Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2017"

Moscow, May 31—June 3, 2017

LEVENSHTEIN DISTANCE AND WORD ADAPTATION SURPRISAL AS METHODS OF MEASURING MUTUAL INTELLIGIBILITY IN READING COMPREHENSION OF SLAVIC LANGUAGES

Stenger I. (ira.stenger@mx.uni-saarland.de),
Avgustinova T. (avgustinova@coli.uni-saarland.de),
Marti R. (rwmslav@mx.uni-saarland.de)

Saarland University, Saarbrücken, Germany

In this article we validate two measuring methods: Levenshtein distance and word adaptation surprisal as potential predictors of success in reading intercomprehension. We investigate to what extent orthographic distances between Russian and other East Slavic (Ukrainian, Belarusian) and South Slavic (Bulgarian, Macedonian, Serbian) languages found by means of the Levenshtein algorithm and word adaptation surprisal correlate with comprehension of unknown Slavic languages on the basis of data obtained from Russian native speakers in online free translation task experiments. We try to find an answer to the following question: Can measuring methods such as Levenshtein distance and word adaptation surprisal be considered as a good approximation of orthographic intelligibility of unknown Slavic languages using the Cyrillic script?

Keywords: Levenshtein distance, word adaptation surprisal, orthographic intelligibility, reading intercomprehension, East and South Slavic languages

РАССТОЯНИЕ ЛЕВЕНШТЕЙНА И МЕРА НЕОЖИДАННОСТИ АДАПТАЦИИ СЛОВА КАК МЕТОДЫ ИЗМЕРЕНИЯ МЕЖЪЯЗЫКОВОЙ ПОНЯТНОСТИ СЛАВЯНСКИХ ЯЗЫКОВ ПРИ ЧТЕНИИ

Штенгер И. (ira.stenger@mx.uni-saarland.de),
Августинова Т. (avgustinova@coli.uni-saarland.de),
Марти Р. (rwmslav@mx.uni-saarland.de)

Университет земли Саар, Саарбрюккен, Германия

В данной статье мы проверяем два метода оценки степени близости родственных языков — расстояние Левенштейна и меру неожиданности адаптации слова — в качестве потенциальных параметров для определения успеха межъязыкового понимания в ситуации, когда читателю необходимо извлечь информацию из текста на незнакомом языке. Мы исследуем, в какой степени орфографические дистанции между русским языком и другими восточнославянскими (украинским, белорусским) и южнославянскими языками (болгарским, македонским, сербским), установленные с помощью алгоритма Левенштейна и меры неожиданности адаптации слова, соотносятся с экспериментальными результатами понимания незнакомых славянских языков носителями русского языка. Сбор данных был выполнен на базе экспериментов в виде заданий по свободному переводу в режиме онлайн. Мы попытаемся найти ответ на следующий вопрос: могут ли такие методы измерения как расстояние Левенштейна и мера неожиданности адаптации слова оптимально оценить понятность орфографии незнакомых славянских языков, использующих кириллицу?

Ключевые слова: расстояние Левенштейна, мера неожиданности адаптации слова, понятность орфографии, взаимопонимание при чтении, восточнославянские и южнославянские языки

1. Introduction

Intercomprehension (Doyé 2005), receptive multilingualism (Braunmüller and Zeevaert 2001) or semi-communication (Haugen 1966) reveals the human ability to understand (but not speak or write) one or more unknown foreign languages that are related to at least one language in the individual's linguistic repertoire. More or less systematic mutual intelligibility investigation was undertaken for certain language groups, e.g., Scandinavian (Gooskens 2006), Germanic (Möller and Zeevaert 2015, Vanhove 2015), West and South Slavic (Golubović and Gooskens 2015). It was shown that the degree of intelligibility of an unknown but (closely) related language depends on both linguistic and extra-linguistic factors (Gooskens 2013).

In all reading activities, orthography is the primary linguistic interface for extracting information from unfamiliar encodings in the stimulus, and it critically affects the transmission of information across languages. The present study focuses on the orthographic intelligibility of written East Slavic (Ukrainian, Belarusian) and South Slavic (Bulgarian, Macedonian, Serbian) languages for Russian subjects. We consider two linguistic distance measurements as potential predictors of successful reading intercomprehension and validate them in web-based experiments.

This article is organized as follows. Section 2 gives a short overview of the Cyrillic alphabet and the main orthographic principles. Section 3 presents the online experiment of testing orthographic intelligibility and describes the experimental data on the basis of which both the normalized Levenshtein distance and the normalized word adaptation surprisal were calculated. In Section 4, the two measuring methods are correlated with the experimental results. Finally, some general conclusions are drawn and future work is outlined.

2. Cyrillic orthographic code

While an alphabet of a language consists of a set of letters (graphemes) used to compose written texts in that language (Sgall 2006), the mechanisms of orthographic code are determined by various principles underlying the established writing systems. All the languages we investigate here employ the Cyrillic alphabet, albeit with slight adaptations, i.e.[1]

Russian (RU)	а б в г д е ё[1] ж з и й к л м н о п р с т у ф х ц ч ш щ ъ ы ь э ю я	(33)
Ukrainian (UK)	а б в г ґ д е є ж з и і ї й к л м н о п р с т у ф х ц ч ш щ ь ю я	(33)
Belarusian (BE)	а б в г д е ё ж з і й к л м н о п р с т у ў ф х ц ч ш ы ь э ю я	(32)
Bulgarian (BG)	а б в г д е ж з и й к л м н о п р с т у ф х ц ч ш щ ъ ь ю я	(30)
Macedonian (MK)	а б в г д ѓ е ж з ѕ и ј к л љ м н њ о п р с т ќ у ф х ц ч џ ш	(31)
Serbian (SR)	а б в г д ђ е ж з и ј к л љ м н њ о п р с т ћ у ф х ц ч џ ш	(30)

Slavic orthographies using the Cyrillic alphabet are based primarily on the phonemic principle, but they also observe other principles, e.g., phonetic, morphological, historical/etymological (Kučera 2009). For example, the Serbian orthography adheres basically to the phonemic principle, with a strong tendency towards the phonetic principle (Kučera 2009, Marti 2014). Despite the fact that Russian orthography is based in general on the phonemic principle, the morphological principle is relevant too, depending on what is understood as a phoneme (Ivanova 1991, Musatov 2012). All Slavic orthographies represent nowadays so-called mixed systems providing the respective languages with a number of general patterns (for more details see Stenger 2016).

[1] The letter *ë* is generally used in dictionaries and schoolbooks only.

3. Material and methods

We investigate reading intercomprehension among Slavic languages and approach the problem of their mutual intelligibility from an information-theoretic perspective in terms of surprisal, taking into consideration information en- and decoding at different linguistic levels.[2]

When research in spoken semi-communication or in reading intercomprehension focuses on testing text understanding (i.e. Beijering et al. 2008, Golubović and Gooskens 2015, Gooskens 2007), the intelligibility scores are based on the text as a whole. This means, in particular, that the influence of different linguistic factors — such as textual and sentence context, syntax, lexis, morphology, phonetics/phonology or orthography — cannot be distinguished in detail, if at all identifiable.

We wanted to determine the role of orthography in written intercomprehension, and for that reason chose to focus on isolated cognate recognition first. Even though it may seem artificial to test cognates without context, since the latter may provide helpful information, the underlying assumption here is that the correct 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.

3.1. Experiment

The orthographic intelligibility between Russian and five other Slavic languages was tested in web-based experiments.[3] 119 native speakers of Russian between 14 and 71 years of age (average 34 years) took part in the challenge. Around three-fourths of them were female. The participants started the experiment with registration and then completed a background questionnaire in their native language. Afterwards 6 challenges were presented randomly: 2 challenges with 60 different BG stimuli in each group, 1 challenge with 60 UK stimuli, 1 challenge with 60 BE stimuli, 1 challenge with 50 MK stimuli, and 1 challenge with 50 SR stimuli. The choice of the stimuli from the manually prepared lists and the order of presentation were also randomized. The participants saw the stimuli on their screen, one by one, and were given 10 seconds to translate each word into Russian. The time limit was carefully piloted taking into consideration the experience of other experiments in reading intercomprehension. During the experiment the participants received feedback in form of emoticons for their answers. The allocated time limit seemed to be sufficient for typing even the longest words, but not long enough for using a dictionary or online translation tools. It was possible to finish before the 10 seconds were over by either clicking on the 'Next' button or pressing 'Enter' on the keyboard. After 10 seconds the participants saw the next stimulus on their screen. The results were automatically categorized as "right" or "wrong" via pattern matching with expected answers. Some stimuli had more

---

[2] This study was carried out within the project INCOMSLAV *Mutual Intelligibility and Surprisal in Slavic Intercomprehension*, which is part of the Collaborative Research Center 1102 *Information Density and Linguistic Encoding*.

[3] The web application is available at http://intercomprehension.coli.uni-saarland.de/ru/

than one possible translation. We also provided a list of so-called alternative correct answers. For example, the BE word *дзіця* (*dzicja*)[4] '*child*' can be translated in Russian as *дитя* (*ditja*) or *ребёнок (rebënok*) or *ребенок (rebenok),* all meaning '*child*'. All these translations were counted as correct.

In the present study we exclude those participants who have indicated knowledge of the stimuli language(s) in the questionnaire and analyze the results only of the initial challenge for each participant in order to avoid any learning effects. The mean percentage of correctly translated items constitutes the intelligibility score of a given language (Table 1).

Table 1. The results of free translation task experiments

Stimuli languages	Participants' native language: RU
UK	80.42%
BE	71.66%
BG	70.88%
MK	61.81%
SR	57.16%

3.2. Material

For the computational transformation experiments on parallel word sets presented in (Fischer et al. 2015), we collected and aligned parallel Slavic word lists, at first for two language pairs: Czech—Polish and Bulgarian—Russian. For each pair, a list of internationalisms and a list of Pan-Slavic vocabulary were freely available from the EuroComSlav website.[5] Additionally we compiled a third parallel list of cognates from Swadesh lists for these languages.[6] All three lists were slightly modified. Thus, formal non-cognates were removed and formal cognates, if existing, were added to the lists where the pairs in the original lists consisted of non-cognates. For example, BG–RU *ние–мы* (*nie–my*) '*we*' were removed and the BG *звяр* (*zvjar*) '*beast*' instead of *животно* (*životno*) '*animal*' was added to its RU formal cognate *зверь* (*zver'*) '*animal*, *beast*'. The linguistic items in these lists belong to different parts of speech, mainly nouns, adjectives, and verbs.

In the second step, we manually collected a cross-linguistic rule set of corresponding orthographical units (transforming both individual letters and letter strings) from comparative historical Slavic linguistics (e.g. Bidwell 1963, Vasmer 1973, Žuravlev et al. 1974–2012). This resulted in sets of diachronically-based orthographic correspondences, e.g. BG–RU: *б*:*бл*, *жд*:*ж*, *я*:*е*, *ла*:*оло* etc. We then tested this set of diachronically-based orthographic correspondences on the parallel word lists mentioned above.

---

[4] Transliteration is given according to DIN 1460.

[5] Pan-Slavic list: http://www.eurocomslav.de/kurs/pwslav.htm;
internationalism list: http://www.eurocomslav.de/kurs/iwslav.htm (accessed 11.07.2015).

[6] Swadesh-list: http://en.wiktionary.org/wiki/Appendix:Swadesh_lists_for_Slavic_languages.

By applying the transformation rules, we categorized the cognates in the pairs as (i) identical, (ii) successfully transformed, or (iii) non-transformable by the rules. In most cases, the automatic transformations were judged to be satisfactory, e.g. BG–RU 128 correctly transformed items excluding doublets of a total of 935 items in all three lists (for more details see Fischer et al. 2015).

In addition, we carried out orthographic transformation experiments on the parallel word lists of Common Slavic vocabulary (Carlton 1991, Mel'nyčuk 1966) for the language pairs UK–RU, BE–RU, BG–RU, MK–RU, and SR–RU. The Common Slavic vocabulary consists of 212 examples for 15 Slavic languages. While the original data have some empty slots for some of the languages, the parallel vocabulary lists for the computational transformation include only 190 items for all languages, consisting mostly of nouns, with a small amount of 23 adjectives and 27 verbs in each language. The number of successfully transformed items differs in the respective pairs: 102 items for BE–RU, 76 items for UK–RU, 68 items for SR–RU, 63 items for BG–RU, and 62 items for MK–RU. The correctly transformed items from all computational transformation experiments are used as the basis for the selection of stimuli in our web-based experiments (Section 3.1). In this way we could exclude possible different derivational morphemes between related languages in order to focus on the impact of mismatched orthographic correspondences in cognate intelligibility.

3.3. Levenshtein distance

Orthographic distances between corresponding cognates are usually measured on the basis of the Levenshtein distance metric (Levenshtein 1966). Kessler (1995) introduced the algorithm for measuring distances between Irish Gaelic dialects. Since then it has been applied successfully not only to different dialects of one language, but also to (closely) related languages (Beijering et al. 2008, Gooskens 2006). Levenshtein distance is considered a fairly good predictor of overall intelligibility in speech semi-communication in related language varieties as well as in reading intercomprehension (Gooskens 2007, Kürschner et al. 2008, Vanhove and Berthele 2015).

The Levenshtein distance between corresponding words is based upon the minimum number of symbols that need to be inserted, deleted or substituted in order to transform the string in one language into the corresponding string in another language. 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*, 1 to map it to a different character, e.g. *a*:*o*. Insertions and deletions of different characters cost 1. In more sensitive versions, base and diacritic may be distinguished. For example, the base of *ë* is *e*, and the diacritic is the diaeresis. Though it is not exactly clear what weight should be attributed to each of the components (Gooskens and Heeringa 2004), it is generally assumed that differences in the base will usually confuse the reader to a much greater extent than diacritical differences (Heeringa et al. 2013). If two characters have the same base but differ in diacritics, we assign them a substitution cost of 0.5.[7] In order

[7] Since we do not have any MK–RU cognates with the following alignment: *ѕ́*:*ѕ* and *ќ*:*к*, we weigh these pairs as 1 in our Levensthein distance matrix.

to obtain distances which are based on linguistically motivated alignments, the algorithm is adapted so that in the alignment a letter representing a vowel (henceforth called a vowel letter) may only correspond to a vowel letter and a consonant letter only to a consonant letter.

We consider the normalized Levenshtein distance with regard to the assumption that a segmental difference in a word of two segments has a stronger impact on intelligibility than a segmental difference in a word of ten segments (Beijering et al. 2008). The normalized Levenshtein distance of BG–RU: *риба–рыба* (*riba–ryba*) '*fish*' is 1:4=0.25 or 25%. We calculate the average of the Levenshtein distance between stimuli of selected Slavic languages and their cognates in RU (Table 2). The assumption is: The larger the distance, the more difficult it is to comprehend the related language (Section 4.1).

Table 2. Normalized Levenshtein distances between Russian and other Slavic languages given as percentages

Slavic languages	RU
UK	23.87%
BE	28.92%
BG	25.61%
MK	28.92%
SR	34.26%

3.4. Word adaptation surprisal

In addition to the Levenshtein distance we use the information-theoretic concept of *surprisal*. The term *surprisal* was introduced by Tribus (1961), who used it to talk about the logarithm of the reciprocal of a probability (Hale 2016). Surprisal allows us to characterize the information value of an observed event and has been shown to correlate in many cases with various metrics of success, such as reading times in eye-tracking experiments (Boston et al. 2008, Smith and Levy 2013). Surprisal is defined as the code length of an optimal prefix-free code for a given probability distribution and thus has an inverse logarithmic relation to the probability values themselves (Shannon 1948). Surprisal values are given in bits and depend heavily on the used probability distribution.

We calculated the letter adaptation surprisal with the following formula (1). Letter adaptation surprisal values allow for quantifying the unexpectedness both of individual letter correspondences and of whole cognate pairs.

(1) $$surprisal(L1 = l1|L2 = l2) = -\log P(L1 = l1|L2 = l2)$$

L1 — native (decoder) language, l1 — letter of the native (decoder) language, L2 — foreign (stimulus) language, l2 — letter of the foreign (stimulus) language

We can also compute the adaptation surprisal for string correspondences in our set. For example, the BG–RU cognate pair *глад–голод* (*glad–golod*) '*hunger*' contains

a string correspondence *ла*:*оло*. The adaptation surprisal of the string correspondence can be calculated by summing up the letter adaptation surprisal of the contained letters: 2.0506 surprisal for *ø*:*о*, 0.0 surprisal for *л*:*л* and 1.8210 surprisal for *а*:*о*.[8]

In this study we investigate the word adaptation surprisal. We compute full word adaptation surprisal by summing up the letter adaptation surprisals. For example, the BG–RU cognate pair *син–сын* (*sin–syn*) '*son*' contains the correspondences *с*:*с* (0.0 surprisal), *и*:*ы* (0.6919 surprisal) and *н*:*н* (0.0 surprisal). Thus, the BG *син* (*sin*) '*son*' has a word adaptation surprisal of 0.6919 bits for Russian readers. This gives a quantification of the (un)expectedness of the correct cognate in Russian *сын* (*syn*) '*son*'. As in the case with the Levenshtein distance, we also normalize the full word adaptation surprisal, e.g. 0.6919:3=0.2306 or 23.06% for BG–RU *син–сын* (*sin–syn*) '*son*'. We calculate the average value of the word adaptation surprisal between stimuli of selected Slavic languages and their cognates in Russian (Table 3): The higher the surprisal, the more difficult it is to comprehend the related language and the more time is needed to complete the translation task (Section 4.2 and 4.3).

Table 3. Normalized word adaptation surprisal between Russian and other Slavic languages given as percentages

Slavic languages	RU
UK	34.79%
BE	48.05%
BG	50.30%
MK	73.01%
SR	79.55%

4. Results

4.1. Levenshtein distance and intelligibility score

To investigate the relationship between intelligibility and Levenshtein distance scores, the results of the orthographic intelligibility tests are correlated with the overall Levenshtein distances (Fig. 1).

There is a negative correlation of -0.89 ($p < 0.05$, $R^2 = 0.79$). In general, the orthographic intelligibility can be predicted well from the overall Levenshtein distances (the larger the distance, the more difficult it is to understand the related language). However, e.g. the BE–RU and MK–RU Levenshtein distances are equal (28.92%), but the intelligibility scores are different, e.g. BE–RU: 71.66% and MK–RU: 61.81%.

[8] The letter adaptation surprisal values are calculated on the basis of 120 Bulgarian stimuli and their Russian cognates.

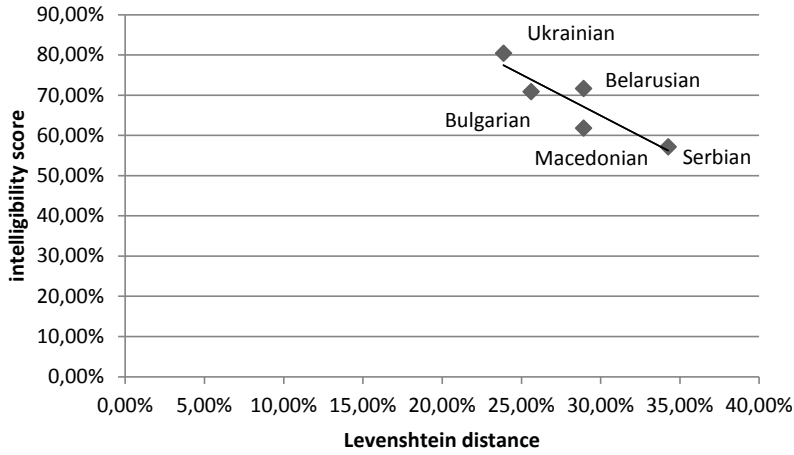

Fig. 1. The correlation between the mean Levenshtein distance and the average of the intelligibility score ($r = -0.89$)

4.2. Word adaptation surprisal and intelligibility score

The correlation between the mean word adaptation surprisal and the average of the intelligibility score (Fig. 2) shows that orthographic intelligibility can be predicted quite reliably ($r = -0.99$, $p < 0.001$, $R^2 = 0.98$) from the overall normalized word adaptation surprisal (the higher the value of surprisal the more difficult it is to comprehend the related language).

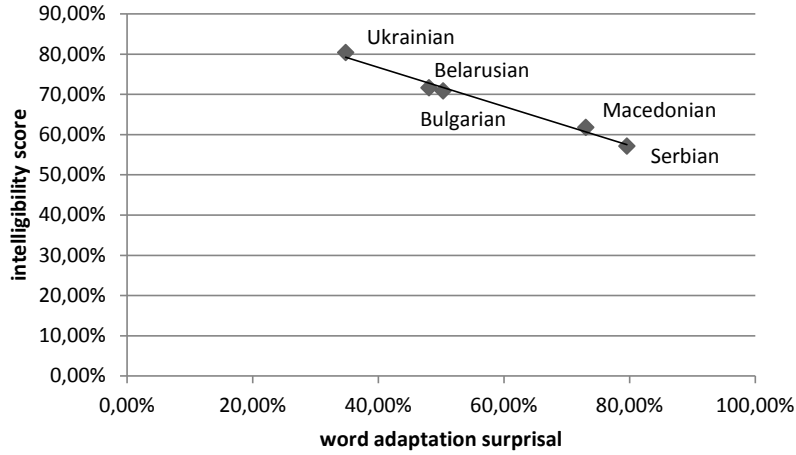

Fig. 2. The correlation between the mean word adaptation surprisal and the average of the intelligibility score ($r = -0.99$)

4.3. Word adaptation surprisal and total time

In addition to the mean intelligibility score we also correlate the normalized word adaptation surprisal with the total time spent per word incl. the initial time, typing time and final hesitation time (Fig. 3). The assumption was that people complete a translation task slower in the case of words with a higher surprisal value. In general, this assumption can be seen as confirmed ($r = 0.87$, $p = 0.05$, $R^2 = 0.76$). The scatterplot in Fig. 3 shows an interesting finding: Though the mean word adaptation surprisal between BE and RU (48.05%) is slightly lower than between BG and RU (50.30%) the total time between BE and RU (5.17 sec.) is slightly higher than between BG and RU (5 sec.). However, the mean value of intelligibility score between BE and RU (72.41%) is slightly higher than between BG and RU (70.88%) (Table 1). This means that participants were more successful in translating from BE into RU than from BG, but they spent slightly more time per word. The same situation concerns the MK–RU (5.72 sec.) and SR–RU (5.38 sec.) pairs. Though participants were more successful in translating from MK into RU than from SR, they spent more time per MK word.

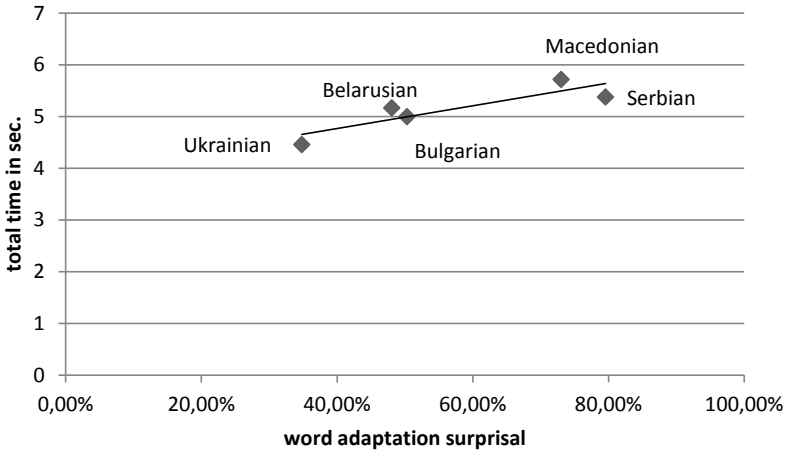

Fig. 3. The correlation between the mean word adaptation surprisal and the average of total reading and translation time (r = 0.87)

5. Conclusions and outlook

The aim of this article was to validate the normalized Levenshtein distance and the normalized word adaptation surprisal by investigating the degree to which these measurements of orthographic distances between Russian and five other Slavic languages, all using the Cyrillic alphabet, correlate with the mean intelligibility score and the total time obtained from Russian readers in web-based experiments. The results suggest that the word adaptation surprisal is a better predictor of orthographic intelligibility than the Levenshtein distance. We see this as a confirmation of the

usefulness of the surprisal method. However, the difference is not significant between Levenshtein distance and word adaptation surprisal ($r = -0.89$ versus $r = -0.99$ respectively). The assumption that people complete a translation task slowly on words whose surprisal value is higher could be confirmed only partly. A possible explanation for this is that participants need more time for the cognitive effort required to process the information and to complete the task correctly.

The Levenshtein distance has often been used as a predictor of mutual intelligibility between related languages in spoken semi-communication. We decided to add the method employing the notion of surprisal in order to test its applicability in our scenario. Both methods have their advantages and disadvantages. The Levenshtein algorithm measures the distance between two cognates within a language pair: nonidentical correspondences contribute to the orthographic distance, identical ones do not. Nonidentical correspondences are regarded as different and cost 1 unit. The surprisal method measures the complexity of a mapping, more precisely, how predictable the correspondence is in a language pair. The surprisal values of correspondences are different. However, they depend on frequency and distribution of correspondences in the particular cognate set. Furthermore, surprisal can be asymmetrical: the surprisal values between language A and language B are not necessarily the same as between language B and language A. This indicates an advantage of the surprisal-based method compared to the Levenshtein distance, which in its basic form is completely symmetrical.

Focusing on orthographic intelligibility, orthographic correspondences themselves, as well as their frequency, their nature or their position can be expected to perform well as predictors of intelligibility (Stenger et al. forthcoming). In future research, using the refined Levenshtein distance and adaptation surprisal models, we will analyze mismatched orthographic correspondences more precisely in order to investigate what kind of correspondences either facilitate or hinder intercomprehension as well as to get qualitatively significant results.

The results of our study are relevant for the areas of written intelligibility as well as of spoken semi-communication. The way in which we tested intelligibility may be relevant for research in other experimental disciplines within the humanities such as psycholinguistics and education science. Furthermore, the adaptation surprisal method can be used to also measure phonetic/phonological as well as morphological intelligibility between related languages.

Acknowledgement: We wish to thank Andrea Fischer and Varvara Obolonchykova for their support in the calculation of the Levenshtein distances and word adaptation surprisal (based on Python 2.7.6).

References

1. *Beijering K., Gooskens C., Heeringa W.* (2008), Predicting intelligibility and perceived linguistic distance by means of the Levenshtein algorithm, Linguistics in the Netherlands, pp. 13–24.
2. *Bidwell C. E.* (1963), Slavic Historical Phonology in Tabular Form, The Hague: Mouton & Co.
3. *Boston M. F., Halle J., Kliegl R., Patil U., Vasishth S.* (2008), Parsing costs as predictors of reading difficulty: an evaluation using the Potsdam Sentence Corpus, Journal of Eye Movement Research 2 (1), pp. 1–12.
4. *Braunmüller K., Zeevaert L.* (2001), Semicommunication, receptive multilingualism and related phenomena. A bibliographical overview, [Semikommunikation, rezeptive Mehrsprachigkeit und verwandte Phänomene. Eine bibliographische Bestandaufnahme], Working papers in multilingualism [Arbeiten zur Mehrsprachigekeit], Series B, No. 19, University Hamburg [Universität Hamburg].
5. *Carlton T. R.* (1991), Introduction to the Phonological History of the Slavic Languages, Slavica Publishers, Inc., Columbus, Ohio.
6. *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.
7. *Fischer A., Jágrová K., Stenger I., Avgustinova T., Klakow D., Marti R.* (2015). An Orthography Transformation Experiment with Czech-Polish and Bulgarian-Russian Parallel Word Sets, in Sharp B., Lubaszewski W., Delmonte R. (eds.), Natural Language Processing and Cognitive Science 2015 Proceedings, Libreria Editrice Cafoscarina, Venezia, pp. 115–126.
8. *Golubović J., Gooskens, C.* (2015), Mutual intelligibility between West and South Slavic languages, Russ Linguist 39, Springer, pp. 351–373.
9. *Gooskens C., Heeringa W.* (2004), Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data, Language Variation and Change, 16, Cambridge University Press, pp. 189–207.
10. *Gooskens C.* (2006), Linguistic and extra-linguistic predictors of Inter-Scandinavian intelligibility, in van de Weijer J. and Los B. (eds.), Linguistics in the Netherlands, 23, John Benjamins, Amsterdam, pp. 101–113.
11. *Gooskens C.* (2007), The contribution of linguistic factors to the intelligibility of closely related languages, Journal of Multilingual and Multicultural Development 28(6), pp. 445–467.
12. *Gooskens C.* (2013), Experimental methods for measuring intelligibility of closely related language varieties, in Bayley R., Cameron R., Lucas C. (eds.), Handbook of Sociolinguistics, Oxford University Press, Oxford, pp. 195–213.
13. *Halle J.* (2016), Information-theoretical Complexity Metrics, Language and Linguistics Compass 10/9, pp. 397–412.
14. *Haugen E.* (1966), Semicommunication: The language gap in Scandinavia, Sociological Inquiry 36, pp. 280–297.

15. *Heeringa W., Golubovic J., Gooskens C., Schüppert A., Swarte F., Voigt S.* (2013), Lexical and orthographic distances between Germanic, Romance and Slavic languages and their relationship to geographic distance, in Gooskens C. and van Bezoijen R. (eds.), Phonetics in Europe: Perception and Production, Peter Lang, Frankfurt a.M., pp. 99–137.
16. *Ivanova V. F.* (1991), Modern Russian orthography [Sovremennaja russkaja orfografija], Vyssšaja škola, Moskva.
17. *Kessler B.* (1995), Computational dialectology in Irish Gaelic, Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics, Dublin: EASCL, pp. 60–67.
18. *Kučera K.* (2009), The Orthographic Principles in the Slavic Languages: Phonetic/Phonological, in Kempgen S., Kosta P., Berger T., Gutschmidt K. (eds.), The Slavic Languages. An International Handbook of their Structure, their History and their Investigation, Vol. 1. Walter de Gruyter, Berlin & New York, pp. 70–76.
19. *Kürschner S., van Bezooijen R., Gooskens C.* (2008), Linguistic determinants of the intelligibility of Swedish words among Danes, International Journal of Humanities and Arts Computing 2(1/2), pp. 83–100.
20. *Levenshtein V. I.* (1965), Binary codes capable of correcting deletions, insertions, and reversals [Dvoičnye kody s ispravleniem udalenij, vstavok i zamen simvolov], Doklady of the Soviet Academy [Doklady Akademii Nauk SSSR], 1965, Vol. 163, No. 4, pp. 845–848.
21. *Marti R.* (2014), Historical Graphemics of the Slavic Languages: the Glagolitic and Cyrillic Writing Systems [Historische Graphematik des Slavischen: Glagolitische und kyrillische Schrift], in Kempgen S., Kosta P., Berger T., Gutschmidt K. (eds.), The Slavic Languages. An International Handbook of their Structure, their History and their Investigation, Vol. 2. Walter de Gruyter, Berlin & New York, pp. 1497–1514.
22. *Mel'ničuk O.S.* (1966), Introduction to comparatively-historical studies of Slavic languages [Vstup do porivnjal'no-istoryčnoho vyvčennja slov"jans'kich mov], Naukova dumka, Kiev.
23. *Möller R., Zeevaert L.* (2015), Investigating word recognition in intercomprehension: Methods and findings, Linguistics 2015 53(2), De Gruyter Mouton, Berlin, Munich & Boston, pp. 313–352.
24. *Musatov V. N.* (2012), Russian language. Phonetics, Phonology, Orphoepy, Graphics, Orthography [Russkij jazyk. Fonetika, Fonologija, Orfoėpija, Grafika, Orfografija], Izdatel'stvo 'Flinta', Moskva.
25. *Sgall P.* (2006), Towards a Theory of Phonemic Orthography, in Sgall P. (ed.), Language in its multifarious aspects, Karolinum Press Charles University, pp. 430–452.
26. *Shannon C. E.* (1948), A mathematical theory of communication, Bell System Technical Journal 27 (379–423), pp. 623–656.
27. *Smith N. J., Levy R.* (2013), The effect of word predictability on reading time is logarithmic, Cognition 128(3), pp. 302–319.
28. *Stenger I.* (2016), How Reading Intercomprehension Works among Slavic Languages with Cyrillic Script, in Köllner M., Ziai R. (eds.), Proceedings of the ESSLLI 2016, pp. 30–42, available at: http://esslli2016.unibz.it/wp-content/uploads/2016/09/esslli-stus-2016-proceedings.pdf

29. *Stenger I., Jágrová K., Fischer A., Avgustinova T.* (Forthcoming), "Reading Polish with Czech Eyes" or "How Russian Can a Bulgarian Text Be?": Orthographic Differences as an Experimental Variable in Slavic Intercomprehension, in Kosta P., Radeva-Bork T. (eds.), (preliminary title) Current developments in Slavic Linguistics. Twenty years after, Peter Lang.
30. *Tribus M.* (1961), Thermostatics and thermodynamics, D. van Nostrand Company.
31. *Vanhove J.* (2015), The Early Learning of Interlingual Correspondences Rules in Receptive Multilingualism. International Journal of Bilingualism, available at: http://homeweb.unifr.ch/VanhoveJ/Pub/papers/Vanhove_Correspondence-Rules.pdf. Vanhove J., Berthele R. (2015), The lifespan development of cognate guessing skills in an unknown related language, International Review of Applied Linguistics in Language Teaching 53(1), pp. 1–38.
32. *Vasmer M.* (1973), Etymological dictionary of the Russian language [Ėtimologičeskij slovar' russkogo jazyka], Progress, Moskva.
33. *Žuravlev A. F.* (ed.) (1974–2012), Etymological dictionary of the Slavic inherited lexikon. Proto-Slavic lexical stock [Ėtimologičeskij slovar' slavjanskich jazykov. Praslavjanskij leksičeskij fond], Vol. 1–37. Nauka, Moskva.