Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 7368–7376
Marseille, 20-25 June 2022

Modeling the Impact of Syntactic Distance and Surprisal on Cross-Slavic Text Comprehension

Irina Stenger, Philip Georgis, Tania Avgustinova, Bernd Möbius, Dietrich Klakow
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, {pgeorgis, avgustinova, moebius}@lst.uni-saarland.de,
dietrich.klakow@lsv.uni-saarland.de

Abstract

We focus on the syntactic variation and measure syntactic distances between nine Slavic languages (Belarusian, Bulgarian, Croatian, Czech, Polish, Slovak, Slovene, Russian, and Ukrainian) using symmetric measures of insertion, deletion and movement of syntactic units in the parallel sentences of the fable "The North Wind and the Sun". Additionally, we investigate phonetic and orthographic asymmetries between selected languages by means of the information theoretical notion of surprisal. Syntactic distance and surprisal are, thus, considered as potential predictors of mutual intelligibility between related languages. In spoken and written cloze test experiments for Slavic native speakers, the presented predictors will be validated as to whether variations in syntax lead to a slower or impeded intercomprehension of Slavic texts.

Keywords: syntactic distance, surprisal, asymmetric intelligibility, Slavic intercomprehension

1. Introduction

1.1 Background

Intercomprehension (Doyé, 2005), semi-communication (Haugen, 1966), lingua receptiva (Rehbein et al., 2012), mutual intelligibility (Gooskens and van Heuven, 2021) or receptive multilingualism (Braunmüller and Zeevaert, 2001) refers to situations where people take advantage of similarities between their native language (L1) and a closely related non-native language (L2) to comprehend speech or text in this L2. Certain degrees of mutual intelligibility are typically due to existing and perceived similarities at different levels of linguistic structure. Investigations of (dis)similarities largely consider lexical, phonetic/phonological, and orthographic distances to predict the ability to comprehend (closely) related languages (Gooskens and van Heuven, 2021). Jágrová et al. (2017) have found that despite small lexical distances measured in terms of the proportion of non-cognates, Czech and Polish (both West Slavic, using Latin script) are orthographically more distant from each other than Bulgarian and Russian (respectively South and East Slavic, both using Cyrillic script). Stenger (2019: 245) shows the Levenshtein distance (Levenshtein, 1966) of cognates to be a reliable predictor of orthographic intelligibility of Slavic languages written in Cyrillic script to Russian native speakers. The role of syntactic (dis)similarities for Slavic intercomprehension is investigated to a considerably lesser extent. Golubović (2016: 105) uses the trigram approach (Nerbonne and Wiersma, 2006; Lauttamus, Nerbonne, and Wiersma, 2007) to illustrate the role of syntactic distance as a predictor of mutual intelligibility in written and spoken cloze tests. Stenger and Avgustinova (2021) adapt measures of word movement, insertion, and deletion from (Heeringa et al., 2017) to investigate the impact of syntactic predictors on contextualized cognate recognition in Slavic. The number of insertions or deletions to obtain the closest wording in subjects' language reveals a negative effect on intelligibility only as a tendency. The movement measure, however, did not influence the results, which contrasts with the significant effect found in (Swarte, 2016) among Germanic languages. According to Stenger and Avgustinova (2021), the possible explanation as to why the movement distance does not explain successful intelligibility of Slavic cognates in context is the greater word order flexibility in Slavic languages. However, the experimental design of these two studies is quite different: Stenger and Avgustinova (2021) tested reading comprehension of cognates in sentences, while Swarte (2016) investigated text intelligibility by means of written and spoken cloze test experiments.

1.2 Contributions of this Paper

We look at (dis)similarities among nine Slavic languages: Belarusian (BE), Bulgarian (BG), Croatian (HR), Czech (CZ), Polish (PL), Slovak (SK), Slovene (SL), Russian (RU), and Ukrainian (UK) with the goal of predicting the impact of syntactic variation on cross-Slavic comprehension of spoken and written texts. For a pair of languages, the difficulty of comprehending a sentence in one language for a speaker of the other depends on the number of *syntactic units* (single words or multi-component equivalents in cross-lingual alignments) to be added or deleted between corresponding sentences or fragments. We apply the first method of Heeringa et al. (2017), the *indel distance* (InDel), which measures the average number of words which are inserted or deleted in parallel sentences. The focus is on linearization differences that have an impact on text intelligibility. Word order variation is accounted for by means of the *movement distance* (Movement), according to the second approach of Heeringa et al. (2017), which measures the average number of words that must be reordered in sentences of one language in order to produce the word order of an equivalent sentence in another language (*movement binary*). Additionally, we consider *linear movement* (Heeringa et al., 2017) which measures the number of word positions a word from a sentence in one language has moved compared to the corresponding word in an equivalent sentence in another language. The assumption is that the more positions a syntactic unit must be moved and the more syntactic units that need to be

added or deleted, the more negative the effect on intercomprehension tends to be. We employ the information theoretical notion of *surprisal* (Shannon, 1948), in particular, *adaptation surprisal*, to measure linguistic asymmetries between related languages (Mosbach et al., 2019, 2021; Stenger et al., 2017, 2020). Adaptation surprisal reveals the complexity of mapping one phonetic or orthographic system onto another, i.e., how predictable the phonetic or orthographic correspondences are in a given language pair. The surprisal values of phonetic and orthographic correspondences are different and depend on the frequencies and distribution of correspondences in the measured material (Stenger, Avgustinova, and Marti, 2017).

Adaptation surprisal values can be asymmetrical, which is advantageous compared to symmetrical syntactic distances (via InDel and Movement). With larger phonetic and orthographic adaptation surprisal, more intercomprehension difficulties in the spoken and written modalities are expected. The research questions are: RQ1: How syntactically distant are these nine Slavic languages from each other? RQ2: What asymmetries are predictable by means of adaptation surprisal between selected languages from phonetic and orthographic viewpoints? RQ3: What is the relation among the measures under study here? After presenting the study material in Section 2, we introduce relevant methods of measuring syntactic (dis)similarities that may influence and explain the cross-lingual text intelligibility in Section 3. The statistical results are discussed in Section 4, before we draw some general conclusions in Section 5.

2. Material

For this study, we use the fable "The North Wind and the Sun", which is available in nine Slavic languages: Belarusian, Bulgarian, Croatian, Czech, Polish, Slovak, Slovene, Russian, and Ukrainian at the International Phonetic Association (1999) for illustration purposes.[1] Since the Croatian translation differs significantly in content from the other versions of the fable, we include an alternative version from the Aesop Language Bank.[2] The chosen text is well-known in both phonetic and experimental research, including studies of receptive multilingualism (e.g., Feleke, Gooskens, and Rabanus, 2020; Tang and van Heuven, 2007; Beijering, Gooskens, and Heeringa, 2008; Gooskens and Heeringa, 2004). Within the INCOMSLAV project[3], written and spoken cloze tests are developed on the basis of this fable in order to validate syntactic distance and surprisal as potential predictors of cross-Slavic textual intelligibility with empirical intelligibility scores obtained from web-based intercomprehension experiments. The nine parallel texts, containing approximately 90 tokens on average, are used to calculate the syntactic distances (Section 3.1) and the orthographic adaptation surprisal (Section 3.2.1). To calculate the phonetic adaptation surprisal we use existing IPA transcriptions. Although the East Slavic languages include both broad and narrow phonetic transcriptions of the text, we only consider the broad transcription. An extra phonetic transcription has been produced for the Croatian version, checked for accuracy by a native speaker phonetician (see Section 3.2.2).

3. Measuring Methods

Our approach includes InDel and Movement syntactic distances as well as orthographic and phonetic adaptation surprisal.[4]

3.1 Syntactic Distances

The syntactic distances used in the present study have been adapted from those proposed by Heeringa et al. (2017) for Germanic texts. Although literal translations are preferable for capturing genuine syntactic differences by avoiding stylistic or content deviations between translations, we have chosen the standard forms of the sentences from the original IPA resources. The parallel translations of the story are split into seven distinct fragments, and each fragment is manually aligned in a multiple alignment scheme, ensuring that all corresponding syntactic units within a fragment match with one another. Words or phrases without any equivalent in other translations align to a gap, reflecting respective deletions in that language. Multi-word expressions, such as the Russian *в конце концов* (*v konce koncov*) [5] 'finally' (lit: 'in the end of ends') or the Polish *w ten sposób* 'so, thus' (lit: 'in this way'), are kept together as single multi-component units in order to facilitate the alignment and reflect their correspondence to single words in other translations, e.g., Czech *konečně* 'finally' or Bulgarian *така* (*taka*) 'so, thus'. Consider the Russian-Bulgarian of the final fragment "And so the North Wind was obliged to confess that the Sun was the stronger of the two" below.[6]

RU		Таким образом				северный ветер	
		Takim obrazom				*severnyj veter*	
BG	И	така				северният вятър	
	I	*taka*				*severnât vâtăr*	

RU		вынужден		был		признать			
		vynužden		*byl*		*priznat'*			
BG	беше	принуден			да	признае			
	beše	*prinuden*			*da*	*priznae*			

RU	что	солнце			сильнее			его	
	čto	*solnce*			*sil'nee*			*ego*	
BG	че	слънцето	е		по-силно	от		него	
	če	*slănceto*	*e*		*po-silno*	*ot*		*nego*	

3.1.1 InDel

The InDel distance describes the number of syntactic units which must be inserted or deleted from a sentence in one language in order to transform it into its equivalent translation in another language. The normalized form of the InDel distance involves dividing this number by the total length of the aligned sentences, as measured by the

[1] https://richardbeare.github.io/marijatabain/ipa_illustrations_all.html

[2] https://www.aesoplanguagebank.com/hr.html

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

[4] The code and the material used for computing syntactic distances and adaptation surprisal are available online https://github.com/slavic-lab/LREC-2022-SynDist-Surprisal

[5] Transliterations according to ISO 9 Latin, *cf.* https://de.wikipedia.org/wiki/ISO_9

[6] Corresponding units appearing in different sentence positions are highlighted in green, deleted words are marked in red.

number of alignment positions. As a refinement to this basic method, a comparison according to part-of-speech (POS) was added to account for cases where aligned units belonged to different syntactic categories. The parallel translations were manually annotated for POS using the Universal Dependencies standard POS tag set.[7] Zero points toward the InDel score were given when aligned units also matched in POS, whereas those that did not match were given 0.5 InDel points, including cases where a multi-word expression was mapped to a single word, or two aligned multi-word expressions did not match exactly in their internal syntactic structure. The example sentence alignment in Russian and Bulgarian would yield a modified InDel distance of 4.5, as four Bulgarian words were deleted (or added, from the Russian perspective), and the multi-word Russian expression *таким образом* (*takim obrazom*) 'so, thus' is aligned with a single Bulgarian word *така* (*taka*).

3.1.2 Movement

The syntactic movement distance describes the sum of distances between corresponding syntactic units in two aligned sentences, as a way of measuring the degree to which words or syntactic units must be reordered when translating from one language to another. The distance is measured in terms of alignment positions, e.g., in the example alignment the Russian word *был* (*byl*) 'was' is two alignment positions removed from the Bulgarian *беше* (*beše*) 'was', yielding a movement distance of 2, or a normalized movement distance of 2/14 = 0.14. Two additional variants of the measure exist that seek to mitigate the effect of very large displacements between corresponding units. The first is the logarithmic movement distance, which instead uses the natural logarithm of the number of alignment positions between corresponding units. The example sentences given above would, thus, have a logarithmic movement distance equal to *ln*(2) = 0.69 (normalized: 0.05). The second variant is the binary movement distance, which simply counts the number of syntactic units which have been reordered at all, regardless of distance, between the two sentences. With this variant the example sentence would yield a binary movement distance of 1 (normalized: 0.07).

3.2 Adaptation Surprisal

Adaptation Surprisal (in particular, Word Adaptation Surprisal) quantifies the degree of unexpectedness of a word form given a possibly related word form and set of transformation probabilities (Stenger, Avgustinova, and Marti, 2017). The measure is flexible, easily applicable to either the orthographic level, operating on character correspondences, or the phonetic level, operating on phonetic segments. In both cases, word alignments are necessary, which have been yielded using the Needleman-Wunsch alignment algorithm (Needleman and Wunsch, 1970) with pairwise alignment costs calculated using the methods outlined in the following sections. The basic formula for Word Adaptation Surprisal (WAS) is given in the equation (1):

$$WAS = \frac{1}{n}\sum_{i}^{n} -\log_2 P(L1_i|L2_i)$$

where $L1_i$ refers to the i^{th} character or phonetic segment in the native (decoder) language, and $L2_i$ refers to the corresponding i^{th} segment in the foreign (stimulus) language. This measure has been applied in intercomprehension contexts to model the difficulty for a native speaker of one language (L1) to comprehend a word in a non-native language (L2) (e.g., Mosbach et al., 2019, 2021; Stenger, 2019). More complex or less consistent mappings of orthographic or phonetic characters between two languages result in higher surprisal values, which reflect the amount of uncertainty that an L1 speaker would experience while perceiving an L2 word. Transformation probabilities for both orthographic and phonetic correspondences were extracted with Lidstone smoothing from alignments of the set of cognate word pairs in the parallel translations of "The North Wind and the Sun" fable. In the case of units in one language without an equivalent in the paired language, surprisal was calculated by aligning the attested unit with a gap of the same length (e.g., the deleted *да* (*da*) 'to' from the Bulgarian example sentence would be aligned with a two-character gap unit '--') and calculating surprisal from the correspondence of each orthographic character or phonetic segment with the gap character.

3.2.1 Orthographic Adaptation Surprisal

A substitution cost matrix was generated between all pairs of Latin and Cyrillic characters in order to align orthographic character sequences. Identical characters had zero alignment cost, whereas characters differing only in their diacritics (e.g., <á> and <a>) were assigned an alignment 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, '>[8] were likewise assigned alignment costs of 1 to each another. All remaining character pairs (e.g., consonant-vowel pairs) were assigned a cost of 4.5 to one another. Using a gap-opening penalty of 2, this disallows consonant and vowel characters from being aligned with one another. Latin and Cyrillic words were aligned by first converting Cyrillic characters to ISO 9 Latin equivalents and then applying the alignment function with the costs as specified above. For example, the Russian word <впечатляющий> would be rendered in Latin characters as <vpečatlâûŝij>. A noteworthy exception is the Bulgarian vowel character <ъ>, transcribed in Latin as <ă> but having no clear equivalent in the other Slavic languages, so this vowel character was assigned an alignment cost of 1 to all other vowels. A further step involved accounting for language-specific digraphs, or orthographic bigrams, in each of the languages, shown in Table 1.

PL	ch, cz, dz, dź, dż, rz, sz
CZ	ch, dz, dž, ou
SK	ch, dz, dž
SL, HR	dz, dž, lj, nj
BG, UK, BE	дз, дж

Table 1: Slavic digraphs

In addition to the canonical digraphs displayed in the table, sequences of Cyrillic consonant characters plus a

[7] https://universaldependencies.org/u/pos/

[8] <ъ> being used as the hard sign in Russian, whereas <'> is used for the same purpose in Ukrainian and Belarusian.

soft or hard sign (e.g., <ль>) were likewise considered orthographic bigrams. The Belarusian sequence <дзь>, consisting of the digraph <дз> plus the soft sign <ь>, was treated as an orthographic trigram. Alignment costs were calculated as zero if bigrams matched completely, 0.5 if at least one of the base characters matched (e.g., Polish <rz> with Czech <ř>), and otherwise calculated in the same way as for unigram character alignments. An example orthographic alignment resulting from this scheme for the word 'friend' in the nine Slavic languages included in this study is shown in Table 2. Given an alignment, WAS is calculated using the equation (1) using the transformation probabilities of the aligned characters and gaps in the parallel corpus.

RU	п	р	и		я	т		е	ль
BE	п	р	ы		я	ц		е	ль
UK	п	р	и		я	т		е	ль
PL	p	rz	y	j	a	c	i	e	l
CZ	p	ř	í			t		e	l
SK	p	r	i		a	t		e	ľ
SL	p	r	i	j	a	t		e	lj
HR	p	r	i	j	a	t		e	lj
BG	п	р	и		я	т		е	л

Table 2: Orthographic alignment of the word 'friend'

3.2.2 Phonetic Adaptation Surprisal

Phonetic adaptation surprisal was calculated in largely the same manner as orthographic surprisal, the only difference being the automatic segmentation method and alignment costs employed. Phonetic sequences transcribed in IPA were first automatically segmented into distinct phonetic segments[9] and then aligned using the Needleman-Wunsch algorithm with alignment costs based on the pairwise dissimilarity of individual segment pairs according to a set of weighted phonological distinctive features (Georgis, 2022). Identical phonetic segments had zero alignment cost, while segment pairs which did not share any phonological features were assigned the maximum cost of 1. The gap-opening penalty was set to 0.7 by default, meaning that the alignment algorithm would insert a gap rather than align two segments sharing less than 30 percent of their features. The same word forms for 'friend' aligned phonetically are given in Table 3.

RU	p	rʲ	ɪ	j	ˈæ	t͡ɕ	ɪ	lʲ
BE	p	r	ˈɨ	j	æ	t͡sʲ	e	lʲ
UK	p	r	ˈɪ	j	ɐ	t	e	lʲ
PL	p	ʂ	ɨ	j	ˈa	t͡ɕ	ɛ	l
CZ	p	r̊	ˈiː			t	ɛ	l
SK	p	r		i̯	a	c	ɛ	ʎ
SL	p	r	i	j	áː	t	ɛ	l
HR	p	r	î	j	a	t	e	ʎ
BG	p	r	i	j	ˈa	t	ɛ	l

Table 3: Phonetic alignment of the word 'friend'

[9] e.g., /fpʲɪt͡ɕɪtˈlʲæjʉɕːɪj/ would be automatically segmented into: [/f/, /pʲ/, /ɪ/, /t͡ɕ/, /ɪ/, /t/, /lʲ/, /ˈæ/, /j/, /ʉ/, /ɕː/, /ɪ/, /j/]. Note that IPA diacritics such as <ʲ>, <ˈ>, and <ː> are kept together with the IPA character which they modify.

4. Results

In this Section we present the results of our measurements applied to the nine parallel translations of "The North Wind and the Sun" fable.

4.1 Syntactic Distances

Table 4 shows the syntactic distance between each language pair according to the InDel measure when averaged across all fragments of the text. Czech and Slovak (West Slavic) are significantly less distant from each other than any other language pair. Other highly similar pairs include Croatian and Slovene (South Slavic), Belarusian and Ukrainian (East Slavic), and Russian and Belarusian (East Slavic). These results are unsurprising, as each of these pairs include closely related languages within the same subgroup of Slavic. The most distant pairs are Slovak and Ukrainian, Slovak and Croatian, and Slovak and Belarusian. Although these language pairs do belong to different subgroups of Slavic and, thus, we might expect larger syntactic distances, it should be noted that the distances are also strongly influenced by differences between translations.

	RU	BE	UK	PL	CZ	SK	SL	HR	BG
RU		4.64	6.07	5.79	6.29	7.07	8.21	7.00	6.71
BE	0.33		4.43	5.50	6.29	7.71	7.29	6.14	5.79
UK	0.41	0.32		6.79	7.36	8.50	8.57	7.21	7.00
PL	0.41	0.41	0.43		4.79	5.29	6.29	5.93	5.57
CZ	0.43	0.43	0.48	0.36		2.93	7.07	7.57	6.57
SK	0.45	0.49	0.52	0.38	0.17		7.64	7.93	7.14
SL	0.48	0.43	0.48	0.37	0.44	0.47		5.00	6.86
HR	0.43	0.41	0.44	0.41	0.48	0.49	0.27		7.00
BG	0.40	0.37	0.42	0.37	0.42	0.44	0.39	0.43	

Table 4: Mean normalized (red) and non-normalized (blue) InDel distance

When plotted as a dendrogram in Figure 1 using agglomerative clustering with the UPGMA algorithm, Croatian and Slovene form a distinctive cluster external to the grouping of the remaining languages. One explanation for this result could be that Croatian and Slovene use an analytic construction with an auxiliary verb to form the past tense, e.g., HR *je puhao*, SL *je pihal* 'he blew' whereas the other languages form the past tense with a single-word synthetic form, e.g., CZ *fúkal*, RU *дул* (*dul*), BG *духаше* (*duchaše*).

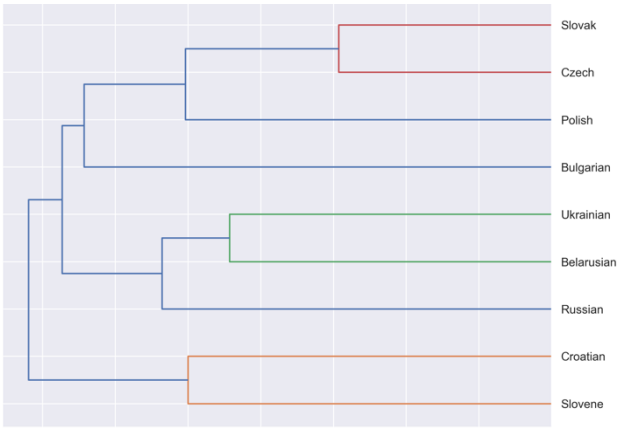

Figure 1: InDel distance dendrogram.

	RU	BE	UK	PL	CZ	SK	SL	HR	BG
RU		10.3	7.1	12.3	4.1	5.4	7.7	11.6	12.1
BE	0.58		9.4	0.3	7.4	4.1	9.6	1.7	0.3
UK	0.43	0.53		7.7	2.6	3.9	6.3	9.6	11.4
PL	0.63	0.02	0.38		8.7	6.0	10.0	2.9	1.7
CZ	0.26	0.28	0.14	0.41		2.6	3.0	9.3	11.1
SK	0.32	0.17	0.20	0.31	0.16		2.0	6.6	5.9
SL	0.40	0.34	0.33	0.38	0.18	0.14		8.3	12.3
HR	0.58	0.09	0.48	0.15	0.42	0.32	0.39		4.6
BG	0.59	0.05	0.59	0.14	0.45	0.29	0.54	0.26	

Table 5: Mean normalized (red) and non-normalized (blue) linear Movement distance

	RU	BE	UK	PL	CZ	SK	SL	HR	BG
RU		1.43	1.00	1.43	1.00	1.14	1.43	2.00	1.71
BE	0.09		1.71	0.14	0.71	0.57	0.86	0.57	0.14
UK	0.06	0.10		1.14	0.71	1.00	1.57	2.00	1.86
PL	0.08	0.01	0.06		1.43	1.14	1.14	0.86	0.43
CZ	0.06	0.03	0.04	0.08		0.71	1.00	1.71	1.29
SK	0.06	0.02	0.05	0.06	0.05		0.57	1.29	1.14
SL	0.07	0.03	0.08	0.06	0.06	0.04		1.29	1.43
HR	0.10	0.03	0.10	0.05	0.09	0.07	0.08		1.00
BG	0.09	0.02	0.10	0.03	0.06	0.06	0.08	0.06	

Table 6: Mean normalized (red) and non-normalized (blue) binary Movement distance

Therefore, each time a past tense verb appears in the text, this necessarily results in at least one point toward the InDel distance between Slovene or Croatian and the other languages, with the auxiliary being always deleted. The remaining cluster of languages splits into a subgroup containing the three East Slavic languages and a subgroup including the three West Slavic languages plus Bulgarian.

The results are somewhat less clear as they pertain to the Movement distance. Table 5 and 6, respectively, show the syntactic distance between each language pair according to the linear and binary Movement measures when averaged across all fragments of the text. There are no obvious patterns in the clusters shown on the dendrograms for either the linear (Figure 2 in the Appendix) or the binary (Figure 3 in the Appendix) syntactic movement distances. A couple factors may be responsible for this. First, Slavic languages are known for having relatively flexible word order. Given that this analysis was performed on literary texts without any standardization of the word order, the results may reflect stylistic differences in the translations more than genuine differences in the ordering of syntactic units as dictated by the grammar of each language. Second, the movement distance can only measure the displacement of equivalent syntactic units. If a syntactic unit has no equivalent in a translated sentence, then its position is irrelevant and results in zero movement distance. Therefore, it seems that the movement distance, at least taken on its own, is less useful in this context.

4.2 Surprisal Asymmetries

Just as they exhibited the smallest syntactic InDel distance to one another, Czech and Slovak also exhibit the least surprisal with one another on average on both the orthographic and phonetic levels. On the orthographic level, Slovak is less surprised by Czech whereas on the phonetic level it is the reverse. This asymmetry appears to be due, at least in part, to the differences in numbers of orthographic and phonetic characters in the two languages. Czech has 37 orthographic characters attested in the text, while Slovak has only 34; Czech has only 33 attested phonetic characters in the transcription while Slovak has 38. Asymmetries also emerge due to uneven correspondences between orthographic or phonetic characters, whereby the mapping is more straightforward in one direction than in the reverse direction. For example, of the 13 instances that the Belarusian orthographic <a> appears in cognates with Russian, it corresponds with Russian <a> six times, with <o> four times, with <e> twice, and with nothing <-> once. Compare this to the Russian letter <a>, which only has two attested orthographic equivalents in Belarusian from this text: <a> six out of seven times, and <ы> once. A Russian native would, thus, experience more difficulty when reading Belarusian than the reverse, as it is less clear which Russian character the Belarusian <a> corresponds with, whereas it is more straightforward for a Belarusian reader. With only a few exceptions, the adaptation surprisal is lower on the orthographic level than on the phonetic level (see Table 7), which can be explained by the fact that there are fewer orthographic characters (mean = 31.3 characters) in each language than phonetic characters (mean = 34 characters). Nevertheless, there is a clear linear relationship between the two types of surprisal, as seen in the scatterplot in Figure 4.[10]

L2/L1	RU	BE	UK	PL	CZ	SK	SL	HR	BG
RU		3.83 3.92	3.93 4.02	3.91 4.21	3.99 4.12	3.89 4.29	3.82 4.33	4.12 4.31	3.83 3.93
BE	3.86 4.10		3.62 3.77	4.23 4.18	4.75 4.53	4.59 4.61	3.84 4.32	4.11 4.37	4.30 4.38
UK	4.00 4.11	3.57 3.64		4.01 3.97	4.55 4.39	4.37 4.58	4.41 4.68	4.32 4.53	4.38 4.41
PL	3.87 4.39	4.13 4.19	4.04 4.15		4.00 4.05	3.87 4.22	3.80 4.44	4.03 4.57	4.01 3.94
CZ	4.27 4.59	4.56 4.49	4.46 4.53	4.17 4.17		2.69 2.84	4.03 4.47	4.36 4.74	4.04 4.23
SK	4.10 4.58	4.54 4.46	4.38 4.54	4.21 4.30	2.70 2.67		3.90 4.51	4.30 4.74	4.19 4.53
SL	4.12 4.36	4.09 4.12	4.78 4.78	4.22 4.35	4.09 4.05	4.06 4.37		3.23 3.60	3.73 3.87
HR	4.33 4.50	4.36 4.46	4.49 4.59	4.30 4.65	4.31 4.45	4.11 4.45	3.22 3.84		4.04 4.22
BG	3.92 4.06	4.28 4.29	4.52 4.54	4.25 4.09	3.93 3.99	4.09 4.48	3.52 4.00	3.94 4.13	

Table 7: Mean normalized orthographic (purple) and phonetic (green) adaptation surprisal (in bits)

Another consideration for the surprisal measurements is that these are strongly impacted by the number of cognate words in the text for each language pair. The explanation for this is two-fold. First, because the transformation probabilities between orthographic and phonetic characters needed for the calculation of surprisal are extracted using only cognate word pairs found in "The North Wind and the Sun" fable (see Table 8), language pairs with fewer attested cognates have less available data to calculate these probabilities, meaning that some genuine character correspondences may never or only seldom appear in this small set of words. Second, whereas cognate word pairs in related languages usually exhibit

[10] In Figure 4 the language pairs are listed in L1-L2 format.

consistent correspondences in their forms, non-cognate words are unrelated to each other and, thus, necessarily yield higher surprisal values, as they do not exhibit the regular correspondences seen in cognate pairs.

	RU	BE	UK	PL	CZ	SK	SL	HR	BG
RU		27	30	24	21	18	21	18	24
BE	39		34	23	16	14	21	13	15
UK	38	50		26	17	16	11	12	16
PL	33	37	36		27	24	17	16	20
CZ	27	20	20	32		47	20	12	19
SK	23	19	19	31	61		15	11	16
SL	35	30	19	26	27	25		23	25
HR	29	20	20	24	19	19	47		21
BG	34	22	22	26	26	24	37	32	

Table 8: Numbers of cognate tokens (red) and cognate types (blue) between languages in the story

As already mentioned in Section 1.2, adaptation surprisal values can be asymmetrical, which is advantageous compared to symmetrical syntactic distances (via InDel and Movement). Indeed, speakers of language A may understand language B better than language C, i.e., [A(B)>A(C)], while speakers of language B may understand language C better than language A, i.e., [B(C)>B(A)]. The degree of asymmetries differs between spoken and written modalities, too. Adaptation surprisal metrics shows different constellations. For example, the orthographic adaptation surprisal is minimal from Russian (L2) to Slovene (L1), while this is not the case at the phonetic level. In contrast, Slovene (L2) is not so close to Russian (L1), see Table 7. The exact prediction potential of adaptation surprisal can be tested only experimentally with speakers of the respective languages.

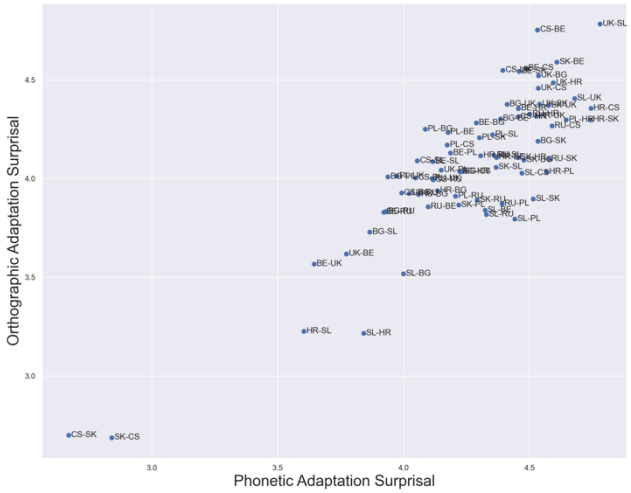

Figure 4: Scatterplot of phonetic vs. orthographic adaptation surprisal (Pearson's $r = 0.91$; $p < 0.001$).

4.3 Relationship between Syntactic Distances and Surprisal Asymmetries

In order to understand the relationships between the syntactic distances and surprisal-based asymmetries employed in this study, we have measured the correlations between them. Similarly to Heeringa et al. (2017), we found no significant correlation between the mean InDel distance and any of the mean movement distances, which confirms that these two types of syntactic measures concern separate aspects of syntactic variation. However, the mean InDel measure is strongly correlated with both mean orthographic adaptation surprisal (Pearson's $r = 0.86$; $p < 0.001$) and even more so with mean phonetic adaptation surprisal (Pearson's $r = 0.91$; $p < 0.001$). A very strong correlation was likewise found between the two measures of adaptation surprisal (Pearson's $r = 0.91$; $p < 0.001$). Concerning the mean movement distances, all three types of movement (linear, binary, logarithmic) were strongly correlated with one another (Pearson's $r > 0.80$ and $p < 0.001$ in all cases). However, this is to be expected given that these measures are simply variations of one another calculated on the same input. No significant correlation was discovered between the movement measures and the adaptation surprisal measures.

5. Conclusions and Discussion

Modeling the impact of syntactic distance and surprisal on cross-Slavic text comprehension, we addressed the following research questions: RQ1: How syntactically distant are the nine Slavic languages from one another? RQ2: What asymmetries are predictable by means of adaptation surprisal between selected languages from phonetic and orthographic viewpoints? RQ3: What is the relation between the measuring methods employed in this study?

Comparing the mean values of syntactic distances within Slavic subgroups (RQ1), we see that in terms of both the mean InDel distance and the mean Movement distances, the West Slavic languages (Polish, Czech, and Slovak) are more similar to one another than the East (Russian, Belarusian, and Ukrainian) and South Slavic languages (Slovene, Croatian, and Bulgarian). The East Slavic languages occupy the second position in terms of within-group similarity, followed by South Slavic languages, as measured by the aggregated InDel distance, though the reverse ordered is observed with the aggregated Movement distances. In general, Czech and Slovak (West Slavic) are the least distant language pair in comparison to other language pairs in terms of the aggregated InDel distance. This means that Czech and Slovak native speakers should experiences a greater degree of success in textual intercomprehension in comparison to speakers of other pairs of Slavic languages. Other highly similar pairs include Croatian and Slovene (South Slavic), Belarusian and Ukrainian (East Slavic). However, the results of the mean Movement distances (linear and binary) are in general less clear. As previously stated in Section 4, this can be explained by the fact that Slavic languages are known for having relatively flexible word order and that this analysis is performed on literary texts with stylistic differences in the translations.

Analyzing asymmetries by means of adaptation surprisal (RQ2), we found that with only a few exceptions, the mean normalized adaptation surprisal is lower on the orthographic level than on the phonetic level. Thus, we assume that native speakers of selected Slavic languages should in general have less difficulty comprehending a written text in another Slavic language than speech. However, it must be mentioned that this finding ignores the fact that two different scripts, Cyrillic and Latin, are used in different Slavic languages. Thus, our assumption that written texts are easier to comprehend should be interpreted with caution, as general trends in Cyrillic (Belarusian, Bulgarian, Russian, and Ukrainian) and Latin

(Croatian, Czech, Polish, Slovak, and Slovene) script intelligibility vary among Slavic native speakers. In general, Czech and Slovak exhibit the least normalized adaptation surprisal with one another on average on both the orthographic and phonetic levels. This means that in a receptive multilingualism scenario Czech speakers should have fewer difficulties than Slovak speakers in perception of a spoken text and more difficulties than Slovak speakers in perception of a written text, and vice versa. However, the differences in asymmetry are very small on both levels.

Additionally, we found high and significant correlations between the mean InDel distance and the mean orthographic and phonetic adaptation surprisal, as well as between the two measures of adaptation surprisal. No significant correlations were found between the mean movement distances and the mean InDel distance, no between the movement distance and the two measures of adaptation surprisal. All three types of movement (linear, binary, logarithmic) were strongly correlated only with one another (RQ3).

The exact prediction potential of syntactic distance and surprisal will be validated with intelligibility scores obtained in our web-based experiments by means of written and spoken cloze tests among speakers of selected Slavic languages.

6. Acknowledgements

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project-ID 232722074 - SFB 1102 and by Saarland University (UdS-Internationalisierungsfonds).

7. Bibliographical References

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*. Amsterdam: John Benjamins, pp. 13--24.

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

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.

Feleke, T. L., Gooskens, C., and Rabanus, S. (2020). Mapping the dimensions of linguistic distance: A study on South Ethiosemitic languages. *Lingua*, vol. 243, https://doi.org/10.1016/j.lingua.2020.102893

Georgis, P. (2022). *Phonetic and Information-Theoretic Distance Methods for Automated Linguistic Phylogenetic Inference*. Master thesis. Saarland University.

Gooskens, C. and Heeringa, W. (2004). Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data. *Language variation and Change*, 16(3):189–207.

Gooskens, C. and van Heuven, V. J. (2021). Mutual intelligibility. In M. Zampieri & P. Nakovi (Eds.), *Similar Languages, Varieties, and Dialects: A Computational Perspective.* Studies in Natural Language Processing. Cambridge: Cambridge University Press, pp. 51--95.

Golubović, J. (2016). *Mutual intelligibility in the Slavic language area*. Doctoral dissertation. Groningen: University of Groningen.

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

Heeringa, W., Swarte, F., Schüppert, A., and Gooskens, C. (2017). Measuring syntactical variation in Germanic texts. *Digital Scholarship in the Humanities*, 33(2):279–296.

International Phonetic Association. (1999). *Handbook of the International Phonetic Association: a guide to the use of the international phonetic alphabet*. Cambridge: Cambridge University Press.

Jágrová, K., Stenger, I., Avgustinova, T., and Marti, R. (2017). Lexical and orthographic distances between Bulgarian, Czech, Polish, and Russian. In *Language Use and Linguistic Structure. Proceedings of the Olomouc Linguistics Colloquium (2016)*, pages 401–416, Olomouc, Palacký University.

Lauttamus, T., Nerbonne, J., and Wiersma, W. (2007). Detecting syntactic contamination in emigrants. The English of Finnish Australians. *SKY Journal of Linguistics*, 21:273–307.

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

Mosbach, M., Stenger, I., Avgustinova, T., and D. Klakow, D. (2019). incom.py - A Toolbox for Calculating Linguistic Distances and Asymmetries between Related Languages. In *Proceedings of Recent Advances in Natural Languages Processing* (*RANLP 2019*), pages 810–818, Varna, Bulgaria, September.

Mosbach, M., Stenger, I., Avgustinova, T., Möbius, B., and Klakow, D. (2021). incom.py 2.0 - A Toolbox for Calculating Linguistic Distances and Asymmetries in Auditory Perception of Closely Related Languages. In *Proceedings of Recent Advances in Natural Languages Processing* (*RANLP 2021*), pages 191–200, Varna, Bulgaria, September.

Needleman, S. B. 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.

Nerbonne, J. and Wiersma, W. (2006). A measure of aggregate syntactic distance. In *Proceedings of the Workshop on linguistic Distances*, pages 82–90. Association for Computational Linguistics.

Rehbein, J., ten Thije, J. D., and Verschik, A. (2012). Lingua receptiva – remarks on the quintessence of receptive multilingualism. *Receptive Multilingualism. Special issue of the International Journal of Bilingualism*, 16(3):248–264.

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*. Doctoral dissertation. Saarbrücken: universaar.

Stenger, I. and Avgustinova, T. (2021). On Slavic cognate recognition in context. In *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference 'Dialogue' (2021)*, Issue 20, pages 660–668, Moscow, Russia, June.

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. In *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference 'Dialogue' (2017)*, Issue 16(23), vol. 1, pages 304–317, Moscow, Russia, May–June.

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 Linguistic* 40(2):175–199.

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 & P. Kosta (Eds.), *Current developments in Slavic Linguistics. Twenty years after (based on selected papers from FDSL 11)*. Frankfurt a.M.: Peter Lang, pp. 483--500.

Swarte, F. (2016). *Predicting the Mutual Intelligibility of Germanic languages from linguistic and extra-linguistic factors.* Doctoral dissertation. Groningen: University of Groningen.

Tang, C. and van Heuven, V. J. (2009). Mutual intelligibility of Chinese dialects experimentally tested. *Lingua* 119:709–732.

7. Language Resource References

Aesop Language Bank, https://www.aesoplanguagebank.com/index.html.

Illustrations of the IPA, https://richardbeare.github.io/marijatabain/ipa_illustrations_all.html.

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

ISO 9:1995, https://de.wikipedia.org/wiki/ISO_9.

Universal POS tags, https://universaldependencies.org/u/pos/.

Syntactic distances and surprisal between nine Slavic languages: Belarusian, Bulgarian, Croatian, Czech, Polish, Slovak, Slovene, Russian, and Ukrainian of the fable "The North Wind and the Sun", https://github.com/slavic-lab/LREC-2022-SynDist-Surprisal

Appendix

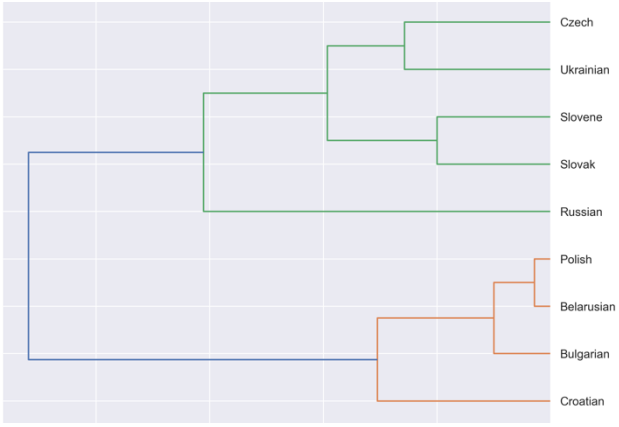

Figure 2: Linear Movement distance dendrogram.

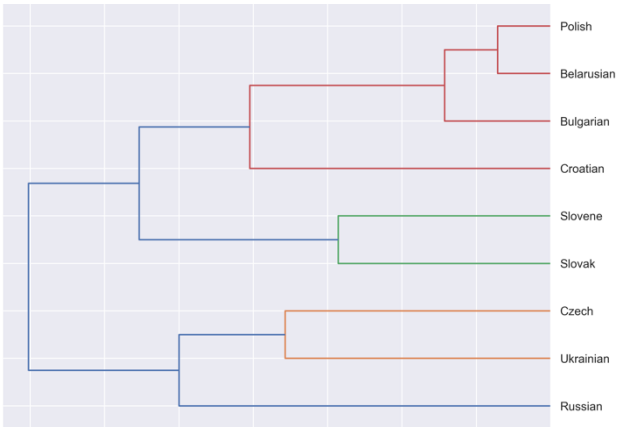

Figure 3: Binary Movement distance dendrogram.