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, Saarbrücken, Germany

SFB 1102, Projekt 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

Background

- Intercomprehension as a special mode of language use.
- Certain degrees of mutual intelligibility between (closely) related languages in written and spoken modalities.
- Existing and perceived similarities at different linguisting levels.

Contributions of this Study

- Modeling intercomprehension among nine Slavic languages: Belarusian, Bulgarian, Croatian, Czech, Polish, Slovak, Slovene, Russian, and Ukrainian information-theoretically:
  - **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 views?
  - **RQ3:** What is the relation among the measures under study here?

Material

- The fable “The North Wind and the Sun” in nine Slavic languages available at the International Phonetic Association (1999) for illustration purposes https://richardbeare.github.io/marijatabain/ipa_illustrations_all.html.
- The parallel translations are split into seven distinct fragments, each fragment is aligned in a multiple alignment scheme, all corresponding syntactic units within a fragment match with one another.
- The Russian-Bulgarian final fragment “*And so the North Wind was obliged to confess that the Sun was the stronger of the two*”:

RU		Таким образом	северный ветер		вынужден	был	
		Takim obrazom	*severnyj veter*		*vynužden*	*byl*	
BG	И	така	северният вятър	беше	принуден		да
	I	*taka*	*severniât vâtăr*	*beše*	*prinuden*		*da*

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

Corresponding units appearing in different sentence positions are highlighted in green, syntactic units without corresponding pairs are marked in red text.

Measuring Methods

- **SYNTACTIC DISTANCE** (Heeringa et al., 2017)
- The *Indel distance* (InDel) measures the average number of words (syntactic units) which are inserted or deleted in parallel sentences.
- The *binary movement distance* measures the average number of words that must be reordered in sentences of L1 in order to produce the word order of an equivalent sentence in L2.
- The *linear movement distance* measures the number of word positions a word from a sentence in L1 has moved compared to the corresponding word in an equivalent sentence in L2.
- **ADAPTATION SURPRISAL** (Stenger, Avgustinova, and Marti, 2017)
- Adaptation Surprisal, in particular *Word Adaptation Surprisal* (WAS), quantifies the degree of unexpectedness of a word form given a possibly related word form and set of transformation probabilities.

$L1_i$ refers to the i^{th} character or sound in the native (decoder) language and $L2_i$ refers to the i^{th} character or sound in the foreign (stimulus) language and

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

Results: Adaptation Surprisal

	East Slavic			West Slavic			South Slavic		
L2/L1	RU	BL	UK	PL	CZ	SK	SL	HR	BG
RU		3.83	3.93	3.91	3.99	3.89	3.82	4.12	3.83
BL	3.86		3.62	4.23	4.75	4.59	3.84	4.11	4.30
UK	4.00	3.57		4.01	4.55	4.37	4.41	4.32	4.38
PL	3.87	4.13	4.04		4.00	3.87	3.80	4.03	4.01
CZ	4.27	4.56	4.46	4.17		2.69	4.03	4.36	4.04
SK	4.10	4.54	4.38	4.21	2.70		3.90	4.30	4.19
SL	4.12	4.09	4.78	4.22	4.09	4.06		3.23	3.73
HR	4.33	4.36	4.49	4.30	4.31	4.11	3.22		4.04
BG	3.92	4.28	4.52	4.25	3.93	4.09	3.52	3.94	

	East Slavic			West Slavic			South Slavic		
L2/L1	RU	BL	UK	PL	CZ	SK	SL	HR	BG
RU		3.92	4.02	4.21	4.12	4.29	4.33	4.31	3.93
BL	4.10		3.77	4.18	4.54	4.61	4.32	4.37	4.38
UK	4.11	3.64		3.97	4.39	4.58	4.68	4.53	4.41
PL	4.39	4.19	4.15		4.05	4.22	4.44	4.57	3.94
CZ	4.59	4.49	4.53	4.17		2.84	4.47	4.74	4.23
SK	4.58	4.46	4.54	4.30	2.67		4.51	4.74	4.53
SL	4.36	4.12	4.78	4.35	4.05	4.37		3.60	3.87
HR	4.50	4.46	4.59	4.65	4.45	4.45	3.84		4.22
BG	4.06	4.29	4.54	4.09	3.99	4.48	4.00	4.13	

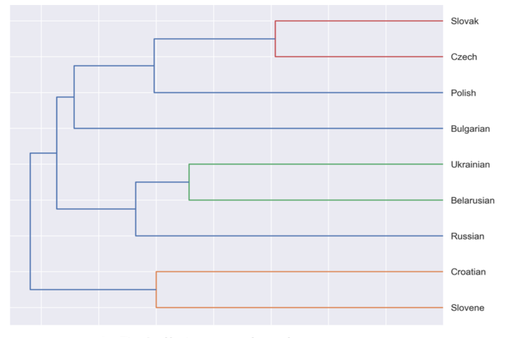
Mean normalized orthographic (purple) and phonetic (green) adaptation surprisal in bits

Results: Syntactic Distances

InDel

	East Slavic			West Slavic			South Slavic		
	RU	BL	UK	PL	CZ	SK	SL	HR	BG
RU		4.64	6.07	5.79	6.29	7.07	8.21	7.00	6.71
BL	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	

Mean normalized (red) and non-normalized (blue) InDel distance



InDel distance dendrogram

Binary Movement

	East Slavic			West Slavic			South Slavic		
	RU	BL	UK	PL	CZ	SK	SL	HR	BG
RU		1.43	1.00	1.43	1.00	1.14	1.43	2.00	1.71
BL	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	

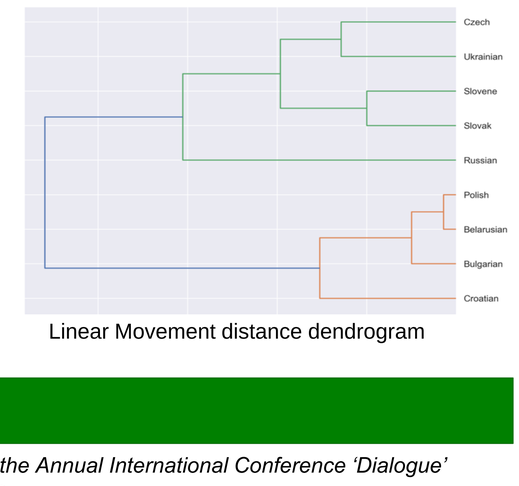
Mean normalized (red) and non-normalized (blue) binary Movement distance

Binary Movement distance dendrogram

Linear Movement

	East Slavic			West Slavic			South Slavic		
	RU	BL	UK	PL	CZ	SK	SL	HR	BG
RU		10.3	7.1	12.3	4.1	5.4	7.7	11.6	12.1
BL	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	

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



Linear Movement distance dendrogram

Conclusions and Outlook

RQ1: The West Slavic languages are more similar to one another than the East and South Slavic languages.

RQ2: With only a few exceptions, the mean normalized adaptation surprisal is lower on the orthographic level than on the phonetic level. Czech and Slovak exhibit the least normalized adaptation surprisal with one another on average on both the orthographic and phonetic levels.

RQ3: High and significant correlations between the mean InDel distance and the mean orthographic (Pearson’s $r = 0.86$; $p < 0.001$) and phonetic adaptation surprisal (Pearson’s $r = 0.91$; $p < 0.001$), as well as between the two measures of adaptation surprisal (Pearson’s $r = 0.91$; $p < 0.001$).

The exact prediction potential of measure methods will be validated with intelligibility scores obtained in web-based experiments among speakers of selected Slavic languages.

References

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.

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.

Syntactic distances and surprisal between nine Slavic languages of the fable “The North Wind and the Sun”, https://github.com/slavic-lab/LREC-2022-SynDist-Surprisal