Formal Aspects of Mutual Intelligibility between Slavic Languages: Output Analysis of Czech-to-Polish and Bulgarian-to-Russian Orthography Transformation Experiments

Tania Avgustinova, Irina Stenger, Klára Jágrová, Roland Marti

Abstract

This paper presents statistical results and linguistic analyses of orthographic transformation experiments with Czech-Polish and Bulgarian-Russian parallel word sets. In a preceding research effort, we carried out large-scale computational transformation experiments on parallel word sets, with orthographic correspondences based on traditional approaches and comparative historical linguistics. Our aim was to investigate to what degree the selected Slavic languages are mutually intelligible at the orthographic level and to analyze the most frequent orthographic correspondences and differences between the respective Slavic language pairs. Based on the insights we gained from our previous experiments, we now investigate the untransformed part of our experiments, taking into consideration orthographical features with their relationships to phonological, morphonological, and morphological features from the perspective of common and comparative historical Slavic linguistics.

1 Background

Similarities in phonology, morphology, syntax and basic vocabulary among Slavic languages are striking. According to (Townsend & Janda 1996), "[m]ost Slavs speak of understanding each other without much difficulty, but this is usually exaggerated and applies mostly to a simple concrete level". While the degree of intelligibility of an unknown but closely related language depends on both linguistic and extra-linguistic factors (Gooskens 2013), *orthography* represents a primary linguistic interface that is crucial for reading comprehension.

We approach the problem of mutual intelligibility from an information-theoretic perspective in terms of *surprisal* (vs. predictability) in linguistic encoding in a reading comprehension scenario. In particular, the central research question could be formulated as follows: To what extent is a speaker of one Slavic language (L1) able to understand a written text in another unknown but closely related Slavic language (L2), using the encoding system of L1 and applying it to decode a message in L2, e.g., a Czech native speaker attempting to read a Polish newspaper or a Bulgarian native speaker being confronted with a written text in Russian.

Our hypothesis is that orthography is a *linguistic determinant of mutual intelligibility* which may facilitate or impede intercomprehension. In order to reveal genuine linguistic distances that would enable information processing in L2 using L1 linguistic competence, we try to avoid the additional cost of adaptation to an unfamiliar writing system, i.e. from Cyrillic to Latin and vice versa. We therefore started with the following language pairs: Czech and Polish (both West Slavic, using the Latin alphabet) vs. Bulgarian and Russian (South and East Slavic, both using the Cyrillic alphabet). When comparing different orthographic systems, special attention must be paid to the various descriptive levels influencing them (i.e. phonetics/phonology, graphemics/graphotactics, morphology/morphosyntax) as well as to historical, etymological and sociolinguistic factors (e.g., spelling reforms) (Penzl 1987; Sgall 1987; Sgall 2006). As a genetically related group, Slavic languages are descendants of a single ancestor language, traditionally referred to as Proto-Slavic or Common Slavic, with characteristics that can be reconstructed by comparing the attested language varieties (Carlton 1991; Comrie & Corbett 1993). The observed similarities and differences are due to the common origin in combination with a slow but steady movement from unity to diversity (Carlton 1991; Mel'ničuk 1986). We started with by establishing diachronically motivated and synchronically attested inter-Slavic *orthographic correlates* for the selected languages.

2 Orthographic Correlates

To investigate to what degree the selected Slavic languages (here Czech-Polish and Bulgarian-Russian) are mutually intelligible at the orthographic level, we analyzed the most frequent orthographic correlates in the respective language pairs. The analyses of historically conditioned cross-linguistic variations between sound sequences, which allow for establishing orthographic correlates, were primarily collected from (Bidwell 1963; Vasmer 1973; Žuravlev 1974-2012). In the process we also attempted to account for the main lines of sound system evolution with regard to (i) the development of vowels and consonants, (ii) the development of specific sound combinations, and in particular (iii) the metathesis of liquids. This resulted in a compilation of *diachronically-based orthographic correspondences* which were automatically tested on parallel word sets for the purposes of an orthography transformation experiment described in (Fischer et al. 2015). These hand-crafted parallel word sets were developed on the basis of the Pan-Slavic and the internationalism lists of the EuroComSlav project[1] and standard Swadesh lists from Wiktionary[2], carefully correcting the errors contained in the EuroComSlav sources.

Focusing only on the formal aspect of the lexemes, all three lists were slightly modified. On the one hand, formal non-cognates (i.e. CS-PL *mnoho – wiele* 'many/much'; BG-RU *ние – мы* 'we') were removed. On the other hand, formal cognates, if existing, were added to the lists where the pairs consisted of non-cognates (i.e., *mężczyzna* 'man' was substituted by *mąż* 'husband' in CS-PL *muž – mąż*; *звяр* 'beast' was added to its Russian formal cognate *зверь* 'animal, beast' for the BG-RU pair *звяр – зверь*). This explains the variation in the amount of words in Table 1 for each list in each language pair.

Word sets	Total number of items	
	CS-PL	BG-RU
Swadesh list	212	227
Panslavic list	455	447
International. list	262	261

Table 1: Word sets with numbers of items

2.1 Method of Implementation

The following strategy has been pursued in the pair-wise orthography transformation, cf. (Fischer et al. 2015). If all characters (orthographic elements) in a given word of the source language L1 are the same as in the corresponding word in the target language L2, the word is automatically listed as *input identical* in the experimental output. If there is a mismatch of one or more positions in the word pair, the computer program performs one or more mappings based not only on single characters but on strings of character too, e.g., CS-PL *ž – ż* and *ře – rze*. In the process, rules for longer strings of characters are preferred before rules for shorter strings or single characters. If all characters in the L1-word can be mapped to characters of the corresponding L2-word, the word pair is listed as *correctly transformed*. If, however, the L1-word contains a character or a string of characters that corresponds to a different character or a string of characters in the L2-word that has not been included in the orthographic correspondences for the experiment, the word pair is classified as *untransformed* in the output.

2.2 Closer Look at the Statistical Results

For the most part, the obtained automatic transformations could be seen as satisfactory for both language pairs. The successful results of the computational application range from 53.63% for CS-PL with the Pan-Slavic list to 67.82% for BG-RU in the internationalism list. Taking into account the two categories *input identical* plus *correctly transformed* words, the best rates are 56.60% for CS-PL (in the Swadesh list) and 67.82% for BG-RU (in the internationalism list). When analyzing the results with more attention to linguistic details, we noticed the different proportion of *orthographically identical words* in the language pairs: a maximum of 33.21% for CS-PL vs. 62.45% for BG-RU, both in

[1] http://www.eurocomslav.de/kurs/pwslav.htm.; http://www.eurocomslav.de/kurs/iwslav.htm.

[2] http://en.wiktionary.org/wiki/Appendix:Swadesh_lists_for_Slavic_languages.

the internationalism lists; a minimum of 9.43% for CS-PL vs. 17.18% for BG-RU, both in the Swadesh lists. In the latter case, it must be kept in mind that the numbers of items in the word lists are different, especially for the Swadesh lists.

Based on these results, one could hypothesize that the degree of mutual intelligibility between Bulgarian and Russian is higher at the formal orthographic level than between Czech and Polish. However, this hypothesis reflects only the formal aspect of word transformation without taking into consideration either the grapheme-phoneme correspondences or the semantics of the words. Therefore it has still to be confirmed by further experiments.

The second obvious finding is the different proportion of *correctly transformed words* in the language pairs: maximum 44.84% for CS-PL vs. 23.04% for BG-RU, both for the Pan-Slavic lists; minimum 20.61% for CS-PL vs. 5.36% for BG-RU, both for the internationalism lists. Here the set of correspondences apparently works more successfully for CS-PL than for BG-RU. However, there already is a high rate of identical words for BG-RU.

In the centre of our attention is now the third category: *untransformed words*. While the proportion of untransformed items is relatively equal throughout the three lists for CS-PL: from 43.40% in the Swadesh list to 46.37% in the Pan-Slavic list, the untransformed part for BG-RU ranges from 64.32% with the Swadesh list to 32.18% with internationalisms. The tables 2 and 3 present these results with numbers of words.

Parallel word sets	**Input identical**	**Corr. transf.**	**Untransf. items**
Swadesh list	20	100	92
Panslavic list	40	204	211
International. list	87	54	121

Table 2: Results of the transformation for CS-PL

Parallel word sets	**Input identical**	**Corr. transf.**	**Untransf. items**
Swadesh list	39	42	146
Panslavic list	95	103	249
International. list	163	14	84

Table 3: Results of the transformation for BG-RU

3 Linguistic Interpretation

Our current goal is to linguistically interpret the statistical results from (Fischer et al. 2015) in an attempt to understand what "went wrong", i.e. focusing on the set of *untransformed items*, in order to sketch the next series of intelligibility experiments.

3.1 Czech and Polish

Czech and Polish use the Latin alphabet with diacritical signs. The Czech letters *á, č, ď, é, ě, ch, í, ň, ř, š, ť, ú, ů, ý, ž* as well as *q*[3], *v*[4], and *x*[5] are not considered part of the Polish alphabet, and the Polish *letters ą, ć, ę, ł, ń, ś, w*[6], *ż* and *ź* do not exist in Czech. However, there are sound correspondences in both languages that are represented differently in orthography. The Czech characters as *č* or *š* can correspond to the digraphs *cz* and *sz* in Polish. Here, Czech orthography can be considered denser than the Polish. On the other hand, where there is a nasal vowel such as *ą* in PL, it would be represented by a diphthong, e.g., *ou,* in Czech, which makes Polish denser in this aspect.

From a diachronical perspective and taking into consideration all four languages, we established 132 correspondences for CS and PL (e.g., *a:a, á:ią, ě:ię, z:dz, hv:gw, lou:łu, o:o, m:m, ou:ą, rů:ró, ří:rze, šť:szcz, c:c*). After omitting the equal-to-equal correspondences for this language pair, there remained 81 correspondences for the transformation experiment, cf. (Fischer et al. 2015).

Throughout all three word lists, there were no significant differences in the rates of cognates that could not be transformed with the help of the orthographic correspondences (min. 43.40% in the

[3] The letters q, w, and x are only used in foreign named entities in Czech. However, they are listed as part of the Czech alphabet in dictionaries and school books.

[4] The letters v, q, and x are not mentioned in the Polish alphabet as in dictionaries or school books. They can only appear in Polish as foreign named entities.

[5] See footnote 3

[6] See footnote 3.

Swadesh list; max. 46.37% in the Pan-Slavic list classified as *untransformed*). When analyzing the untransformed category in the experiment output for each list, the results show some basic tendencies.

The untransformed cognates of the Pan-Slavic and Swadesh lists suggest that we need to extend the rule set to account for correspondences involving characters with or without diacritics in both transformational directions:

For example, the set of correspondences allows a transformation of the CS *é* to the PL *a* or *ie* only. However, the pairs CS-PL *plést – pleść* 'to knit' or *děšť – deszcz* 'rain' demand a rule that tolerates the absence of the diacritical sign above the grapheme *e* to make them successfully transformable. The same applies accordingly for the underlined positions in pairs such as *jazyk – język* 'language/tongue', *se – się* 'oneself', *zvíře – zwierzę* 'beast', *široký – szeroki* 'broad'. These pairs were categorized as *untransformed*, because the set of correspondences allowed only transformations of CS-PL *á:ę*, *ě:ię*, *e:e* and *í:e*.

Another correspondence that becomes apparent in those two lists is the CS-PL *kd – gd* pair which can be explained by the historical principle that Czech orthography is following in this case vs. the phonetical principal of Polish orthography (Kellner 1936) here: the historical *k* is kept before *d*, although there is an assimilation in pronunciation of the voiceless *k* to a voiced /*g*/ when it is followed by a voiced consonant, such as in *kde – gdzie* 'where'.

The results further demand an addition of phonetic correlates to the set of correspondences, respectively an addition of grapheme-phoneme correspondences within a language. In all three lists, the most frequently lacking rules appeared to be CS-PL *i:y*, *st:ść*, *s:ś*, e.g., in the pairs CS-PL *živý – żywy* 'alive', *mladost – młodość* 'youth', *světlý – świątły* 'bright'. Previously formulated correspondences allowed only: *i:i*, *í:i*, *sť:szcz* (here: tolerating diacritics would be necessary), *s:sz*.

The internationalism list unifies points made above and adds other important insights about the (orthographic) distance of the two languages.

There are different ways in which loan words are represented in both orthography and phonetics in the two languages, with more or less orientation on the original, e.g., CS-PL *mač – mecz* 'match', *leasing – lis* 'leasing', *apartmá – apartament* 'appartment'.

Most of the internationalisms categorized as *untransformed* differ in their endings, sometimes because of being of different gender in the two languages, e.g., CS-PL *univerzita – uniwersytet* 'university', *teritorium – terytoria* 'territory', *recept – recepta* 'recipe', *sál – sala* 'hall', *salát – sałata* 'salad', but sometimes having different endings despite same gender, such as *legitimace – legitymacja* 'legitimation', *penze – pensja* 'pension'. The performance of correlation rules on such pairs could be improved by adding morphological correspondences to the set.

Polish uses *ks* instead of *x*: CS-PL *maximum – maksymum*, *export – eksport*. Furthermore, there are no exceptions for internationalisms in Polish orthography as there are in Czech. This becomes apparent when comparing the pairs CS-PL *legitimace – legitymacja*, *kredit – kredyt*, *praktika – praktyka*, *medicína – medycyna*. Although the Czech internationalisms use the letter combinations *ti* and *di*, *t* and *d* are not palatalized by the *i* as they would be in non-internationalisms. The phonetic principle seems to be stronger represented in Polish orthography than in Czech, when looking at internationalisms.

Consequently, the set of diachronically-based orthographic correspondences that was compiled for the orthographic transformation experiment should be extended by the information gained from the *untransformed* vocabulary in the points mentioned. An addition of morphological correspondences will be necessary for a successful application of correlates between the two languages, for instance for machine translation.

3.2 Bulgarian and Russian

Bulgarian and Russian use the Cyrillic alphabet. Three letters of the Russian alphabet do not occur in Bulgarian: *ы, э, ё*[7]. The Bulgarian alphabet thus consists of the following letters: *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*. For the computational transformation, the small case letters are used.

From the formal visual perspective, the forms (printed and handwritten) of the Bulgarian letters do not differ from their Russian counterparts. However, the use and pronunciation of a number of letters

[7] The letter *ё* is used mostly only in dictionaries and schoolbooks.

is not the same as it is in Russian (Gribble 1987; Gribble 2013; Ivanova et al. 2011). For the computational transformation we took into account only the written (= printed) text itself, without regard to its relationship to spoken language.

Ignoring the technical details, let us summarize that 126 diachronically-based orthographic correspondences have been formulated for BG-RU, including equal-to-equal correspondences (e.g., *б:б, г:г, к:к, п:п, т:ть, б:бл, в:вл, жд:ж, м:мл, п:пл, а:а, е:е, ъ:у, и:ы, я:е, ла:оло* etc.). However, only those correspondences were used in the experiment which represented a mismatch between target and source language units (e.g., *т:ть, б:бл, в:вл, жд:ж, м:мл, п:пл, ъ:у, и:ы, я:е, ла:оло etc.*). Thus only 48 correspondences were applied on parallel word lists for the BG-RU mapping.

While the proportion of the transformation of identical words (this means equal-to-equal correspondences) between Bulgarian and Russian was high (see 2.2), the transformation set of correspondences performed less successfully for BG-RU than for CS-PL. Based on the comparative analyses of the untransformed parts, the following observations can be made.

The diachronically-based orthographic correspondences that were applied do not include all possible orthographic correlates (e.g., the internationalism list shows the lowest rate with 5.36% of correctly transformed words based on the applied set of correspondences). The internationalism list that is based on EuroComSlav requires additional rules for the following systematic BG-RU correspondences, well known from common and comparative Slavic studies (Gribble 1987; Gribble 2013; Ivanova et al. 2011; Valgina et al. 2002):

- *ьо:ё* (*актьор – актёр* ‘actor’, *партньор – партнёр* ‘partner’, *шофьор – шофёр* ‘driver’);
- *е:э* (*економия – экономия* ‘economy’, *експорт – экспорт* ‘export’, *енергия – энергия* ‘energy’ etc.);
- *л:лл* (*алигатор – аллигатор* ‘alligator’, *колега – коллега* ‘colleague’);
- *п:пп* (*апарат – аппарат* ‘administration, mechanism’, *апетит – аппетит* ‘appetite’ etc.);
- *с:сс* (*бос – босс* ‘boss’, *дискусия – дискуссия* ‘discussion’ etc.);
- *р:рр* (*перон – перрон* ‘platform’ etc.);
- *н:нн* (*тунель – туннель* ‘tunnel’ etc.).

It is necessary to collect further orthographic correlates, based on comparative and diachronic Slavic studies, taking into consideration systematic phonological as well as morphonological correspondences, e.g., Bulgarian /ə/ will most often correspond to /u/ in Russian (both from the back nasal vowel of Common Slavic */ǫ/): BG *зъб, път, ръка* – RU *зуб, путь, рука* ‘tooth’, ‘road’, ‘hand’/‘arm’. In suffixes, and rarely in roots, when ъ is or was a mobile vowel it will correspond to *о* in Russian (Gribble 1987; Gribble 2013): BG *зъл*, *зла* – RU *зол*, *зла* ‘wicked’ (this case is an example for Russian short adjective forms). Our diachronically-based orthographic correspondences already include both mentioned correlates. However, in the Pan-Slavic word list, there are long forms of adjectives as cognates for Russian: BG *зъл* – RU *злой* ‘wicked’. These two points: (i) the lack of some BG-RU orthographic correlates, e.g., *ъ:ø* and (ii) different morphological features could explain the unsuccessful transformation of adjectives in the Pan-Slavic and Swadesh lists, the same holds true for the BG-RU verb forms in these lists. From the Common Slavic */ь/ (also as a mobile vowel) we get *е* in both Bulgarian and Russian: BG *отец*, *ден* – RU *отец*, *день* ‘father’, ‘day’. However, in a few words, Bulgarian ъ may correspond to Russian *е* vs. *ё* (Gribble, 1987, 2013): BG *пъстър* – RU *пёстрый* ‘colorful’; BG *тъмно* – RU *темно* ‘dark’.

In general, Bulgarian has gone through major changes in nominal morphology, whereas in verbal morphology it has kept and expanded the old system in contrast to Russian and other Slavic languages (Gribble 1987). The most characteristic feature of Bulgarian inflectional morphology is its loss of case in all declensions, (except for vocative forms and remnants (nominative, accusative, dative) in the pronoun system) – cf. (Gribble 1987; Townsend & Janda 1996).

4 Conclusion and Future Work

The diachronically-based correspondences have originally been explored on roots of Common Slavic vocabulary. After feeding them to the computer program, some of the correspondences also applied to other parts of the lexemes, such as suffixes and endings.

The next step ahead is the comparison of morphology in the four languages mentioned, as well as other Slavic languages. Taking into consideration that *morphology* is the science of the smallest meaningful units of language it is typical to distinguish between derivational and inflectional morphological features. As the different morphological processes are inseparable for inflectional languages, we deal in both cases with a certain ensemble of units (Akhmanova 1971). Therefore, inflectional and derivational aspects have to be considered jointly to formulate the correspondences based on morphological features.

The computational orthographic analysis described here is implemented in the framework of the ICOMSLAV project launched in October 2014 at Saarland University (Avgustinova et al. 2014-2018). The outcomes will be tested in human reading intercomprehension experiments (e.g., free translation tasks, multiple choice, cloze tests – isolated words vs. words in context). The results will be used for building a feature-based language model mapping the encoding system of one language to another.

Online sources

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

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

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

References

Akhmanova, Olga. 1971. Phonology, Morphonology, Morphology The Hague, Paris: Mouton.

Avgustinova, Tania, Dietrich Klakow & Roland Marti. 2014-2018. Mutual Intelligibility and Surprisal in Slavic Intercomprehension. Project C4 INCOMSLAV, SFB 1102 Information Density and Linguistic Encoding, Collaborative Research Center at Saarland University: funded by the German Science Foundation (DFG).

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

Carlton, Terence R. 1991. Introduction to the Phonological History of the Slavic Languages Columbus, Ohio: Slavica Publishers, Inc.

Comrie, Bernard & Greville G. Corbett. 1993. Introduction. The Slavonic Languages, ed. by B. Comrie & G.G. Corbett, 1-20. London and New York: Routledge.

Fischer, Andrea, Klara Jagrova, Irina Stenger, Tania Avgustinova, Dietrich Klakow & Roland Marti. 2015. An Orthography Transformation Experiment with Czech-Polish and Bulgarian-Russian Parallel Word Sets. Paper presented at the Natural Language Processing and Cognitive Science, Krakow 2015.

Gooskens, Charlotte. 2013. Methods for measuring intelligibility of closely related language varieties. Handbook of Sociolinguistics, ed. by R. Bayley, R. Cameron & C. Lucas, 195-213: Oxford University Press.

Gribble, Charles E. 1987. Reading Bulgarian through Russian Columbus, Ohio: Slavica Publishers, Inc.

—. 2013. Reading Bulgarian through Russian. 2nd Revised Edition Columbus, Ohio: Slavica Publishers, Inc.

Ivanova, E. Ju., Z. K. Šanova & D. Dimitrova. 2011. Bolgarskij jazyk St. Petersburg: Karo.

Kellner, Adolf. 1936. Revise polského pravopisu. Slovo a slovesnost 2.246-48.

Mel'ničuk, A.S. 1986. Istoričeskaja tipologija slavjanskich jazykov. Kiev: Naukova dumka.

Penzl, Herbert. 1987. Zur alphabetischen Orthographie als Gegenstand der Sprachwissenschaft. Orthography and Phonology, ed. by P.A. Luelsdorff, 225-38. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Sgall, Petr. 1987. Towards a Theory of Phonemic Orthography. Orthography and Phonology, ed. by P.A. Luelsdorff, 1-31. Amsterdam/Philadelphia: John Benjamins Publishing Company.

—. 2006. Towards a Theory of Phonemic Orthography. Language in its multifarious aspects, ed. by P. Sgall, 430-52. Prague: Karolinum Press, Charles University.

Townsend, Charles E. & Laura A. Janda. 1996. Common and Comparative Slavic: Phonology and Inflection with special attention to Russian, Polish, Czech, Serbo-Croatian, Bulgarian Columbus, Ohio: Slavica Publishers, Inc.

Valgina, N. S., D. Ė. Rosental' & M. I. Fomina. 2002. Sovremennyj russkij jazyk Moscow: Logos.

Vasmer, Max. 1973. Etimologičeskij slovar' russkogo jazyka Moscow: Progress.

Žuravlev, A. F. (ed.) 1974-2012. *Etimologičeskij slovar' slavjanskich jazykov. Praslavjanskij leksičeskij fond* (Vyp. 1-37). Moscow: Nauka.