Nor Jnl Ling 40.2, 175–199 © Nordic Association of Linguists 2017 doi:10.1017/S0332586517000130

Stenger, Irina, Klára Jágrová, Andrea Fischer, Tania Avgustinova, Dietrich Klakow & Roland Marti.
2017. Modeling the impact of orthographic coding on Czech–Polish and
Bulgarian–Russian reading intercomprehension. *Nordic Journal of Linguistics*
40(2), 175–199.

# Modeling the impact of orthographic coding on Czech–Polish and Bulgarian–Russian reading intercomprehension

Irina Stenger, Klára Jágrová, Andrea Fischer,
Tania Avgustinova, Dietrich Klakow & Roland Marti

Focusing on orthography as a primary linguistic interface in every reading activity, the central research question we address here is how orthographic intelligibility can be measured and predicted between closely related languages. This paper presents methods and findings of modeling orthographic intelligibility in a reading intercomprehension scenario from the information-theoretic perspective. The focus of the study is on two Slavic language pairs: Czech–Polish (West Slavic, using the Latin script) and Bulgarian–Russian (South Slavic and East Slavic, respectively, using the Cyrillic script). In this article, we present computational methods for measuring orthographic distance and orthographic asymmetry by means of the Levenshtein algorithm, conditional entropy and adaptation surprisal method that are expected to predict the influence of orthography on mutual intelligibility in reading.

**Keywords:** entropy, Levenshtein distance, mutual intelligibility, reading intercomprehension, Slavic orthographic code, surprisal

*Collaborative Research Center (SFB) 1102: Information Density and Linguistic Encoding Project C4: INCOMSLAV Mutual Intelligibility and Surprisal in Slavic Intercomprehension Saarland University, Postfach 151150, 66041 Saarbrücken, Germany.*
*Irina Stenger: ira.stenger@mx.uni-saarland.de*
*Klára Jágrová: kjagrova@coli.uni-saarland.de*
*Andrea Fischer: andrea.fischer@lsv.uni-saarland.de*
*Tania Avgustinova: avgustinova@coli.uni-saarland.de*
*Dietrich Klakow: dietrich.klakow@lsv.uni-saarland.de*
*Roland Marti: rwmslav@mx.uni-saarland.de*

## 1. INTRODUCTION

Intercomprehension (Doyé 2005), receptive multilingualism (Braunmüller & Zeevaert 2001) or semi-communication (Haugen 1966) reveals a remarkable tolerance of human language processing mechanisms to deviations in linguistic encoding. It can be defined as the ability to understand unknown foreign languages,

https://doi.org/10.1017/S0332586517000130 Published online by Cambridge University Press

only due to their relatedness with at least one language that is already contained in the user's linguistic repertoire. This robust process is receptive in nature and does not include the ability to speak or write in the respective unknown language. The degree of intelligibility of an unknown but closely related language depends on both linguistic and extra-linguistic factors (Gooskens 2013). Research on receptive multilingualism has considered extra-linguistic factors such as subjective linguistic attitudes, exposure (Gooskens & van Bezooijen 2006, Schüppert & Gooskens 2012), fluid and crystallized intelligence, and age (Vanhove & Berthele 2015a). The literature shows different results of the relationship between extra-linguistic factors and intelligibility. While in some studies no correlation (e.g. between language attitudes and intelligibility) could be found (van Bezooijen & Gooskens 2007, Gooskens & Hilton 2013), other studies did find a positive correlation (Gooskens & van Bezooijen 2006, Schüppert, Hilton & Gooskens 2015).

In order to formally model the intercomprehension scenario in our project, we systematically investigate the linguistic determinants of successful information transmission across languages from the Slavic family, which are 'sufficiently similar and sufficiently different to provide an attractive research laboratory' (Corbett 1998:42).

Realistic scenarios of a reader fluent in one language trying to decipher a text written in an unknown but related language could involve, for example, a Czech native speaker attempting to read a Polish newspaper or a Russian native speaker being confronted with hotel information on Bulgarian websites. We assume that there should be a systematic way to successfully decode a given message depending on a quantifiable proximity between the language of the message and the decoder's own language, if the two are genetically related. In particular, the initial challenge is to decode the encountered signs into meaningful units in order to form a coherent understanding of the encoded information. Thus, orthography critically affects the success of transmitting information across languages: Unsuccessful or incomplete decoding of orthographic representations in an unknown language may lead to a situation where even cognate words become incomprehensible.

According to the *Handbook of Orthography and Literacy* (Joshi & Aaron 2006), orthography is the 'visual representation of language as conditioned by phonological, syntactical, morphological, and semantic features of the language' (Joshi & Aaron 2006:xiii). Indeed, orthography reflects a unique relationship to the characteristics of the respective language and a comparison of different orthographic systems involves various descriptive levels (phonetics/phonology, graphemics/graphotactics, morphology/morphosyntax, semantics) as well as historical, etymological, and sociolinguistic factors (e.g. spelling reforms) (Sgall 2006). In fact, the degree to which orthography provides a visual representation of (spoken) language varies between languages. Written and spoken language intercomprehension are traditionally considered different experiential domains, as pointed out by Gooskens (2013) and

https://doi.org/10.1017/S0332586517000130 Published online by Cambridge University Press

Möller & Zeevaert (2015), as well as in the critique of the representational idea by Kravchenko (2009). Yet, as mentioned by Gooskens (2013:1), methods for measuring the intelligibility of spoken language are applicable to written language too.

In our view, the following question is crucial for investigating the role of orthography in Slavic intercomprehension: How can orthographic intelligibility be measured and predicted between related languages? Our approach includes three computational methods: Levenshtein distance, conditional entropy and adaptation surprisal. Using them for modeling the orthographic mutual intelligibility should help us explain and predict how written stimuli are processed. As the Slavic languages use two different scripts, we investigate two representative language pairs here: Czech–Polish (West Slavic, using the Latin script; henceforth referred to as CS and PL) and Bulgarian–Russian (South Slavic and East Slavic, respectively, using the Cyrillic script; henceforth referred to as BG and RU).

This article is organized as follows. In Section 2 we give a short overview of the Slavic orthographic code as the linguistic basis for our modeling. In Section 3 we present the collected material (Section 3.1) and describe the computational methods by means of the Levenshtein algorithm (Section 3.2), conditional entropy (Section 3.3) and adaptation surprisal (Section 3.4). Finally, in Section 4, some general conclusions are drawn, including the discussion of advantages and disadvantages of the methods presented, and future work is presented.

## 2. SLAVIC ORTHOGRAPHIC CODE

In this section, we give a short overview of the orthographic code of the selected Slavic languages. In Section 2.1 we present the Latin and Cyrillic alphabets. In Section 2.2 we describe briefly the main orthographic principles of the respective orthographic systems. The way information is orthographically encoded in a word is influenced by the phonetic/phonological as well as the morphological structure of the language (Frost 2012). We omit the detailed presentation of the phonetic/phonological systems of the respective languages, because our focus lies on reading intercomprehension. However, we consider phonetics/phonology a possible factor in so far as readers will try to pronounce what they read in a process that is known as ‘inner speech’ (Harley 2008).

### *2.1 Latin and Cyrillic scripts*

Modern Slavic languages use two scripts, Latin and Cyrillic, which provide the Slavic orthographies with two rather different bases (Kučera 2009:72). An alphabet of a language constitutes the material basis of the respective orthography and can be described as a set of characters[1] (i.e. horizontally separable units; on the treatment of diacritics, see Section 3.2) that are used to constitute a written text in the language.

https://doi.org/10.1017/S0332586517000130 Published online by Cambridge University Press

| | | |
|---|---|---|
| CS | *a á b c č d ď e é ě f g h ch i í j k l m n ň o ó p q r ř s š t ť u ú ů v w x y ý z ž* | (42) |
| PL | *a ą b c ć e ę f g h i j k l ł m n ń o ó p r s ś t u w y z ż ź* | (31) |

**Table 1. Czech (CS) and Polish (PL) alphabets (vowels are underlined).**

| | | |
|---|---|---|
| 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) |
| Transliteration | *a b v g d e ž z i j k l m n o p r s t u f ch c č š št ă ' ju ja* | (30) |
| RU | *а б в г д е ё 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) |
| Transliteration | *a b v g d e ë ž z i j k l m n o p r s t u f ch c č š šč '' y ' ė ju ja* | (33) |

**Table 2. Bulgarian (BG) and Russian (RU) alphabets with transliteration according to the DIN (German Institute for Standardization) norm (vowels are underlined).**

The Latin and Cyrillic alphabets of the respective language pairs CS–PL and BG–RU are described in Sections 2.1.1 and 2.1.2. The characters of the alphabets are written in italics here in order to make it easier to compare these with examples. For a better visualization, we underlined the representations of vowels in Tables 1 and 2.

### *2.1.1 Czech and Polish alphabets*

Although both CS and PL use the Latin script (Table 1), they differ in their diacritical systems and the use of digraphs (combinations of two characters representing one phoneme). While PL makes extensive use of digraphs (which are not considered alphabetical characters), CS prefers diacritics. There are 42 characters in the CS alphabet: 14 representations of vowels (underlined) plus 28 representations of consonants. The PL alphabet consists of only 31 characters, of which nine are representations of vowels and 22 representations of consonants.

CS uses two basic diacritical signs: the *čárka* (acute accent symbol) ´ for marking a long vowel (plus the *kroužek* (circle) above *u*, i.e. *ů*, to designate a long *u* that historically goes back to *o*) and the *háček* (caron) ˇ, which was an innovation (originally in the shape of a superscript dot) (Comrie 1996a:664).

CS has a *háček* (i) for the palato-alveolar fricatives *š* /ʃ/, *ž* /ʒ/ and for the affricate *č* /ʧ/; (ii) for the fricative trill *ř* /r̝/; (iii) for palatal Ď /ɟ/, ň/Ň /ɲ/, and Ť /c/, except before *ě* and *i*; and (iv) on *ě* to palatalize the preceding consonant. In the case of *ď* and *ť*, a *klička* (apostrophe) is used as an alternation of the *háček* (Comrie 1996a, Skorvid 2005).

PL uses the digraphs *cz* for /t͡ʂ/, *sz* for /ʂ/, and either the diacritic *ż* (with the original CS dot) or the digraph *rz*, depending on etymology, for /ʐ/. PL has four different diacritical signs: (i) the *kreska* (acute accent symbol) ´, marking the vowel *ó* /u/ (which is likely to be mispronounced as /oː/ by Czech readers) and performing a similar function to the CS *háček* in *ć*, *ń*, *ś*, and *ź*; (ii) the *kropka* (overdot) used

https://doi.org/10.1017/S0332586517000130 Published online by Cambridge University Press

only in *ż* /ʐ/; (iii) the *ogonek*˛ in *ą and* ę; and (iv) the *kreska ukośna* (stroke) used in *ł*. Digraphs such as *ch*, *cz*, *dz*, *dź*, *dż*, *rz*, *sz* are not considered characters of the PL alphabet. The digraph *ch* for /x/ is used in both languages and is considered a character of the CS alphabet.

The CS characters *á*, *č*, *ď*, *é*, *ě*, *í*, *ň*, *ř*, *š*, *ť*, *ú*, *ů*, *v*, *ý*, *ž* as well as *q*, *v* and *x* are not part of the PL alphabet (the character *v* is only used in PL texts when it is part of a named entity or a foreign word), and the PL characters *ą*, *ć*, *ę*, *ł*, *ń*, *ś*, *ż*, *ź* do not exist in the CS alphabet. Still, these characters are expected to be legible for readers of the respective target language (i.e. by ignoring or substituting diacritical signs) and thus are not expected to impair reading intercomprehension heavily – especially when the actual phonetic representation is similar (e.g. *á* vs. *a*), although this fact might not be known to the reader.

### 2.1.2 Bulgarian and Russian alphabets

Comparing BG and RU alphabets (Table 2), we see that there are only slight differences. There are 30 characters in BG: eight representations of vowels (underlined), 21 representations of consonants and one sign without independent phonetic value (*ь*). RU has 33 characters: 10 representations of vowels, 21 representations of consonants and two signs without independent phonetic values: the so-called soft (*ь*) and hard (*ъ*) signs. Three characters of the RU alphabet do not occur in BG: *ы*, *э*, *ё*.[2]

The use of digraphs or diacritics is rare in the Cyrillic script. BG uses the digraph *дж* for [ʤ] and the digraph *дз* for [ʣ] (sounds not found in RU), for example, BG *джоб* (*džob*) ‘pocket’ and *дзифт* (*dzift*) ‘tar’. These two digraphs are not listed separately in the BG alphabet.

In an intercomprehension reading scenario, all BG characters seem to be familiar to readers who know the RU alphabet, but not vice versa. However, the nature as well as the use and pronunciation of a number of BG characters are not the same as in RU. The following BG characters have approximately the same value as in RU: *а*, *б*, *в*, *г*, *д*, *з*, *и*, *й*, *к*, *л*, *м*, *н*, *о*, *п*, *р*, *с*, *т*, *у*, *ф*, *х*, *ю*, *я* (Comrie 1996b, Cubberley 1996). The BG characters *ъ* and *щ* seem to be familiar to Russian readers, but the character-sound correspondences are different: *ъ* and *щ* in BG are pronounced [ɤ] and [ʃt] (Ternes & Vladimirova-Buhtz 2010), while in RU *ъ* has no phonetic, but an orthographic function, and *щ* is pronounced [ʃʲː] (Yanushevskaya & Bunčić 2015).

## 2.2 Orthographic principles as mechanisms of Slavic orthographic code

In an ideal phonographic[3] writing system, one graphic unit corresponds to a single sound unit and vice versa. However, the main lines of the evolution of the sound system in the Slavic languages are not always reflected in writing. Depending on the language and its history, orthographies were on the one hand adapted to sound changes in order to achieve harmony with the spoken language; on the other hand, orthographic systems were built upon the morphological[4] principle (e.g. with morphemes being written the same way, if possible, despite the differences in pronunciation) or historical/etymological principles (e.g. reflecting either the spelling of a language from which a word has been borrowed or an older state of the same language) (Kučera 2009).

Most, if not all, Slavic orthographies can be primarily described as phonemic, i.e. based on the correspondences between graphic and phonemic units (Kučera 2009). However, Kučera (2009:74) points out that '[t]he boundaries between the implementations of the phonological and morphological principles vary in different Slavic orthographies'. Thus, the Moscow Phonological School considers the phonemic principle to be the main principle in RU orthography in contrast to the St. Petersburg Phonological School that regards the morphological principle as the basic orthographic principle, depending on what is understood as a phoneme (Ivanova 1991, Musatov 2012). The BG orthography generally followed the RU model with a number of changes in the alphabets (Kempgen 2009, Marti 2014) and is defined as phonemic, although the morphological principle is also very important (Maslov 1981). The CS and PL orthographies can also be described as being primarily phonemic with some exceptions according to the morphological principle (Kučera 2009).

The particular orthographic principles used in the writing systems determine the mechanisms of the orthographic code using characters to represent speech – words and grammatical forms – in writing. The computational methods presented here (Levenshtein algorithm, conditional entropy and adaptation surprisal) should help us to discover a number of general patterns of interlingual orthographic encoding in order to model orthographic intelligibility in reading intercomprehension.

## 3. MODELING ORTHOGRAPHIC INTELLIGIBILITY

Successful reading intercomprehension is very closely linked to the amount of common vocabulary among genetically related languages (Möller & Zeevaert 2015). However, very often these cognates[5] (i.e. historically/etymologically related words) are not identical (e.g. orthographically). For instance, CS *štěstí*, PL *szczęście*, BG *щастие* (*štastie*), and RU *счастье* (*sčast'e*) 'happiness' (Vasmer 1973) are cognates, but spelled (and also pronounced) differently. A correct cognate recognition can be seen as a precondition of success in reading intercomprehension. In this section we present three methods of measuring orthographic intelligibility: the Levenshtein algorithm (Section 3.2) and conditional entropy (Section 3.3), and the resulting

https://doi.org/10.1017/S0332586517000130 Published online by Cambridge University Press

| | Total number of items | |
|---|---|---|
| Word list | Czech–Polish | Bulgarian–Russian |
| Swadesh list | 212 | 227 |
| Pan-Slavic list | 455 | 447 |
| Internationalism list | 262 | 261 |
| Frequency list | 100/100 | 100/100 |
| Dictionary | 80963 | 3705 |
| Phrasebook | ø | 5000 |

**Table 3. Sizes of word sets for applicability experiments with cross-lingual orthographic correspondences.**

adaptation surprisal (Section 3.4) as potential explanatory variables of orthographic intelligibility. These methods are applied to collections of parallel cognates in the selected language pairs.

### 3.1 Material

For our models, we make the following assumption: If a cognate pair in two languages L1 (native language) and L2 (stimulus language[6]) is orthographically (nearly) identical, then its similarity is determined by the chance that a L1 reader is able to decipher the characters of the L2 word. This implies that there exists a mapping between the characters of the two languages.

In the first step of our work, we set out to find these mappings. To this end, we created a maximally large collection of parallel cognate lists in the two language pairs (Table 3). We chose to use vocabulary lists instead of parallel sentences or texts in order to exclude the influence of other linguistic factors as much as possible. We compiled parallel cognate lists, consisting of internationalisms, lists of Pan-Slavic vocabulary (both adapted from the EuroComSlav website), and cognates from the Swadesh lists. All three lists were slightly modified in a way that non-cognates (i.e. CS–PL *mnoho–wiele* ‘many/much’; BG–RU *ние–мы* (*nie–my*) ‘we’) were removed or replaced by the orthographically closest cognates, if existing, in the respective other language (i.e. *mężczyzna* ‘man’ was substituted by *mąż* ‘husband’ in CS–PL *muž–mąż*; *звяр* (*zvjar*) ‘beast’ was added to its RU formal cognate *зверь* (*zver’*) ‘animal, beast’ for the BG–RU pair *звяр–зверь* (*zvjar–zver’*). The linguistic items in these lists belong to different parts of speech, mainly nouns, adjectives, and verbs.

For CS–PL we used a large, freely available online dictionary list (Kazojć 2010). For BG–RU, we included cognate pairs of a freely available online RU–BG phrasebook as well as a RU–BG dictionary from the website www.lexicons.ru.

We also included parallel lists of the 100 most frequent nouns (‘frequency lists’) from a previous study (Jágrová et al. 2017), in which readily available frequency lists of the languages (*Frequency Dictionaries of Bulgarian* (2011), available at http://dcl.bas.bg/en/tchestotni-retchnitsi-na-balgarskiya-ezik-2/, Křen (2010) for CS, the *Frequency List* (2016) for PL (Broda & Piasecki 2013), and Ljaševskaja & Šarov (2009) for RU), each based on large national corpora, were translated and compared. These lists were compiled for each source language separately and were then translated into the respective other language, always choosing the orthographically closest cognate if there was more than one option for translation (for details, see Jágrová et al. 2017). Thus, all four source lists are different.

To incorporate recognized theoretical findings of traditional Slavic linguistics, we decided to collect a systematic cross-linguistic rule set of orthographic correspondences from historical comparative studies (Bidwell 1963, Vasmer 1973, Žuravlev 1974–2012). These rules explain which L1 substrings correspond to which L2 substrings. We selected only those rules which we determined to have purely orthographic effects, excluding morphology and lexis as far as possible. This resulted in a compilation of diachronically based orthographic correspondences: 103 unique correlates for CS–PL and 77 for BG–RU, including one-to-one (matching) correspondences and mismatches consisting of single characters or of strings of characters (e.g. CS–PL: *a*:*a*, *á*:*ią*, *ě*:*ię*, *z*:*dz*, *hv*:*gw*, *lou*:*łu*, etc.; BG–RU: *б*:*б*, *т*:*ть*, *б*:*бл*, *ъ*:*у*, *и*:*ы*, *я*:*е*, *ла*:*оло*, etc.). Only correspondences which represented orthographic mismatches were automatically tested for applicability on parallel vocabulary lists. For the most part, the obtained automatic transformations could be seen as satisfactory for both language pairs. The outcomes of this applicability experiment can already be considered a measure for orthographic intelligibility: The fact that there are more cross-lingual correspondence rules in the CS–PL pair than in the BG–RU pair suggests that BG and RU are orthographically closer to each other than CS and PL. For more details concerning the transformation experiment see Fischer et al. (2015). From these lists (Table 3), we obtained 3404 CS–PL and 1182 BG–RU word pairs,[7] consisting of either identical words or words to which the cross-lingual correspondence rules apply.[8] These resulting word pairs were used for further calculations of conditional entropy and adaptation surprisal in Sections 3.3 and 3.4 below.[9] From these lists, selected word pairs containing such cross-lingual correspondences will be used as stimuli in our web-based experiments.

### 3.2 Levenshtein distance

Orthographic distances between cognates can be calculated by means of the Levenshtein algorithm (Levenshtein 1966). The latter has been developed for measuring linguistic distances between dialects (Heeringa et al. 2006) and successfully used to measure phonetic distances between Scandinavian language varieties (see Gooskens 2007).

| (a) | Language | 1 | 2 | 3 | 4 | 5 | 6 | 7 |  |  | Normalized LD |
|---|---|---|---|---|---|---|---|---|---|---|---|
|  | Czech | *m* | *l* | *a* | *d* | *o* | *s* | *t* |  |  |  |
|  | Polish | *m* | *ł* | *o* | *d* | *o* | *ś* | *ć* |  |  |  |
|  |  | 0 | 0.5 | 1 | 0 | 0 | 0.5 | 1 |  |  | 3/7 → 0.43 |
| (b) | Language | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | Normalized LD |
|  | Bulgarian | *м* |  | *л* | *а* | *д* | *о* | *с* | *т* |  |  |
|  | Russian | *м* | *о* | *л* | *о* | *д* | *о* | *с* | *т* | ь |  |
|  |  | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 3/9 → 0.33 |

**Table 4. Normalized LD (Levenshtein distance) for the word 'youth' in (a) Czech–Polish and (b) Bulgarian–Russian.**

Levenshtein distance (hereafter referred to as LD) is a string similarity measure which is equal to the number of operations needed to transform one string of characters into another through character insertions, deletions, and substitutions. These three operations are assigned weights. In the simplest form of the algorithm, all operations have the same cost. LD is a simple measure that (given reasonably similar alphabets) works out-of-the-box, is easy to apply, and requires no training data. In order to perform the alignment automatically, the algorithm is fed with character weight matrices for each language combination. Each matrix contains the complete alphabets of a language pair together with the costs assigned for every possible character alignment. 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. The actual edit costs are calculated after the automatic character alignment. Substitutions, insertions and deletions of different characters cost 1. In more sensitive versions, base and diacritic may be distinguished for a given character. For example, the base of CS *á* is *a*, and the diacritic is the *čárka*. It is not clear what weight exactly should be attributed to each of the components, but it is generally assumed that differences in the base will usually confuse the reader more than diacritical differences (Heeringa et al. 2013). Thus, if two characters have the same base but differ in diacritics, we assign them a substitution cost of 0.5, and no difference is made between the costs of the different diacritical signs (e.g. *čárka* vs. *háček*). For example, the following differences between CS and PL and between BG and RU are automatically calculated for the word 'youth' (see Table 4).

In our analysis we consider normalized LD 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 & Heeringa 2008). Per word pair, the non-normalized LD is computed by

| | Normalized LD | |
|---|---|---|
| Word list | CS-PL | BG-RU |
| Swadesh list | 42% | 33% |
| Pan-Slavic list | 39% | 31% |
| Internationalism list | 17% | 8% |
| Frequency list | 34% (PL for CS) | 14% (RU for BG) |
| | 35% (CS for PL) | 13% (BG for RU) |

LD = Levenshtein distance; CS = Czech; PL = Polish; BG = Bulgarian; RU = Russian

**Table 5. Average orthographic distance by means of the Levenshtein algorithm.**

means of the minimum number of operations needed to transform the string from one language into the other. The normalized LD is obtained by dividing the non-normalized distance by the total length of the alignment with the minimum costs (see Table 4).

With this method, we calculated the average orthographic distance within the respective Slavic language pairs on four of the parallel vocabulary lists: Swadesh list, Pan-Slavic list, Internationalism list (Stenger et al. in press) and the frequency lists (Jágrová et al. 2017). We found that throughout all lists the average orthographic distance between BG and RU is smaller than that between CS and PL (see Table 5), which confirms our previous findings from the applicability of rules experiment (Section 3.1above, Fischer et al. 2015).

There is a general assumption that the higher the LD, the more difficult it is to comprehend a given word in a translation task (see Gooskens 2007, Vanhove & Berthele 2015b, Vanhove 2016). This in particular means that, according to the average LD, CS and PL are mutually less intelligible than BG and RU as far as orthography is concerned.

Employing LD as a model of orthographic similarity means to impose no assumed costs for retaining symbols, while imposing costs on any non-alignable character and on every substituted character. While the use of matrices specifying edit costs of particular substitutions is an intuitive way of inputting prior knowledge about these substitutions, typically these costs are chosen ad-hoc on a subjective basis. Another disadvantage is that LD does not capture the systematicity, the complexity, and the asymmetry of the correspondences. Systematic correspondences, such as *v*:*w* in CS-PL are assigned a substitution cost of 1, just as any other character substitution. Also, LD does not capture asymmetries in language pairs. If, for instance, the RU vowel *а* always corresponds to *а* for a Bulgarian reader, but in the other direction, BG *а* can correspond to *а*, *о* or *я* for a Russian reader, then we would desire a measure of linguistic distance to reflect both this difference in adaptation possibilities and the uncertainty involved in transforming *а*. Heeringa et al. (2013) and Jágrová et al.

(2017) came up with asymmetric orthographic distance measures calculated by means of the Levenshtein algorithm, but this asymmetry is due to the fact that there were different source lists for each direction of reading in both studies (see Frequency list in Table 5).

The next two measurements for orthographic intelligibility that we discuss are conditional entropy and adaptation surprisal. Conditional entropy and surprisal stem from the field of information theory (Shannon 1948) and conditional entropy has previously been applied to modeling asymmetric linguistic divergences in Frinsel et al. (2015) and Moberg et al. (2006). To our knowledge, adaptation surprisal has not been calculated in previous research on receptive multilingualism.

## 3.3 Conditional entropy

With the Levenshtein method we confirmed that BG and RU are orthographically closer to each other than CS and PL (see Section 3.2). LD is a mathematical distance and thus completely symmetric. It cannot capture any asymmetries between related languages (Moberg et al. 2006, Frinsel et al. 2015), but asymmetries are something we do expect: Previous research has shown that speakers of two (closely) related languages do not always understand each other to the same degree (Budovičová 1987, Jensen 1989, Gooskens & van Bezooijen 2013a). In this section we present conditional entropy as a potential explanatory variable of orthographic asymmetries.

### 3.3.1 Surprisal as prefix-free code length

In order to explain conditional entropy, we begin with the notion of prefix-free code lengths. Let us assume that we have a string of characters we wish to communicate to a receiver who neither knows the identity nor the order of the character sequence. Let us assume that for this communication, we have only two symbols available, from which we need to construct code words for our characters: 0 and 1. Let us further assume that we have knowledge on the general distribution of characters, i.e. we know for example that 75% of all characters will be *a* and 25% will be *b*. We denote the proportion, or probability, of a specific character *c* occurring as *p(c)*. Then the theoretically-optimal length of the code word for each character *c* is given as $-log_2(p(c))$ bits. In practice, we cannot use fractions of 0s and 1s, so this is a theoretical lower limit.

Log probabilities are also often called ‘surprisal’ values (see Section 3.4 below) for intuitive reasons: The higher the probability of a specific character *c*, the smaller its code length; and the lower the probability of *c*, the greater its code length. These code lengths appear to correlate naturally with some measures of cognitive processing complexity (Smith & Levy 2013), and are called ‘surprisal’ in these contexts.

https://doi.org/10.1017/S0332586517000130 Published online by Cambridge University Press

| Language | 1 | 2 | 3 | 4 | 5 | 6 | 7 | | |
|---|---|---|---|---|---|---|---|---|---|
| CS | *m* | *l* | *a* | *d* | *o* | *s* | *t* | | |
| PL | *m* | *ł* | *o* | *d* | *o* | *ś* | *ć* | | |
| Reader | | | | | | | | | |
| CS reader | 1:1 | 1:1 | 1:2 | 1:1 | 1:2 | 1:1 | 1:1 | | |
| PL reader | 1:1 | 1:1 | 1:1 | 1:1 | 1:1 | 1:1 | 1:1 | | |
| Language | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| BG | *м* | | *л* | *а* | *д* | *о* | *с* | *т* | |
| RU | *м* | *о* | *л* | *о* | *д* | *о* | *с* | *т* | *ь* |
| Reader | | | | | | | | | |
| BG reader | 1:1 | 1:3 | 1:1 | 1:3 | 1:1 | 1:3 | 1:1 | 1:1 | 1:1 |
| RU reader | 1:1 | 1:2 | 1:1 | 1:1 | 1:1 | 1:1 | 1:1 | 1:1 | 1:2 |

CS = Czech; PL = Polish; BG = Bulgarian; RU = Russian

**Table 6. Entropy calculation of a corpus of one word pair for the word 'youth' in Czech–Polish and Bulgarian–Russian.**

### *3.3.2 Entropy of distributions*

The ENTROPY of a distribution is defined as the weighted average of the surprisal values for this distribution, i.e.

(1) $$H\left(X\right) = -\sum_{x \in X} P\left(X = x\right) log_2 P\left(X = x\right)$$

Entropy is the average length of a code word when encoding data sets whose compositions follow the given distribution exactly. As such, it is commonly considered a measure of the information content in that distribution.

In an intercomprehension setting, the task is to transform, or adapt, one unknown word into a known one. Thus, we condition our character probabilities on the corresponding characters seen in the stimulus word. The formula for this is:

(2) $$P\left(L1 = c1 | L2 = c2\right) = \frac{count(L1{=}c1 \wedge L2{=}c2)}{count(L2{=}c2)}$$

L1 – native language, c1 – character of L1

L2 – stimulus language, c2 – character of L2

In this way, we get different entropy for each of the characters in the alphabet of the (foreign) language of the stimulus. This satisfies our expectation of asymmetry, since transforming characters between languages is not necessarily symmetric. To illustrate this, Table 6 above shows two sample adaptations of the word 'youth' between our languages, which we shall use to illustrate conditional entropy.

Table 6 shows an alignment of one cognate pair in CS–PL and BG–RU, character by character. In order to obtain measurements which are based on linguistically motivated alignments, the characters are aligned here in the same way as for the Levenshtein algorithm – a character representing a vowel may only correspond to a character representing a vowel, and a consonant character only to a consonant character. We assume that a non-native reader would map the characters in the same way.

The word ‘youth’ produces seven character pairs in CS–PL and nine character pairs in BG–RU. From a CS perspective, the mappings would be *m* with *m*, *ł* with *l*, *o* with *a* and so forth. In the examples in Table 6, most of the mappings constitute 1:1 correspondences in both directions of reading. This means that each character in one language corresponds only to one particular character in the other language. In these cases it means that the entropy for these alignments is 0. However, there are two exceptions in CS–PL and five exceptions in BG–RU.

From a PL perspective, all adaptations are unambiguous. Reading the CS word, the Polish reader sees an *a* corresponding only to an *o* (alignment 3) and an *o* corresponding only to an *o* (alignment 5). In these cases, both $p(o|a)$ and $p(o|o)$ are 1.0 as are all the other adaptation probabilities. Thus, the conditional entropies of all characters, including *o* and *a*, is 0, which gives a total conditional entropy of 0 for this direction. However, a Czech reader finds an *o* twice (alignment 3 and 5), which corresponds to an *o* (alignment 3) and to an *a* (alignment 5). Therefore, both $p(o|o)$ and $p(a|o)$ is 0.5, the entropy of *o* is $-(1/2^{*}\log_2(0.5) + 1/2^{*}\log_2(0.5))/2 = 1$, and the overall entropy for this direction is $(2^{*}1 + 5^{*}0)/7 \approx 0.286$. This result is asymmetric: The entropy of PL for Czech readers is higher than the entropy of CS for Polish readers. In other words, a Czech reader has to deal with a higher amount of uncertainty and is expected to find it more difficult to read the PL word ‘youth’ than a Polish reader deciphering the respective CS cognate.

In the BG–RU alignment, we have five exceptions. From a BG perspective, the Bulgarian reader sees an *o* three times (alignment 2, 4 and 6), which corresponds to nothing (alignment 2), to an *a* (alignment 4), and to an *o* (alignment 6). In this case $p(\mathrm{x}|\mathrm{y})$ is 0.33 for each of these alignments and the overall entropy is about 0.579. The Russian reader sees twice nothing in the BG word (alignment 2 and 9), which corresponds to an *o* (alignment 2) and to an *ь* (alignment 9). In this case $p(\mathrm{x}|\mathrm{y})$ is 0.5 for each of these alignments and the entropy is about 0.222. In this example, both Bulgarian and Russian readers have to deal with some uncertainty when reading one foreign word. However, the amount of entropy for Bulgarian readers is higher than for Russian readers and the characters which cause the entropy are different.

### 3.3.3 Conditional entropy of CS–PL and RU–BG

An overview of some conditional entropy values for our language pairs is given in Table 7 below. For space reasons, we present only the entropies for the vowel characters.

Our entropy calculations based on 3404 CS–PL word pairs reveal that the entropy, e.g. of the CS *o* for Polish readers is 0.140681385 and that of the PL *o* for Czech

| CS entropy for PL readers | Characters | | PL entropy for CS readers |
|---|---|---|---|
| 0.078681926 | a | | 1.155763237 |
| | ▨ | ą | 1.143150615 |
| 0.195806274 | á | ▨ | |
| 0.907250209 | e | | 0.866205118 |
| | ▨ | ę | 0.7225543 |
| 0.07001205 | é | ▨ | |
| 0.741011679 | ě | ▨ | |
| 0.087827011 | i | | 0.736118208 |
| 1.692804843 | í | ▨ | |
| 0.140681385 | o | | 0.20901951 |
| 0 | ó | | 1.689482267 |
| 0.635610854 | u | | 0.01812382 |
| 0.918295834 | ú | ▨ | |
| 0 | ů | ▨ | |
| 0.085367347 | y | | 1.634948831 |
| 0 | ý | ▨ | |
| BG entropy for RU readers | Characters | | RU entropy for BG readers |
| 0.277333815 | a | | 0 |
| 0.593576323 | e | | 0.474721623 |
| | ▨ | ё | 0 |
| 0.421052502 | и | | 0 |
| 0 | o | | 0.871023454 |
| 0 | y | | 0.654664053 |
| | ▨ | ы | 0 |
| 1.457594511 | ъ | ▨ | |
| 0 | ю | | 0 |
| 0.828272583 | я | | 1.06700715 |

CS = Czech; PL = Polish; BG = Bulgarian; RU = Russian

**Table 7. Vowel character entropies for Czech–Polish and Bulgarian–Russian (shading in some of the cells indicates that this vowel character is not present in the respective alphabet).**

readers is 0.20901951. This means that the mapping of the PL *o* to possible CS characters is more complex than vice versa. More precisely, the PL *o* can map into 6 CS characters (*o*, *e*, *a*, *á*, *ů*, or *í*) and the CS *o* can map only into two PL characters (*o* and *ó*) or to nothing. Of course, in an intercomprehension scenario a Czech reader or a Polish reader does not know these mappings and the respective probabilities. However, the assumption is that the measure of complexity of the mapping can be used as an indicator for the degree of intelligibility (Moberg et al. 2006), because it reflects the difficulties with which a reader is confronted in 'guessing' the correct correspondence. In this case Czech readers will have more uncertainty in adapting the PL *o* than Polish readers adapting the CS *o*.

Table 7 also shows that the entropy of the BG *o* for Russian readers is 0, but the entropy of the RU *o* for Bulgarian readers is 0.871023454 (the calculation is based on 1182 word pairs). This means that Bulgarian readers have to deal with some uncertainty in transforming the RU *o*, while Russian readers should have no difficulties with the BG *o*.

Overall, the conditional entropy between language L1 and L2 is given as the weighted averages of all character entropies:

(3) $$H(L1|L2) = \sum_{c2 \in L2} P(L2 = c2)\mathrm{H}(L1|L2 = c2) = -\sum_{c1 \in L1, c2 \in L2} P\,(L1{=}c1 \wedge L2{=}c2)\log_2 P(L1{=}c1|L2{=}c2)$$

L1 – native language, c1 – character of L1
L2 – stimulus language, c2 – character of L2

For our languages, full conditional entropy gives us the following entropy values: 0.45 for the PL to CS transformation and 0.37 for the CS to PL transformation. Thus, a Czech reader may have more difficulties reading PL than a Polish reader reading CS. For the BG–RU language pair the difference in the entropies is very small for both directions: 0.16 for the BG to RU transformation and 0.17 for the RU to BG transformation, with a very small amount of asymmetry of 0.01. This calculation predicts that speakers of BG reading RU words are facing only a slightly higher amount of uncertainty than speakers of RU reading BG words. The low orthographic entropy for BG and RU can be explained by a higher number of orthographically identical (848) vs. orthographically non-identical items (335) in the material for these two languages.

### 3.4 Adaptation surprisal: A refined measure

In addition to conditional entropy 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 values are given in bits and depend heavily on the probability distribution used. In our setting, we get CHARACTER ADAPTATION SURPRISAL (CAS) from our character adaptation probabilities (see Section 3.3.2 above). The character adaptation surprisal is calculated with the following formula:

(4) $$\mathit{surprisal}(L1 = c1|L2 = c2) = -\log_2 P(L1 = c1|L2 = c2)$$

L1 – native language, c1 – character of L1
L2 – stimulus language, c2 – character of L2

CAS values allow us to quantify the unexpectedness both of individual character correspondences and of the whole cognate pair. We can compute full WORD ADAPTATION SURPRISAL (WAS) by summing up the CAS values of the contained characters (Table 8). This gives a quantification of the overall (un)expectedness of the correct cognate.

https://doi.org/10.1017/S0332586517000130 Published online by Cambridge University Press

| Language | 1 | 2 | 3 | 4 | 5 | 6 | 7 | | | Normalized WAS |
|---|---|---|---|---|---|---|---|---|---|---|
| CS | *m* | *l* | *a* | *d* | *o* | *s* | *t* | | | |
| PL | *m* | *ł* | *o* | *d* | *o* | *ś* | *ć* | | | |
| Reader | | | | | | | | | | |
| CS reader | 0.001 | 0 | 6.724 | 0 | 0.036 | 0 | 0.002 | | | 6.78/7 → 0.97 |
| PL reader | 0 | 2.229 | 6.984 | 0 | 0.026 | 2.968 | 1.700 | | | 13.9/7 → 1.99 |
| Language | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | Normalized WAS |
| BG | *м* | | *л* | *а* | *д* | *о* | *с* | *т* | | |
| RU | *м* | *о* | *л* | *о* | *д* | *о* | *с* | *т* | *ь* | |
| Reader | | | | | | | | | | |
| BG reader | 0 | 4.960 | 0.009 | 4.960 | 0 | 0.225 | 0 | 0 | 0.029 | 10.18/9 → 1.13 |
| RU reader | 0 | 2.883 | 0 | 4.985 | 0 | 2.968 | 0 | 0 | 0.288 | 8.16/9 → 0.91 |

WAS = word adaptation surprisal; CS = Czech; PL = Polish; BG = Bulgarian; RU = Russian

**Table 8. Adaptation surprisal values for the word ‘youth’ in Czech–Polish and Bulgarian–Russian.**

For example, the CS–PL cognate pair *mladost–młodość* 'youth' contains the following correspondences for Czech readers: *m*:*m* (surprisal: 0.001), *ł*:*l* (surprisal: 0.0), *o*:*a* (surprisal: 6.724), *d*:*d* (surprisal: 0.0), *o*:*o* (surprisal: 0.036), *ś*:*s* (surprisal: 0.0), *ć*:*t* (surprisal: 0.002). Thus, *mladost–młodość* 'youth' gets a WAS value of 6.78 for Czech readers or 13.9 for Polish readers (Table 8) with a predicted advantage for Czech readers. As in the case with the LD, we also normalize the full WAS for the cognate pair 'youth' in the selected languages. For instance, 6.78/7 = 0.97 for Czech readers and 13.9/7 = 1.99 for Polish readers or 10.18/9 = 1.13 for Bulgarian readers and 8.16/9 = 0.91 for Russian readers.

We are now in position to meaningfully compare the individual CAS values and the full WAS values to the results of our web-based experiments revealing respective intercomprehension scores. The CAS should help us to predict and explain the effect of mismatched orthographic correspondences on cognate recognition in a reading intercomprehension scenario. We assume that the smaller the adaptation surprisal, the easier it is to guess the correct orthographic correspondence. However, it must be mentioned here that identical orthographic correspondences may still have a small surprisal value, for example, from a CS perspective the correspondence *o*:*o* has a surprisal value of 0.036 bits (Table 8). In our setting, this can be explained by the fact that the PL *o* could become *o* (2268 times), *e* (23 times), *a* (22 times), *á* (9 times), *ů* (2 times) or *í* (1 time) in CS. On the other hand, non-identical orthographic correspondences may have no surprisal value at all, for example from a CS perspective the correspondence *ł*:*l* has the surprisal of 0.0 (Table 8). This means that the PL *ł* always corresponds to CS *l* (252 times) in our training corpus.

We can also calculate the average value of the normalized WAS between stimuli of selected languages and their cognates in L1. The assumption is that the higher the mean normalized WAS, the more difficult it is to comprehend a related language. This assumption was confirmed by free translation task experiments of written East Slavic (Ukrainian, Belarusian) and South Slavic (Bulgarian, Macedonian, Serbian) languages for Russian subjects (Stenger, Avgustinova & Marti 2017).

In our training corpora we get the following average values of the normalized WAS within the respective language pairs: 0.48 for the PL to CS transformation and 0.40 for the CS to PL transformation, as well as 0.18 for the BG to RU transformation and 0.19 for the RU to BG transformation. This calculation results let us conclude again that CS and PL are orthographically less mutually intelligible in comprehension of written stimuli than BG and RU. In addition, the WAS calculation allows us to predict that a Czech reader may have more difficulties reading PL stimuli than a Polish reader reading CS. For the BG–RU language pair, the difference in the WAS is very small for both directions, with a small predicted advantage of 0.01 for Russian readers of BG stimuli.

## 4. DISCUSSION AND OUTLOOK

Previous research in intercomprehension has shown that closely related languages may be differently distant from each other. In this article we presented computational methods for measuring and predicting the orthographic mutual intelligibility of Slavic languages by means of LD, conditional entropy and adaptation surprisal. All three methods have their advantages and disadvantages as summarized in the following discussion.

LD as a measure of orthographic similarity of cognate pairs is considered a fairly good predictor of overall intelligibility in spoken semi-communication as well as in reading intercomprehension (Gooskens 2007, Beijering et al. 2008, Kürschner, van Bezooijen & Gooskens 2008, Vanhove & Berthele 2015b, Vanhove 2016). The simplest versions of this method are based on a distance measure in which orthographic overlap is binary: Non-identical characters contribute to orthographic distance and identical characters do not. In this way, for instance, the BG–RU pair *а*:*о* counts as different to the same degree as *а*:*я*, namely 1 unit.

The conditional entropy can reflect the difficulties humans encounter when mapping one orthographic system on another. However, it measures only the regularity of correspondences and does not measure the similarity between two languages as, for example, the LD does (Frinsel et al. 2015). Nonetheless, such regularity could very well have an effect on the reader's ability to understand a related language (ibid.). The full conditional entropy between two languages, H(L1|L2), allows for measuring the complexity of the ENTIRE mapping between these two languages. It reflects both the frequency and regularity of correspondences between them and can reveal asymmetries in overall adaptation difficulties, as has been demonstrated for Swedish and Danish by Frinsel et al. (2015) and for Danish, Swedish and Norwegian by Moberg et al. (2006). The underlying hypothesis is that high predictability improves intelligibility, and therefore a low entropy value should correspond to a high intelligibility score.

The adaptation surprisal method measures the complexity of a mapping, in particular, how predictable the particular correspondence in a language pair is. The surprisal values of correspondences are indeed different. However, they depend on their frequency and distribution in the particular cognate set. Surprisal can be as asymmetric as entropy: 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 LD, which in its basic form is completely symmetrical. Furthermore, the adaptation surprisal method can be used to measure phonetic/phonological as well as morphological intelligibility between related languages.

The exact predictive potential of the LD, conditional entropy and adaptation surprisal methods for Slavic intercomprehension is an open question, which we

https://doi.org/10.1017/S0332586517000130 Published online by Cambridge University Press

explore in web-based experiments.[10] A human may not succeed in transforming the substrings of words of an unknown language into the correct corresponding forms in their L1. The goal of our efforts is to design a metric of linguistic distance which takes into account the human decoding process. On the basis of the data collected from our experiments, we intend to improve the Levenshtein algorithm by optimizing the cost of each operation, including that of insertions and deletions. We have observed strong asymmetries of classic averaged Levenshtein costs in each of the language pairs. Our findings for CS and PL orthographic distance confirm those of Heeringa et al. (2013). For BG and RU we discovered that BG adjectives in the masculine form which employ zero endings are a major factor of asymmetric Levenshtein values. Furthermore, normalizing LD asymmetrically, i.e. dividing edit costs by the length of a word, leads to lower values for source languages that tend to use shorter words. If we assume that a lower Levenshtein cost implies higher intelligibility of a word, readers should be more successful when comprehension requires dropping characters rather than adding them.

In this contribution we calculated orthographic distances and asymmetries by means of LD, conditional entropy and adaptation surprisal based on large corpus data. There are a number of arguments why LD and conditional entropy scores should be calculated on a large cognate set rather than only on the stimuli that might be presented to readers in intercomprehension experiments. Van Heuven, Gooskens & van Bezooijen (2015:132) point out that

> distance measures become more stable and correlate better with mutual intelligibility scores ... as the materials the distances are computed on get larger ..., but this relation may well be different if the distance measures are specifically based on the stimulus materials used in the intelligibility tests.

According to Moberg et al. (2006), at least 800 word pairs are needed to reach stable entropy measures. Not only do we expect more representative entropy and adaptation surprisal measures by using large corpus data, but it also allows us to choose appropriate stimulus material for intelligibility tests, taking into consideration the complexity of the mapping of correspondences as an experimental variable. For the upcoming experiments we hypothesize that the most regular and frequent correspondence rules (those with little adaptation surprisal values) should be more transparent for readers. Of course, the calculations of conditional entropy and adaptation surprisal on a greater sample of cognate pairs than an actual stimulus set in future experiments may also represent a greater exposure of a reader to correspondences in the other language. An important argument for using conditional entropy and adaptation surprisal instead of Levenshtein distance is that the first two can model asymmetric intelligibility. In future research, we will refine the entropy and surprisal method in several ways and enhance the experiments by measurements

https://doi.org/10.1017/S0332586517000130 Published online by Cambridge University Press

based on bigrams or trigrams, focusing on the position of correspondences in cognates as well as on their nature.

While our upcoming analyses will shed light on the degrees of transparency of the different orthographic correspondences, it remains to be seen which other factors are likely to play a role in mutual intelligibility. Thus, for example, word frequency, word length, neighborhood density[11] and different orthographic correspondences themselves (their nature, frequency and position) can influence the correct recognition of cognate elements between related languages. Vanhove & Berthele (2015b) demonstrated that the frequency of the word in question in the reader's language is a reliable predictor for its intelligibility. Word length was shown to influence intelligibility of individual words, too (Kürschner et al. 2008). Kürschner et al. (2008) find that longer words are recognized more easily than shorter words. Gooskens (2013) points out that words with a high neighborhood density are often more difficult to recognize than those with few competitors.

The combination of computational methods of Levenshtein, conditional entropy and adaptation surprisal with such factors as word frequency, word length, neighborhood density, as well as the orthographic correspondences themselves, will help us anticipate, analyze and evaluate the results of our web-based experiments with speakers of Slavic languages. Once our human data collection is completed, we will construct regression models as predictors of linguistic distance. Such models allow to analyze the effects of e.g. character context, within-word position, representations of consonants vs. representations of vowels, and dialects or archaic terms individually, and thus to gain insight into the importance of each of the factors considered. Isolated investigations at the orthographic level reach their limits as soon as we examine continuous texts. The inevitable next steps of our work will be devoted to analyzing the influence of grammatical differences.

## ACKNOWLEDGEMENTS

This study was carried out within the framework of the project INCOMSLAV (Mutual Intelligibility and Surprisal in Slavic Intercomprehension) funded by Deutsche Forschungsgemeinschaft (DFG) under grant SFB 1102: Information Density and Linguistic Encoding. We thank three anonymous reviewers for their useful comments on an earlier version of this article.

## NOTES

1. See Sgall (2006) for a discussion and definition of the alphabet and its basic units.
2. The character *ё* is generally used in dictionaries and schoolbooks only.
3. Alphabetic according to Daniels (2001), segmental phonographic according to Sampson (1985).

https://doi.org/10.1017/S0332586517000130 Published online by Cambridge University Press

[4] Some linguists prefer the term ‘morphemic’.

[5] ‘Cognates are historically related word pairs that still bear the same meaning in both languages’ (Kürschner et al. 2008:86). In this work we are investigating cognates not only including shared inherited words from Proto-Slavic, but also shared loans, for example, internationalisms.

[6] In our contribution, L2 is defined as stimulus language and not as second foreign language, as it would be in language learning contexts.

[7] We make the resource available under: http://www.coli.uni-saarland.de/~tania/incomslav.html. An access code can be requested from the authors.

[8] Duplicates were removed.

[9] We make the computer code available at http://www.coli.uni-saarland.de/~tania/incomslav.html.

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

[11] Kürschner et al. (2008:90) define neighbors linguistically as ‘word forms that are very similar to the stimulus’. Thus, neighborhood density is the amount of similar words that might interfere in the process of cognate identification.

## CORPORA

*Czech National Corpus*: Srovnávací frekvenčni seznamy. 2010. http://ucnk.ff.cuni.cz/srovnani10.php (accessed 1 January 2016).

*Frequency Dictionaries of Bulgarian*. 2011. Department of Computational Linguistics, Bulgarian Academy of Sciences. http://dcl.bas.bg/en/tchestotni-retchnitsi-na-balgarskiya-ezik-2 (accessed 5 April 2016).

*Internationalism list*. http://www.eurocomslav.de/kurs/iwslav.htm (accessed 11 July 2015).

*Lista frekwencyjna* [Frequency list]. 2016. Grupa Technologii Językowych G4.19 Politechniki Wrocławskiej. http://www.nlp.pwr.wroc.pl/narzedzia-i-zasoby/zasoby/lista-frekwencyjna (accessed 8 September 2016).

*Novyj Častotnyj Slovar’ Russkoj Leksiki* [New frequency dictionary of Russian vocabulary] (*NČS*). 2009. Ol’ga N. Ljaševskaja & Sergej A. Šarov. http://dict.ruslang.ru/freq.php (accessed 5 April 2016).

*Otwarty słownik czesko-polski* [Open Czech–Polish dictionary] *V.03.2010 (c)*. 2010. J. Kazojć. http://www.slowniki.org.pl/czesko-polski.pdf (accessed 22 April 2015).

*Pan-Slavic list*. http://www.eurocomslav.de/kurs/pwslav.htm (accessed 11 July 2015).

*Russko-bolgarskij Razgovornik* [Russian–Bulgarian phrase book]. Izdatel’stvo ‘Chermes’. https://drive.google.com/file/d/0B3ZsKnxnxCJNSUd3RzNnOVYydlU/view (accessed 15 April 2016).

*Russko-bolgarskij Slovar’* [Russian–Bulgarian dictionary]. http://www.lexicons.ru/modern/b/bulgarian/index.html (accessed 5 April 2016).

*Swadesh-list*. http://en.wiktionary.org/wiki/Appendix:Swadesh_lists_for_Slavic_languages (accessed 11 July 2015).

## REFERENCES

Beijering, Katrin, Charlotte Gooskens & Wilbert Heeringa. 2008. Predicting intelligibility and perceived linguistic distance by means of the Levenshtein algorithm. In Marjo van

https://doi.org/10.1017/S0332586517000130 Published online by Cambridge University Press

Koppen & Bert Botma (eds.), *Linguistics in the Netherlands 2008*, 13–24. Amsterdam: John Benjamins.

Bidwell, Charles E. 1963. *Slavic Historical Phonology in Tabular Form*. The Hague: Mouton & Co.

Braunmüller, Kurt & Ludger Zeevaert, L. 2001. *Semikommunikation, rezeptive Mehrsprachigkeit und verwandte Phänomene. Eine bibliographische Bestandaufnahme* (Arbeiten zur Mehrsprachigkeit, Folge B, 19). Hamburg: Universität Hamburg.

Broda, Bartosz & Maciej Piasecki. 2013. Parallel, massive processing in SuperMatrix: A general tool for distributional semantic analysis of corpora. *International Journal of Data Mining, Modelling and Management* 5(1), 1–19.

Budovičová, Viera. 1987. Literary languages in contact: A sociolinguistic approach to the relation between Slovak and Czech today. In Jan Chloupek & Jiří Nekvapil (eds.), *Reader in Czech Sociolinguistics*, 156–175. Amsterdam: John Benjamins.

Comrie, Bernard. 1996a. Adaptations of the Roman alphabet: Languages of Eastern and Southern Europe. In Daniels & Bright (eds.), 663–675.

Comrie, Bernard. 1996b. Adaptations of the Cyrillic alphabet. In Daniels & Bright (eds.), 700–726.

Corbett, Greville G. 1998. Agreement in Slavic. Presented at the workshop Comparative Slavic Morphosyntax, Indiana University, Bloomington. [Position paper]

Cubberley, Paul. 1996. The Slavic alphabets. In Daniels & Bright (eds.), 346–355.

Daniels, Peter T. 2001. Writing systems. In Mark Aronoff & Janie Rees-Miller (eds.), *The Handbook of Linguistics*, 43–80. Malden, MA: Blackwell.

Daniels, Peter T. & William Bright (eds.). 1996. *The World's Writing Systems*. New York & Oxford: Oxford University Press.

Doyé, Peter. 2005. Intercomprehension. *Guide for the Development of Language Education Policies in Europe: From Linguistic Diversity to Plurilingual Education* (Reference Studies). Strasbourg: Council of Europe.

Fischer, Andrea, Klára Jágrová, Irina Stenger, Tania Avgustinova, Dietrich Klakow & Roland Marti. 2015. An orthography transformation experiment with Czech–Polish and Bulgarian–Russian parallel word sets. In Bernadette Sharp, Wiesław Lubaszewski & Rodolfo Delmonte (eds.), *Natural Language Processing and Cognitive Science 2015 Proceedings*, 115–126. Venezia: Libreria Editrice Cafoscarina.

Frinsel, Felicity, Anne Kingma, Charlotte Gooskens & Femke Swarte. 2015. Predicting the asymmetric intelligibility between spoken Danish and Swedish using conditional entropy. *Tijdschrift voor Slandinavistiek* 34(2), 120–138.

Frost, Ram. 2012. Towards a universal model of reading. *Behavioral and Brain Sciences* 35(5), 263–329.

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

Gooskens, Charlotte. 2013. Experimental methods for measuring intelligibility of closely related language varieties. In Robert Bayley, Richard Cameron & Ceil Lucas (eds.), *Handbook of Sociolinguistics*, 195–213. Oxford: Oxford University Press.

Gooskens, Charlotte & Nanna H. Hilton. 2013. The effect of social factors on the comprehension of a closely related language. In Jani-Matti Tirkkonen & Esa Anttikoski (eds.), *Proceedings of the 24th Scandinavian Conference of Linguistics*, 201–210. Joensuu: University of Eastern Finland.

https://doi.org/10.1017/S0332586517000130 Published online by Cambridge University Press

Gooskens, Charlotte & Renée van Bezooijen. 2006. Mutual comprehensibility of written Afrikaans and Dutch: Symmetrical or asymmetrical? *Literary and Linguistic Computing* 21(4), 543–557.

Gooskens, Charlotte & Renée van Bezooijen. 2013a. Explaining Danish–Swedish asymmetric word intelligibility: An error analysis. In Gooskens & van Bezooijen (eds.), 59–82.

Gooskens, Charlotte & Renée van Bezooijen (eds.). 2013b. *Phonetics in Europe: Perception and Production*. Frankfurt a.M.: Peter Lang.

Hale, John. 2016. Information-theoretical complexity metrics. *Language and Linguistics Compass* 10(9), 397–412.

Harley, Trevor. 2008. *The Psychology of Language: From Data to Theory*. New York: Psychology Press.

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

Heeringa, Wilbert, Jelena Golubovic, Charlotte Gooskens, Anja Schüppert, Femke Swarte & Stefanie Voigt. 2013. Lexical and orthographic distances between Germanic, Romance and Slavic languages and their relationship to geographic distance. In Gooskens & van Bezooijen (eds.), 99–137.

Heeringa, Wilbert, Peter Kleiweg, Charlotte Gooskens & John Nerbonne. 2006. Evaluation of string distance algorithms for dialectology. In John Nerbonne & Erhard Hinrichs (eds.), *Linguistic Distances Workshop at the Joint Conference of International Committee on Computational Linguistics and the Association for Computational Linguistics*, 51–62. The Association for Computational Linguistics (ACL).

Ivanova, Vera F. 1991. *Sovremennaja russkaja orfografija* [Contemporary Russian orthography]. Moskva: Vysšaja škola.

Jágrová, Klára, Irina Stenger, Roland Marti & Tania Avgustinova. 2017. Lexical and orthographic distances between Bulgarian, Czech, Polish, and Russian: A comparative analysis of the most frequent nouns. In Joseph Emonds & Markéta Janebová (eds.), *Language Use and Linguistic Structure: Proceedings of the Olomouc Linguistics Colloquium 2016*, 401–416. Olomouc: Palacký University.

Jensen, John B. 1989. On the mutual intelligibility of Spanish and Portuguese. *Hispania* 72(4), 848–852.

Joshi, R. Malatesha & P. G. Aaron. 2006. Introduction to the volume. In R. Malatesha Joshi & P. G. Aaron (eds.), *Handbook of Orthography and Literacy*, xiii–xiv. Mahwah, NJ & London: Lawrence Erlbaum.

Kazojć, Jerzy. 2010. *Otwarty słownik czesko-polski* [Open Czech–Polish dictionary], V.03.2010 (c). http://www.slowniki.org.pl/czesko-polski.pdf (accessed 22 April 2015).

Kempgen, Sebastian. 2009. Phonetik, Phonologie, Orthographie, Flexionsmorphologie. In Kempgen et al. (eds.), 1–14.

Kempgen, Sebastian, Peter Kosta, Tilman Berger & Karl Gutschmidt (eds.). 2009. *The Slavic Languages: An International Handbook of their Structure, their History and their Investigation*, vol. 1. Berlin & New York: Walter de Gruyter.

Kravchenko, Alexander V. 2009. The experiential basis of speech and writing as different cognitive domains. *Pragmatics & Cognition* 17(3), 527–548.

Křen, Michal. 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 11 September 2016).

Kučera, Karel. 2009. The orthographic principles in the Slavic languages: Phonetic/phonological. In Kempgen et al. (eds.), 70–76.

https://doi.org/10.1017/S0332586517000130 Published online by Cambridge University Press

Kürschner, Sebastian, Renée van Bezooijen & Charlotte Gooskens. 2008. Linguistic determinants of the intelligibility of Swedish words among Danes. *International Journal of Humanities and Arts Computing* 2(1/2), 83–100.

Levenshtein, Vladimir I. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady* 10(8), 707–710.

Ljaševskaja, Ol'ga N. & Sergej A. Šarov. 2009. *Častotnyj slovar' sovremennogo russkogo jazyka* [Frequency dictionary of the contemporary Russian language]. Moskva: Azbukovnik.

Marti, Roland. 2014. Historische Graphematik des Slavischen: Glagolitische und kyrillische Schrift. In Karl Gutschmidt, Sebastian Kempgen, Tilman Berger & Peter Kosta (eds.), *The Slavic Languages: An International Handbook of their Structure, their History and their Investigation*, vol. 2, 1497–1514. Berlin & New York: Walter de Gruyter.

Maslov, Jurij S. 1981. *Grammatika bolgarskogo jazyka* [A grammar of the Bulgarian language]. Moskva: Vysšaja škola.

Moberg, Jens, Charlotte Gooskens, John Nerbonne & Nathan Vaillette. 2006. Conditional entropy measures intelligibility among related languages. In Peter Dirix, Ineke Schuurman, Vincent Vandeghinste & Frank Van Eynde (eds.), *Computational Linguistics in the Netherlands 2006: Selected Papers from the 17th CLIN Meeting*, 51–66. Utrecht: LOT.

Möller, Robert & Ludger Zeevaert. 2015. Investigating word recognition in intercomprehension: Methods and findings. *Linguistics 2015* 53(2), 313–352.

Musatov, Valerij N. 2012. *Russkij jazyk. Fonetika, fonologija, orfoėpija, grafika, orfografija* [The Russian language: Phonetics, phonology, orphoepy, graphetics, orthography]. Moskva: Izdatel'stvo 'Flinta'.

Sampson, Geoffrey. 1985. *Writing Systems: A Linguistic Introduction*. Stanford, CA: Stanford University Press.

Schüppert, Anja & Charlotte Gooskens. 2012. The role of extra-linguistic factors for receptive bilingualism: Evidence from Danish and Swedish pre-schoolers. *International Journal of Bilingualism* 16(3), 332–347.

Schüppert, Anja, Nanna H. Hilton & Charlotte Gooskens. 2015. Swedish is beautiful, Danish is ugly? Investigating the link between language attitudes and spoken word recognition. *Linguistics* 53(2), 375–403.

Sgall, Petr. 2006. Towards a theory of phonemic orthography. In Petr Sgall (ed.), *Language in its Multifarious Aspects*, 430–452. Prague: Charles University; Karolinum Press.

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

Skorvid, Sergej S. 2005. Češskij jazyk [The Czech language]. In Aleksandr M. Moldovan, Sergej S. Skorvid, Andrej A. Kibrik, Natal'ja V. Rogova, Ekaterina I. Jakuškina, Anatolij F. Žuravlëv & Svetlana Tolstaja (eds.), *Jazyki mira. Slavjanskie jazyki* [The languages of the world: Slavic languages], 234–274. Moskva: Academia.

Smith, Nathaniel J. & Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition* 128(3), 302–319.

Stenger, Irina, Tania Avgustinova & Roland Marti. 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, 304–317.

https://doi.org/10.1017/S0332586517000130 Published online by Cambridge University Press

Stenger, Irina, Klára Jágrová, Andrea Fischer & Tania Avgustinova. In press. 'Reading Polish with Czech eyes' or 'How Russian can a Bulgarian text be?': Orthographic differences as an experimental variable in Slavic intercomprehension. In Peter Kosta & Teodora Radeva-Bork (eds.), *Current Developments in Slavic Linguistics: Twenty Years After* [preliminary title]. Frankfurt am Main: Peter Lang.

Ternes, Elmar & Tatjana Vladimirova-Buhtz. 2010. Bulgarian. In IPA (ed.), *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*, 55–57. Cambridge: Cambridge University Press.

Tribus, Myron. 1961. *Thermostatics and Thermodynamics*. Princeton, NJ: D. van Nostrand Company.

van Bezooijen, Renée & Charlotte Gooskens. 2007. Interlingual text comprehension: Linguistic and extralinguistic determinants. In Jan D. ten Thije & Ludger Zeevaert (eds.), *Receptive Multilingualism: Linguistic Analyses, Language Policies and Didactic Concepts*, 249–264. Amsterdam: John Benjamins.

van Heuven, Vincent J., Charlotte Gooskens & Renée van Bezooijen. 2015. Introduction Micrela: Predicting mutual intelligibility between closely related languages in Europe. In Judit Navracsics & Szilvia Batyi (eds.), *First and Second Language: Interdisciplinary Approaches* (Studies in Psycholinguistics 6), 127–145. Budapest: Tinta konyvkiado.

Vanhove, Jan. 2016. The early learning of interlingual correspondences rules in receptive multilingualism. *International Journal of Bilingualism* 20(5), 580–593.

Vanhove, Jan & Raphael Berthele. 2015a. The lifespan development of cognate guessing skills in an unknown related language. *International Review of Applied Linguistics in Language Teaching* 53(1), 1–38.

Vanhove, Jan & Raphael Berthele. 2015b. Item-related determinants of cognate guessing in multilinguals. In Gessica De Angelis, Ulrike Jessner & Marija Kresić (eds.), *Crosslinguistic Influence and Crosslinguistic Interaction in Multilingual Language Learning*, 95–118. London: Bloomsbury.

Vasmer, Max. 1973. *Ėtimologičeskij slovar' russkogo jazyka* [Etymological dictionary of the Russian language]. Moskva: Progress.

Yanushevskaya, Irena & Daniel Bunčić. 2015. Russian. *Journal of the International Phonetic Association* 45(2), 221–228.

Žuravlev, Anatolij F. (ed.). 1974–2012. *Ėtimologičeskij slovar' slavjanskich jazykov. Praslavjanskij leksičeskij fond* [Etymological dictionary of the Slavic languages: The Common Slavic lexical basis], vols. 1–37. Moskva: Nauka.

https://doi.org/10.1017/S0332586517000130 Published online by Cambridge University Press