the most completed experiments in different language combinations for the word level, we decided to focus here on the individual word translation tasks. We need to collect and further analyze the experimental results at the phrasal and sentence levels, too. Recently, the INCOMSLAV platform also provides the LADO experiments (Language Analysis for Determination of Origin) and collects experimental data evaluating in fact the listening interpretation ability of the participants not only in foreign languages, but also in their own language, for example, recognition of RU segments (LADO 1) and prosody (LADO 2) among Russian native speakers.

Related research has already shown that *inherent* intelligibility can be predicted quite well by linguistic distance and that a short word list provides sufficient input for computing the distance measures needed (Gooskens and van Heuven, 2019). Therefore it may be an option to rely on distance measurements rather than on costly functional testing in order to investigate how well speakers of closely related languages will be able to understand each other (ibid). We presented two measurements of linguistic distance and asymmetry as potential predictors of mutual intelligibility between (closely) related languages: normalized Levenshtein distance (nLD) as orthographic distance and normalized word adaptation surprisal (nWAS) as orthographic asymmetry between Slavic languages. As already discussed in Section 3, the mean nWAS at the language level appears to be a better predictor than the aggregated nLD when the same stimuli sets in different language pairs are compared (Stenger, Avgustinova, and Marti, 2017). In this contribution we were also able to show that the mean nWAS can be a reliable measure when explaining small asymmetries in intelligibility between BG and RU (see Section 4.3). However, at the cognate level, the nLD correlates better with the experimental results as nWAS. As other inter-comprehension research shows, each pair of cognates has its own constellation of factors that influence intelligibility, whereby one factor can overlay another (Kürschner, van Bezooijen, and Gooskens, 2008). In addition, factors and corresponding models are language-dependent, as each language combination poses different challenges to the readers. In summary, this means that each model has its limits and there is room for improvement by taking into account the influence of additional factors, for example, neighborhood density (the number of word forms that are similar to the stimulus word), the effects of character context, within-word position, consonants vs. vowels, dialects or archaic terms etc.

Our resources, including *incom.py*[7] – a toolbox for calculating linguistic distances and asymmetries between related languages, can be of interest to other researchers working on intercomprehension and to teachers of multilingual language courses. In the next phase, we plan to extend the *SlavMatrix* resources by an *IncomSlavCorpus*, providing researches of receptive multilingualism with the experimental material used in our tests and with all correlated intercomprehension results. In addition to structural characteristics of the languages a broader approach will include extra-linguistic factors (e.g. language exposure) and individual factors (e.g. age, linguistic repertoire, language learning experience, education level) that contribute to understanding unknown but related languages.

6. Acknowledgements

We wish to thank Hasan Alam for his support in the implementation of the SlavMatrix. This work has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102.

7. Bibliographical References

Angelov, A. (2004). EuroComSlav Basiskurs – der panslavische Wortschatz. http://www.eurocomslav.de/BIN/inhalt.htm, accessed 2016-02-17.

Beijering, K., Gooskens, C. and Heeringa, W. (2008). Predicting intelligibility and perceived linguistic distance by means of the Levenshtein algorithm. In M. van Koppen & B. Botma (Eds.), *Linguistics in the Netherlands* 2008. John Benjamins, Amsterdam pp. 13–24.

Berruto, G. (2004). Sprachvarietät – Sprache (Gesamtsprache, historische Sprache). In U. Ammon et al. (Eds.), *Soziolinguistik*. *Ein internationales Handbuch zur Wissenschaft von Sprache und Gesellschaft*, 1. Teilband. Walter de Gruyter, Berlin, New York, pp. 188–195.

Branets, A., Bahtina, D., and Anna Verschik. (2019). Mediated receptive multilingualism: Estonian-Russian-Ukrainian case study. *Linguistic Approaches to Bilingualism*: 1–32.

Braunmüller, K. and Zeevaert, L. (2001). Semikommunikation, rezeptive Mehrsprachigkeit und verwandte Phänomene. Eine bibliographische Bestandsaufnahme, Arbeiten zur Mehrsprachigkeit, Folge B, 19, Universität Hamburg, Hamburg.

Block, C. K. and Baldwin, C. L. (2010). Cloze probability and completion norms for 498 sentences: Behavioral and neural validation using event-related potentials. *Behavior Research Methods* 42(3): 665–670.

Carlton, T. R. (1991). Introduction to the phonological history of the Slavic languages. Slavica Publishers, Inc, Columbus, Ohio.

[7] The licence of incom.py is freely available: https://github.com/uds-lsv/incompy.

Comrie, B. and Corbett, G. G. (1993). Introduction. In B. Comrie & G. G. Corbett (Eds.), *The Slavonic Languages*. Routledge, London/New York pp. 1–20.

Crocker, M., Demberg, V., and Teich, E. (2016). Information Density and Linguistic Encoding (IDeaL), *Künstliche Intelligenz* 30: 77–81.

Doyé, P. (2005). Intercomprehension. Guide for the development of language education policies in Europe: from linguistic diversity to plurilingual education. Reference study, Strasbourg, DG IV, Council of Europe.

Fischer, A., Jágrová, K., Stenger, I., Avgustinova, T., Klakow, D., and Marti, R. (2015). An orthography transformation experiment with Czech–Polish and Bulgarian–Russian parallel word sets. In B. Sharp, W. Lubaszewski & R. Delmonte, editors, Natural Language Processing and Cognitive Science 2015 Proceedings, pages 115–126, Libreria Editrice Cafoscarina, Venezia.

Fischer, A., Jágrová, K., Stenger, I., Avgustinova, T., Klakow, D., and Marti, R. (2016). Orthographic and Morphological Correspondences between Related Slavic Languages as a Base for Modeling of Mutual Intelligibility, Proceedings Language Resources and Evaluation Conference (LREC), pages 4202–4209, Portorož.

Golubović, J. (2016). Mutual intelligibility in the Slavic language area. PhD thesis. University of Groningen (Netherlands).

Gooskens, C. (2019). Receptive multilingualism. *Multidisciplinary perspectives on multilingualism: The fundamentals* LCB 19: 149–174.

Gooskens, C. (2007). The contribution of linguistic factors to the intelligibility of closely related languages. *Journal of Multilingual and Multicultural Development* 28(6): 445–467.

Gooskens, C. and van Heuven, V. J. (2019). How well can intelligibility of closely related languages in Europe be predicted by linguistic and non-linguistic variables? *Linguistic Approaches to Bilingualism*: 1–29.

Gooskens, C. and Swarte, F. (2017). Linguistic and extralinguistic predictors of mutual intelligibility between Germanic languages. *Nordic Journal of Linguistics* 40(2): 123–147.

Haugen, E. (1966). Semicommunication: The language gap in Scandinavia. *Sociological Inquiry* 36: 280–297.

van Heuven, V. J., Gooskens, C., and van Bezooijen, R. (2015). Introduction Micrela: Predicting mutual intelligibility between closely related languages in Europe. In: J. Navracsics & S. Batyi (Eds.), *First and Second Language: Interdisciplinary Approaches* (Studies in Psycholinguistics (6)), Tinta konyvkiado, Budapest, pp. 127–145.

Jágrová, K. and Avgustinova, T. (2019). Intelligibility of highly predictable Polish target words in sentences presented to Czech readers. To appear in Proceedings of CICLing: International Conference on Intelligent Text Processing and Computational Linguistics.

Jágrová, K., Avgustinova, T., Stenger, I., and Fischer, A. (2018). Language Models, Surprisal and Fantasy in Slavic Intercomprehension. In R. K. Moore, P. Fung & S. Narayanan (Eds.), *Computer Speech and Language* 53: 242–275.

Jágrová, K., Stenger, I., Marti, R., and Avgustinova, T. (2017). Lexical and orthographic distances between Bulgarian, Czech, Polish, and Russian: A comparative analysis of the most frequent nouns. In J. Emonds & M. Janebová (Eds.), Language Use and Linguistic Structure. Proceedings of the Olomouc Linguistics Colloquium 2016, pages 401–416, Palacký University, Olomouc.

Křen, M. (2010). Srovnávací frekvenční seznamy [Comparative frequency lists]. Prague: Institute of the Czech National Corpus Faculty of Arts, Charles University Prague. http://ucnk.ff.cuni.cz/index.php, accessed 2016-09-11.

Kürschner, S., van Bezooijen, R. and Gooskens, C. (2008). Linguistic determinants of the intelligibility of Swedish words among Danes. *International Journal of Humanities and Arts Computing* 2(1/2): 83–100.

Levenshtein, V. I. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady* 10(8): 707–710.

Levy, R. (2008). Expectation-Based Syntactic Comprehension. *Cognition* 106(3): 1126–1177.

Likomanova, I. (2004). EuroComSlav Basiskurs – der internationale Wortschatz. http://www.eurocomslav.de/kurs/iwslav.htm, accessed 2016-02-17.

Ljaševskaja, O. N. and Šarov, S.A. (2009). Častotnyj slovar' sovremennogo russkogo jazyka [Frequency dictionary of the contemporary Russian language]. Moskva: Azbukovnik.

Mosbach, M., Stenger, I., Avgustinova T. and Klakow, D. (2019). incom.py – A Toolbox for Calculating Linguistic Distances and Asymmetries between Related Languages. In: G. Angelova, R. Mitkov, I. Nikolova & I. Temnikova, editors, Proceedings of Recent Advances in Natural Languages Processing (RANLP 2019), pages 811–819, Varna, Bulgaria.

Muikku-Werner, P. (2014). Co-text and receptive multilingualism Finnish students comprehending Estonian. *Journal of Estonian and Finno-Ugric Linguistics* 5(3): 99–103.

Ringbom, H. (2007). Cross-linguistic similarity in foreign language learning. Multilingual Matters LTD, Clevedon.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal* 27: (379–423), 623–656.

Stenger, I. (2019). Zur Rolle der Orthographie in der slavischen Interkomprehension mit besonderem Fokus auf die kyrillische Schrift. Dissertation. Universaar, Saarbrücken.

Stenger, I., Avgustinova, T., and Marti, R. (2017). Levenshtein distance and word adaptation surprisal as methods of measuring mutual intelligibility in reading comprehension of Slavic languages. Computational Linguistics and Intellectual Technologies: International Conference 'Dialogue 2017' Proceedings. Issue 16(23), vol. 1, pp. 304–317.

Stenger, I., Jágrová, K., Fischer, A., and Avgustinova, T. (2020). "Reading Polish with Czech Eyes" or "How Russian Can a Bulgarian Text Be?": Orthographic Differences as an Experimental Variable in Slavic Intercomprehension. In T. Radeva-Bork and P. Kosta (Eds.), *Current developments in Slavic Linguistics. Twenty years after. (based on selected papers from FDSL 11)*, Peter Lang, Bern, pp. 483–500.

Stenger, I., Jágrová, K., Fischer, A., Avgustinova, T., Klakow, D. and Marti, R. (2017). Modeling the impact of orthographic coding on Czech–Polish and Bulgarian–Russian reading intercomprehension. *Nordic Journal of Linguistics* 40(2): 175–199.

Vanhove, J. (2014). Receptive multilingualism across the lifespan. Cognitive and linguistic factors in cognate guessing. PhD thesis. University of Fribourg (Switzerland).

8. Language Resource References

incom.py – A toolbox for calculating linguistic distances and asymmetries between related languages. SFB 1102 – projects B4 and C4, available at: https://github.com/uds-lsv/incompy.

Intercomprehension Website (2014–2019). SFB 1102 – project C4 INCOMSLAV, availabe at: http://intercomprehension.coli.uni-saarland.de/de/.

Slavic Intecomprehension Matrix (2019). SFB 1102 – Project C4 INCOMSLAV, available at: http://intercomprehension.coli.uni-saarland.de/de/SlavMatrix/Results/.

Slavic Swadesh lists, https://en.wiktionary.org/wiki/Appendix:Slavic_Swadesh_lists, accessed on 2015-04-22.