An Orthography Transformation Experiment with Czech-Polish and Bulgarian-Russian Parallel Word Sets

Andrea Fischer, Klara Jagrova, Irina Stenger,
Tania Avgustinova, Dietrich Klakow and Roland Marti

Saarland University, Collaborative Research Center (SFB) 1102:
Information Density and Linguistic Encoding,
Campus A 2.2, 66123 Saarbrücken, Germany
Project C4: INCOMSLAV
Mutual Intelligibility and Surprisal in Slavic Intercomprehension
{kjagrova, avgustinova}@coli.uni-saarland.de,
{ira.stenger, rwmslav}@mx.uni-saarland.de,
{andrea.fischer, dietrich.klakow}@lsv.uni-saarland.de
http://www.sfb1102.uni-saarland.de

Abstract. This article presents the methods and findings of a computational transformation of orthography within two Slavic language pairs (Czech-Polish and Bulgarian-Russian) on different word sets. The experiment aimed at investigating to what extent these closely related languages are mutually intelligible, concentrating on their orthographies as linguistic interfaces to the written text. Besides analyzing orthographic similarity, the aim was to gain insights into the applicability of rules based on traditional linguistic assumptions for the purposes of language modelling.

1 Introduction

We are interested in identifying the mechanisms by which languages en- and decode information, focusing on the phenomenon of receptive multilingualism observed within the Slavic language group. We are framing the problem as one of (statistical) language model adaptation from a L1 to L2, incorporating results from traditional approaches and comparative historical linguistics. The key idea is that comprehension of a text in an unknown, but related language should be better when the language model adapted for processing the unknown language exhibits relatively low average surprisal.

This contribution elaborates on an inter-language orthographic transformation experiment[1] for which, based on orthographic features, different mappings between selected language pairs were tested. Two language pairs for which a relatively high degree of mutual intelligibility could be expected were chosen: Czech-Polish (CS-

[1] The experiment took place in the initial phase of the INCOMSLAV project at Saarland University, launched in October 2014. Morphology, lexis and syntax will be subject to later project phases.

PL, both West Slavic, and both using the Latin script with a number of additional diacritic signs) and Bulgarian-Russian (BG-RU, South and East Slavic, both using Cyrillic script). The probably best known and most obvious example for such orthographic correspondences of characters between Czech and Polish are *v:w*, *h:g*, *č:cz*, etc.

We collected and systematized traditional linguistic assumptions about how Slavic languages developed from a reconstructed parent language – referred to as Proto-Slavic or Common Slavic – to the modern varieties of Czech, Polish, Bulgarian and Russian (Schenker 1993). Although this parent language existed before any Slavic script appeared, historical linguistics was able to reconstruct how and in which stages the individual modern varieties moved from unity to diversity in the course of several centuries (Carlton 1991:9). The key features which reflect this development and now distinguish one Slavic language from another had their origin in Proto-Slavic times. Thus, there is a common base in the linguistic systems of the individual languages.

The existing orthography rules can be considered a result of both linguistic and sociolinguistic factors (Sgall 1987; Penzl 1987). Orthography does not only follow phonological, morphological and diacritical principles. It is also the syntactic, semantic, etymological and historical factors that are reflected in the graphematic representation of a language. Apart from this, written language is subject to manipulation by rules and laws created by governing authorities (e.g. in the process of spelling reforms). Kučera explains the specific character of Cyrillic as follows:

> *"Like Glagolitic and unlike the Latin alphabet, Cyrillic was a script customized to the contemporaneous Slavic languages, with a highly efficient and systematic one-to-one correspondence between its graphemes and the Slavic set of phonemes. […] [T]here have been few exceptions from the correspondence, a fact that was in marked contrast with the widespread use of digraphs in the systems based on the Latin alphabet. Thus, there was significantly more asymmetry, and consequently more looseness in the relation of the Slavic phonemes to the Latin graphemes than in their relation to Cyrillic graphemes." (Kučera 2009:74)*

2 Experimental Setup

Parallel contemporary vocabulary lists were analysed in terms of their orthographic similarity and the applicability of the correspondence patterns that are assumed in comparative Slavic linguistics. The objective of our transformation experiment was, in the first place, to validate (confirm or reject) the traditional assumptions by applying orthographic correlation rules, which were formulated on the basis of historical comparative linguistics, on contemporary word material. As a result of this experiment it should be possible, with the help of the validated rules, to describe or even predict the written representation of units of the source language a target language. If however the traditional assumptions appear not to hold for certain vocabu-

lary (sub-) sets, the relevant orthographic correlations were to be directly derived from the compiled parallel word lists.

2.1 Rules Inferred from Traditional Linguistic Assumptions

To account for the historically conditioned variation between the languages under investigation, we first collected and worked out orthographic correlations reflecting the development of the sound systems as established in historical comparative linguistics. We attempted to accommodate the main lines of the sound system evolution, from Common Slavic to individual modern Slavic languages, focusing on the following aspects: (i) development of vowels and consonants, (ii) development of specific sound combinations, and (iii) the metathesis of liquids.

The next step when designing the rule sets for the transformation experiment was a change of perspective, away from the perspective *Common Slavic vs. all other* towards a comparison of language pairs. In the diachronically-based language-family-oriented collection of correspondences[2] there were 132 for CS-PL vs. 126 for BG-RU (i.e. *h:g:ʔ:ʔ* for CS-PL-BG-RU). A considerable number of these rules stated regular one-to-one correspondences for the respective language pairs, for example such rules as *p:p* for CS-PL and *ʔ:ʔ* for BG-RU. Consequently, only those rules were applied in the experiment that represent a mismatch between target and source language units (e.g. *č:cz* for CS-PL and *ʔ:ʔ* for BG-RU), so that only 81 rules for CS-PL and 48 rules for BG-RU were applied to the word lists. This suggests a greater orthographic diversity between Czech and Polish than between the other two languages. Equal-to-equal grapheme correspondences were not considered a transformation. Such a situation in fact represents a reading intercomprehension scenario in which equal graphemes are not expected to cause any additional surprisal for readers. The remaining transformation rules were then applied on parallel word lists and checked for their practical usability.

Czech and Polish: Although both use the Latin script, they differ in their diacritical systems and the use of digraphs. While CS sibilants are usually represented by a single character, PL uses digraphs instead, at least for hard sibilants, e.g. *č:cz*. In the experimental setup, a letter is defined as an independent unit including diacritics, if applicable. For the purposes of the current automatic transformation, digraphs are considered two characters, e.g., PL *sz* and CS *ch*. There are 15 Czech letters (*á, č, ď, é, ě, í, ň, ř, š, ť, ú, ů, v*[3], *ý, ž)* that do not exist in PL, and 9 Polish letters (*ą, ć, ę, ł, ń, ś, w*[4]*, ż, ź*) that do not exist in CS. Still, these letters are expected to be legible for readers of the respective target language (i.e. by ignoring diacritical signs) and thus should not impair reading intercomprehension to a large extent – especially when the actual phonetic representation is similar (e.g. *á* vs. *a*, although this fact might not be known to the reader).

---

[2] The analyses were primarily collected from Bidwell (1963), Žuravlev et al. (1974-2012) and Vasmer (1973).

[3] The letter *v* is only used in Polish texts when it is part of a named entity or a foreign word.

[4] The letter *w* is only used in Czech texts when it is part of a named entity or a foreign word.

Bulgarian and Russian both use the Cyrillic script and there are only slight differences in the alphabets. The use of digraphs and diacritics is rare in the Cyrillic-based systems. The Russian letters *?*, *?*[5], *?* do not appear in BG. Generally, one can distinguish two important orthographic differences: unfamiliar graphemes representing unfamiliar or familiar phonemes (these differences only apply to a limited number of graphemes between BG and RU); graphemes that seem to be familiar, but in fact the grapheme-phoneme correspondences are different (e.g. *?* and *?* in BG are pronounced [ə] and [ʃt], while their RU counterpart *?* has no phonetic, but an orthographic function (hard sign) and *?* is pronounced [ʃtʃ], different rules for the reduction of unstressed vowels etc.).

2.2 Word Sets Used

In the initial phase of INCOMSLAV, we started collecting all parallel word lists and corpora that were available to us in digital format. The main inspiration and the first source of Slavic word lists was the EuroComSlav website. We decided to test the traditional assumptions on word lists instead of full texts in order to focus on the orthographic level only and thus exclude such influences that are caused by individual morphological rules from our analysis as far as possible.

Verb forms play a special role in the BG-RU comparison. While we analyzed infinitive verb forms in the CS-PL lists, we had to replace all infinitives in the BG-RU lists with the 3rd person present tense forms of the verbs. This was done to ensure a more appropriate comparison of RU with BG, as there are no infinitive forms in BG and 1st person forms are highly irregular, which makes them less suitable for an orthographic comparison.

There were three types of basic parallel lists available for all four languages: a Pan-Slavic list and a list of internationalisms on the EuroComSlav website, and the online version of the Swadesh list. The EuroComSlav lists had to be corrected for errors. All lists were slightly modified, as formal non-cognates (i.e. CS-PL *mnoho – wiele* [*many*/*much*]; BG-RU *??? – ??* [*we*]) were removed and formal cognates, if existing, were added to the lists, where the pairs consisted of non-cognates (i.e. *mężczyzna* [*man*] substituted by *mąż* [*husband*] in CS-PL *muž – mąż*; *????* [*beast*] added to its RU formal cognate *?????* [*animal*, *beast*] for the BG-RU pair *????? – ????*). Focusing only on the formal aspect of the lexemes, we did not take semantics into account. This explains the variation in the amount of words for each list in each language pair.

Table 1. Word sets with numbers of items

	Total number of items	
Word list	CS-PL	BG-RU
Swadesh list	212	227
Panslavic list	455	447

[5] The letter *?* is used mostly only in dictionaries and schoolbooks.

Internationalism list	262	261
Homonyms	1553	X
Dictionary	80963	X

For the CS-PL pair we implemented two additional large word lists which might have a statistically more representative effect: A set of homonyms, extracted from (Szałek and Nečas 1993), as well as an open-source digital version of a CS-PL dictionary containing more than 80,0000 lexemes (Kazojć 2010).

2.3 Method

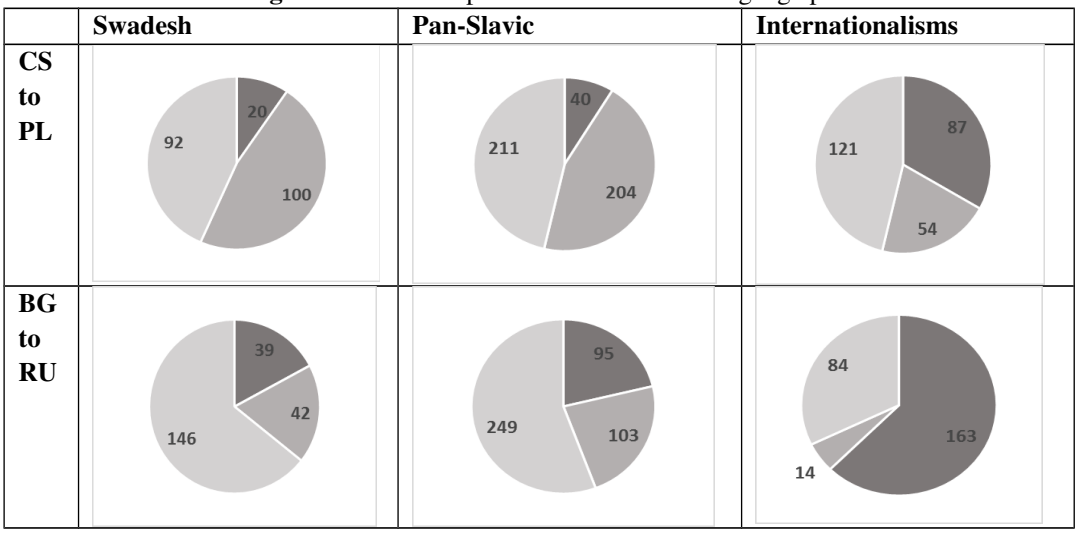
If all characters in a word of L1 are the same as in the corresponding word in L2, the word was automatically listed as *input identical*. If there is a mismatch of one or more positions in the word pair, the computer tries to apply one or more rules from the transformation rule set. If all characters in a word of L1 can be transformed with the help of the rules into the L2 word, the word pair is listed as *correctly transformed*. Rules for strings of characters take precedence over rules for single characters. There is also a chance that a unit from L1 corresponds to a different unit (character or string of characters) in L2, which is not part of the traditional linguistic rule set entered for this experiment. In such a case, these words are classified as *untransformed*.

The computer code for the implementation of the orthographic transformation rules between language pairs (by Andrea Fischer and Ali Shah) is provided below.

```
method Transformations(w, T)
--------------
input: a word w from language L1, the set T of admissible transformation rules
output: all L2 transformations of w obtained by applying rules from T
--------------

transformations = {(w, [])} // initialize the set of transformations with just
the word and no applied transformations

new_variants = {} // temporary iteration variable

while True: // iterate until no new transformations are found anymore
        for t in T: // process each transformation rule
                for variant, path in transformations: // apply this rule to all
                currently known variants of the original word
                        for new_word, application_pattern in
                        TransformWithRule(variant, t): // apply the rule t in
                        each combination of positions where it is applicable
                                new_variants.add((new_word, path,
                                application_pattern)) // record the new variant
                                plus the path by which it was obtained
        if words(new_variants) + words(transformations) ==
        words(transformations): // after processing all rules, see if there are
        any new words
                break // stop iteration if no new words were found
        else:
                transformations.addAll(new_variants) // record the newfound
                transformations and continue iterating otherwise

return transformations
```

2.6 Results of the Implementation of the Rules

Fig. 1. Results of implementation for both language pairs



Legend: ■ input identical, ■ correctly transformed, □ untransformed

The most obvious finding is the different proportion of orthographically identical words in the language pairs (max.: 33.21 % for CS-PL vs. 62.45 % for BG-RU). The internationalism lists, consisting only of nouns, show the highest proportion of orthographically identical words in both pairs. An explanation for the low rate of identical words in the BG-RU Swadesh list is the high rate of morphological differences reflected in orthography, e.g. different endings of male adjectives and verb forms in 3rd person singular – here the orthographic rule set can be applied only in very few cases. However, the rule set works well for the CS-PL Swadesh list (best transformation rate of the experiment: 47.17 %).

The Swadesh lists consist of a relatively high rate of verbs and adjectives and they are the only lists containing a number of pronouns, prepositions and numerals. The Pan-Slavic lists include nouns, verbs and adjectives. While for CS-PL the proportion of untransformed items is relatively constant throughout the three lists, the untransformed part for BG-RU ranges from 64.32 % with the Swadesh list to 32.18 % with internationalisms.

Tables 2a and 2b display the five most frequently used rules for each word set, given the rule from L1 to L2 along with the number showing how often this rule was applied in words that were classified as *correctly transformed*. Directly under the rules, one example word pair for which this rule holds is provided.

Table 2a. Most frequent transformation rules applied on the different lists

Czech to Polish

Swadesh	Pan-Slavic	Internat.	Homonyms	Dictionary
t:ć,24	ý:y,45	á:a,15	v:w,307	v:w,991
dát–dać	dýně–dynia	bál–bal	věc–wiec	kráva–krowa
ý:y,21	v:w,42	e:a,12	ý:y,175	t:ć,663
nový–nowy	voda–woda	linie–linia	výlet–wylot	prát–prać
v:w,20	t:ć,37	v:w,8	t:ć,163	á:a,515
dva–dwa	bolet–boleć	káva–kawa	tma–ćma	pára–para
á:a,10	l:ł,24	í:e,5	á:a,142	e:a,353
já–ja	zlý–zły	talíř–talerz	čára–czara	duše–dusza
l:ł,9	h:g,20	l:ł,4	l:ł,111	ý:y,336
teplý–ciepły	hlava–głowa	kanál–kanał	látka–łatka	dým–dym
		rá:ra,4		
		rádio–radio		

CS-PL: The success of *t:ć* can be explained by the high rate of verb endings (morphological feature reflected in orthography) in all lists except in internationalisms, although this rule was originally inferred from the diachronically-based rule for *deset - dziesięć*. Another outstanding rule is *ý:y* which is due to a high rate of adjective endings in the lists, although this rule was originally derived from a historical correspondence in word stems. For some rules such as *á:a* and *l:ł* it may be assumed that they will not pose a problem to reading intercomprehension because the diacritics can be ignored. The *v:w* rule represents characters that would not appear in the other language. The success of applying these rules in this experiment depends strongly on their overall frequency of the individual characters in the word lists, i.e. there is a higher frequency of *h:g* in Pan-Slavic vocabulary relative to *h:g* in all other lists. The strongest tendencies for vowel changes are from *e:a*, from *í:e*, which both apply for noun endings. As a result, the findings reveal a strong applicability of rules that refer to endings, to letters that are not part of the inventory of one language and to letters that are only distinguished by the absence or presence of diacritical signs.

Table 2b. Most frequent transformation rules applied on the different lists

Bulgarian to Russian

Swadesh	Pan-Slavic	Internation.
?: ?,8	?: ?,17	?: ??,9
???– ????	????– ????	???– ????
?: ?,7	?: ?,10	?: ?,1
????– ????	???– ???	??????– ??????
?: ?,6	?: ?,10	?: ??,1
??– ??	????– ????	a???– ?????
?: ?,6	?: ?,9	?: ??,1

????– ????	??– ???	????????– ?????????
?.?,4	?.?,8	?.?,1
?????– ?????	??– ??	????– ????

BG-RU: The most frequent orthographic correspondences of the transformation experiment in the Swadesh and Pan-Slavic lists are between the orthographic representations of vowels: *ъ:y; е:я; я:е; и:ы; е:о; е:ё*. The orthographic differences could generally be explained, on the one hand, by the different development of the vowels from Common Slavic to the modern Slavic languages and, on the other hand, by subsequent spelling reforms in these languages with the aim to harmonize their writing system to the sound system, i.e.:
ъ:y – explained by the different development of the back nasal vowel */ǫ/ of Common Slavic to /ə/ in Bulgarian and to /u/ in Russian.
е:я – also explained by the different development of the front nasal vowel */ę/ of Common Slavic to /e/ in Bulgarian and to /ˊa/ in Russian.

The most frequent orthographic correspondences in the internationalism list, besides the correlations of orthographic representations of the vowels *a:?* and *?:?* , here concern the orthographic representation of consonants, e.g. *?:??* , *?:??* , *?:??* , which can be explained by the difference between non-palatalized consonants in Bulgarian and palatalized consonants in Russian. It must be kept in mind, however, that most internationalisms in the list are borrowings from other languages and thus constitute a rather specific problem. Usually, in orthographies using Cyrillic the pronunciation of the borrowing may be preserved and the spelling may be changed to correspond to the orthographic rules of the borrowing language (Kučera 2009). Borrowings were generally handled in harmony with the phonological and morphological principles of each particular language, which could be presented by other orthographic correspondences that are distinct from our diachronically-based transformation rules. This could be an explanation for the fact that only five of the transformation rules could be successfully applied on the internationalism list. However, there already is a high rate of identical words in this list.

The overall results for both language pairs show that there are different principles in how the diachronically-based transformation rules work. For BG-RU, the results confirm the validity of the rule set to a high degree for the reasons mentioned above. For CS-PL we found that the traditional rules were valid not only for word stems as explained in historical comparative research, but also for other parts of words, mainly endings. The rules do not only cover orthographic features, but also those morphological features to which the same rules apply. The words classified as *correctly transformed* were much lower in number for BG-RU. This could be explained by the fact that in the experiment, words in which there was only one unit that could not be transformed with the rule set, were sorted out by the program as *untransformed*. For example the adjective pair *???* (BG) vs. *?????* (RU) could not be correctly transformed, because there is no rule saying ø [nothing] in BG corresponds to -*??* in RU – this would require a morphological rule set.

The difference in the language pairs confirms the isolated position of Bulgarian in contrast to the other languages under focus, especially because of its morphology.

3 Conclusions

In the present application of diachronically-based orthographic transformation rules between the two language pairs CS-PL and BG-RU we tried to find out to what extent traditional linguistic assumptions explain the differences between parallel word sets in the languages. The computational transformation experiment revealed that there are different percentages of orthographically identical words in both language pairs. For all word sets, the initial orthographic similarity is greater for BG and RU (max.: 62.45 % for internationalisms) than for CS and PL (max.: 33.21 % for internationalisms), which suggests a greater degree of mutual intelligibility for BG-RU by the presence of internationalisms than in the other pair.

For those words in the parallel lists that were not identical in terms of orthography, a rule set of inter-language orthographic correspondences was applied. For the CS-PL combination, these orthographic transformation rules led to better results – 44.84 % for the Pan-Slavic vocabulary list, while the results for BG-RU in the same list amounted to only 23.04 %. The low success rate for the BG-RU orthographic transformations suggests a high influence of morphological differences between these languages (zero endings for BG adjectives, different verb endings, etc.). While investigating the CS-PL orthographic correspondences, we found that the morphological features are reflected in the respective orthographies to a similar degree and are therefore comparable. This suggests that knowledge of those orthographic correspondence rules might improve reading comprehension, e.g., for a Czech native speaker reading Polish. The knowledge of orthographic correspondences between BG and RU, in contrast, is not expected to lead to such large improvement in reading comprehension as in the other pair, when the respective other language is unknown to the reader. However, knowledge of morphological cross-language correspondence principles might be much more helpful here.

4 Outlook

Orthography was subject to the first of six work packages in the INCOMSLAV project. In the near future a series of on-line reading (inter-)comprehension experiments with Slavic native speakers is planned to validate the findings of this and other computational experiments. The results from the experiments with human readers will be discussed in the framework of several other computational estimations and calculations of similarity and distance. The upcoming project phases will cover morphology, lexis and syntax. On the linguistic level, more similarities and discrepancies in the subsystems of the languages will be investigated. Both the nature of the phenomena and the strength of the effects are relevant at these levels.

For the information-theoretic part of the project, the aim will be to adapt feature-based n-gram language models for cross-language use via latent space and similarity. The information-theoretical results will then be analyzed again from a linguistic point

of view and interpreted together with the results of reading intercomprehension experiments with Slavic native speakers.

References

Bidwell, C.E. (ed) *Slavic Historical Phonology in Tabular Form.* Mouton & Co., The Hague, 1963

Carlton, T. R. (ed) *Introduction to the Phonological History of the Slavic Languages.* Slavica Publishers, INC. Columbus, Ohio, 1991

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*. Volume 1. Walter de Gruyter, Berlin/New York, pp. 70-76

Penzl, H. (1987) Zur alphabetischen Orthographie als Gegenstand der Sprachwissenschaft. In: Luelsdorff, P. A. (ed.): *Orthography and Phonology*. John Benjamins Publishing Company, Amsterdam/Philadelphia, pp. 225-238

Schenker, A.M. (1993) Proto-Slavonic. In: Comrie, B., Corbett, G.G. (eds.) *The*

Slavonic Languages, Routledge, London and New York, pp. 60-125

Sgall, P. (1987) Towards a Theory of Phonemic Orthography. In: Luelsdorff, P. A. (ed.) *Orthography and Phonology*. John Benjamins Publishing Company, Amsterdam/Philadelphia, pp. 1-31

Dictionaries

Szałek, M.; Nečas, J. (eds) *Czesko-Polska Homonymia*. Poznań, 1993

Vasmer, M (ed) *Etimologičeskij slovar' russkogo jazyka*. Moscow, 1973

Žuravlev, A. F., et al. *Etimologičeskij slovar' slavjanskich jazykov*. Vyp. 1-37. Moscow, 1974-2012

Online documents

Swadesh list:
http://en.wiktionary.org/wiki/Appendix:Swadesh_lists_for_Slavic_languages. Accessed 22/04/2015

Pan-Slavic list:
http://www.eurocomslav.de/kurs/pwslav.htm. Accessed 22/04/2015

Internationalism list:
http://www.eurocomslav.de/kurs/iwslav.htm. Accessed 22/04/2015

Kazojć, J. (2010) Otwarty słownik czesko-polski V.03.2010 (c)
http://www.slowniki.org.pl/czesko-polski.pdf. Accessed 22/04/2015