Orthographic and Morphological Correspondences between Related Slavic Languages as a Base for Modeling of Mutual Intelligibility

Andrea Fischer, Klára Jágrová, 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

In an intercomprehension scenario, typically a native speaker of language L1 is confronted with output from an unknown, but related language L2. In this setting, the degree to which the receiver recognizes the unfamiliar words greatly determines communicative success. Despite exhibiting great string-level differences, cognates may be recognized very successfully if the receiver is aware of regular correspondences which allow to transform the unknown word into its familiar form. Modeling L1-L2 intercomprehension then requires the identification of all the regular correspondences between languages L1 and L2.

We here present a set of linguistic orthographic correspondences manually compiled from comparative linguistics literature along with a set of statistically-inferred suggestions for correspondence rules. In order to do statistical inference, we followed the Minimum Description Length principle, which proposes to choose those rules which are most effective at describing the data. Our statistical model was able to reproduce most of our linguistic correspondences (88.5% for Czech-Polish and 75.7% for Bulgarian-Russian) and furthermore allowed to easily identify many more non-trivial correspondences which also cover aspects of morphology.

Keywords: comparative linguistics, Minimum Description Length, receptive multilingualism

1. Introduction

Similarities in phonology, morphology, syntax and basic vocabulary among Slavic languages are striking, and the possibility of mutual intercomprehension within the Slavic language family is a generally accepted hypothesis. However, similarities and differences must be studied from different linguistic perspectives: orthography, morphology, syntax, and lexis. In the present work, we focus on orthographic and morphological relations in two language pairs that we consider representative for the Slavic language family: Czech and Polish (CS-PL) and Bulgarian and Russian (BG-RU). Czech and Polish are both West Slavic languages that use the Latin alphabet, but differ in their use of digraphs and diacritical signs. Russian and Bulgarian are East and South-East Slavic, respectively, and both use the Cyrillic alphabet.

Our hypothesis is that both orthography and morphology are linguistic determinants of mutual intelligibility which may facilitate or impede intercomprehension. Thus, we need to analyze the most frequent orthographic and morphological correspondences in order to uncover systematic similarities and differences between the two language pairs.

We analyzed lists of cognates in a previous experiment, investigating the statistical significance of rules formulated on the basis of comparative historical linguistics literature. This revealed that proceeding purely on the basis of knowledge from comparative linguistics did not give us all necessary rules – many occurring correspondences were missing.

In the present contribution, we discuss missing correspondences and turn our attention to the statistical inference of correspondence rules. Specifically, we use the Minimum Description Length principle (Grünwald, 2007) in order to supplement our comparative-linguistic correspondence rules. The lists and the compiled rules should serve as a resource for future comparative research.

The paper is structured as follows: we first outline our data and previous experiments in Section 2. Then, in Section 3., we briefly outline the model with which we obtain statistical correspondences. In Section 4., we discuss examples of correspondences suggested by our statistical model from a linguistic perspective before concluding in Section 5.

2. Preparatory Work

Data: In previous work, we carried out large-scale computational transformation experiments on parallel word sets. For this, we manually compiled orthographic correspondences based on traditional approaches and comparative historical linguistics. The first step was to assemble a suitable resource for an examination of orthography. We chose to use vocabulary lists instead of parallel sentences, as the latter allow for too many degrees of freedom. Thus, we collected and aligned parallel Slavic word lists for the two language pairs. For each pair, a list of internationalisms and a list of Pan-Slavic vocabulary were freely available from the EuroComSlav website.[1] A third parallel list of cognates we compiled from Swadesh lists for these languages (Swadesh, 1952).[2] Focusing mainly on the linguistic form

[1] Refer to (Angelov, 2004) and (Likomanova, 2004).

[2] http://en.wiktionary.org/wiki/Appendix:Swadesh_lists_for_Slavic_languages, accessed on 2015-04-22.

and ignoring some semantic shifts, we thoroughly modified and corrected these lists. Firstly, formal non-cognates, such as CS-PL *mnoho – wiele 'many/much'*; BG-RU *ние (nie) – мы (my)*[3] *'we'*, were removed. Secondly, existing formal cognates were added to the lists where the pairs consisted of non-cognates. As an example, *звяр (zvjar) 'beast'* was added to its Russian formal cognate *зверь (zver') 'animal, beast'* in order to obtain the BG-RU pair *звяр – зверь* instead of the 'semantic' pair *животно – зверь (životno – zver')* while e.g. *mężczyzna 'man'* was substituted by *mąż 'husband'* in CS-PL *muž – mąż*. This explains the variation in the amount of words in Table 1 for each list in each language pair.

language pair	CS-PL	RU-BG
internationalisms	262	261
Pan-Slavic vocabulary	455	447
Swadesh list	210	227

Table 1: Number of cognate pairs for both language pairs.

The words in the lists belong to different parts of speech. All adjectives are in their masculine form. All verbs in the BG-RU lists are 3rd person singular in order to deal with the non-existing infinitive in Bulgarian and for future comparison of all four languages. For the CS-PL Swadesh and Pan-Slavic lists, we compiled two versions: one with all verbs in infinitives and one with all verbs in 3rd person singular, for future comparison of all four languages.

Diachronically-based orthographic correspondences: In the second step, we manually collected a cross-linguistic rule set of corresponding orthographical units (transforming both individual letters and letter strings) from comparative historical Slavic linguistic literature (Bidwell, 1963; Žuravlev, 1974 2012; Vasmer, 1973). This resulted in sets of diachronically-based orthographic correspondences: 81 unique rules for CS-PL and 48 unique rules for BG-RU. These rules bridge mismatches between target and source language units, such as *(á:ią), (ě:ię), (z:dz), (hv:gw), (lou:łu), (ou:ą), (rů:ró), (ří:rze), (šť:szcz)* etc. (CS-PL); *(т:тъ) [t,t'], (б:бл) [b:bl], (в:вл) [v:vl], (жд:жс) [žd:ž], (м:мл) [m:ml], (п:пл) [p:pl], (ъ:у) [ă:u], (и:ъı) [i:y], (я:е) [ja:e], (ла:оло) [la:olo]* etc. (BG-RU).[4]

We then tested this set of diachronically-based orthographic correspondences in terms of their statistical importance on the parallel word lists mentioned above. The results of this purely orthography-based transformation experiment are described in (Fischer et al., 2015). By applying the transformation rules, we categorized the cognates in the pairs as either (i) identical, (ii) successfully transformed, or (iii) non-transformable by the rules. The rate of covered pairs (identical + successfully transformed) of the computational application of the above-mentioned three parallel lists ranged from 53.63% for CS-PL (for the Pan-Slavic list) to 67.82% for BG-RU (for the internationalisms lists). We found a notably higher rate of identical words within the BG-RU pair throughout all lists, whereas there was overall a greater range and quantity of transformation rules that could be successfully applied for the CS-PL combination. The successful results also allowed us to supplement the linguistic rules with additional, correct one-to-one correspondences. Detailed results are described in (Fischer et al., 2015).

We were left with fairly high numbers of words not covered by the comparative historical rules. We next aim to gain a better coverage of the cognate lists by more rules. The compiled transformation rules perform satisfactorily in that they do allow a glimpse into one facet (orthography) of linguistic similarity. Formal aspects beyond orthography, however, are equally important when accounting for differences between the languages and must therefore be taken into account. Most obviously, a large part of the words are non-transformable because of differing morphological elements. Examples for this are (i) zero endings vs. different endings of adjectives in the masculine form for BG-RU; (ii) different endings of verb forms in 3rd person singular present for BG-RU; (iii) most of the internationalisms for CS-PL categorized as non-transformable differ in suffixes and endings – both in pairs with same and different gender. More generally, differences may arise both from different characteristics such as grammatical gender and from differences simply not being orthographical, but structural. Thus, we add the morphological perspective to our analysis.

We next turn our attention towards the statistical model to aid in the task.

3. MDL for Correspondence Discovery

As our previous experiment suggests, proceeding purely on the basis of hand-crafted rules compiled from comparative linguistics literature is tedious work and still results in low coverage of word lists. Many necessary correspondences from different levels appear to be missing. This motivates the need for a statistical model to assist in the task.

The proposition to infer regular correspondences statistically is fairly old (Kay, 1964). Initial ideas focused on simply learning the most frequent correspondences, and many approaches from the past few decades are based on Levensthein alignments (Levenshtein, 1966). However, such approaches suffer from several drawbacks. Firstly, there is a combinatorial number of possible correspondences for any given set of cognate pairs, and therefore, distinguishing useful patterns from noise in the data is hard. Secondly, the use of Levenshtein distance is not well suited to comparing languages with differing alphabets.

There have also been a few attempts at implementing tools to aid in correspondence discovery (Lowe and Mazaudon, 1994; Covington, 1996). However, these tools only assist in finding examples for hypothesized correspondences and do not allow to easily find completely new ones.

We are looking for correspondences at different linguistic levels, so we require our model to be able to identify both grapheme-to-grapheme and morpheme-to-morpheme correspondences. There is a combinatory number of possible correspondences of this kind and designing a model is far from trivial. For our model, we employ the Minimum Description Length principle (MDL) (Grünwald, 2007). MDL

[3] Transliteration of Bulgarian and Russian words follows DIN 1460.

[4] Transliterations of rules given in square brackets.

proposes that good rules are those that *succinctly* describe the data.

Specifically, we use a model realized with a two-part code (Grünwald, 2007). The basic formula to use such a code is

$$\mathbf{M} = \arg\min_{M \in \mathcal{M}} \{L(M) + L(D|M)\}$$

where D is the data, in our case a list of parallel words, and M, the model, is one of the potential explanations for D. $L(M)$ is called the *description length* of the model, while $L(D|M)$ is the description length of the data given the model. By minimizing description length, we search for those correspondences that yield the most *precise* description of our word lists. Note that description lengths are, by Shannon's source coding theorem (Shannon, 2001), log-likelihoods. Thus, two-part MDL can be thought of as a regularized maximum likelihood approach.

We want our model to treat correspondences simply as associated strings of characters from the individual alphabets. Describing the model then boils down to specifying the sizes and usage counts of both alphabets along with the list of rules. An alphabet here is simply a collection of enumerated symbols, the exact identities of which are irrelevant. Thus, alphabets are specified simply via their lengths. The usage counts of each symbol from the alphabets are necessary to use the Shannon-optimal codes during transmission of the rules. Similarly, the usage counts of the correspondence rules must be transmitted.

In the following, count(x) denotes the occurrence count of x, and code(x) is the function that assigns x its Shannon-optimal code word.

In order to transmit lengths, sizes, and total counts, we utilize the *universal code for the integers*, $L_{\mathbb{N}}$, which is the best way of transmitting an integer of arbitrary size when no further information is known (Grünwald, 2007). To specify the information necessary to use Shannon-optimal codes, we transmit the distribution of counts via a data-to-model code (Grünwald, 2007).

We call our alphabets Σ_1 and Σ_2, respectively, and denote the list of rules $[(\pi_{1,1}, \pi_{1,2}), (\pi_{2,1}, \pi_{2,2}), ...]$ by Π. Our model then has a total description length $L(M = (\Sigma_1, \Sigma_2, \Pi))$ of

$$L(M) = L(\Sigma_1) + L(\Sigma_2) + L(\Pi).$$

Each alphabet is described with

$$L(\Sigma_i) = L_{\mathbb{N}}(|\Sigma_i|) + L_{\mathbb{N}}(T_{\Sigma_i}) + \log \binom{T_{\Sigma_i} - 1}{|\Sigma_i| - 1}$$

where $T_{\Sigma_i} = \sum_{\sigma \in \Sigma_i} \text{count}(\sigma)$; and the rule table Π is modeled via

$$L(\Pi) = L_{\mathbb{N}}(|\Pi|) + \sum_{\pi \in \Pi} L(\pi) + L_{\mathbb{N}}(T_{\Pi}) + \log \binom{T_{\Pi} - 1}{|\Pi| - 1} \quad ,$$

with $T_{\Pi} = \sum_{\pi \in \Pi} \text{count}(\pi)$. In order to describe a correspondence rule $\pi = (\pi_1 \in \Sigma_1^*, \pi_2 \in \Sigma_2^*)$, we transmit

$$L(\pi) = L_{\mathbb{N}}(|\pi_1| + 1) + L_{\mathbb{N}}(|\pi_2| + 1) + \sum_{\sigma \in \pi} L(\text{code}(\sigma)) \quad ,$$

i.e., we model rules simply by specifying the two strings they associate. In order to use the $L_{\mathbb{N}}$ to specify the lengths of the strings, we must offset the numbers by one, since $L_{\mathbb{N}}(x)$ is defined only for $x > 0$.

The data then can be modeled simply as lists of rules:

$$L(D|M) = L_{\mathbb{N}}(|D|) + \sum_{d \in D} L(d|M)$$

$$\text{where } L(d|M) = L_{\mathbb{N}}(|d|) + \sum_{\pi \in d} L(\text{code}(\pi)) \quad .$$

We infer rules by way of Expectation-Maximization (Dempster et al., 1977). New rules are constructed by merging two previously-known ones. The partitioning of each data entry into the different correspondence rules is computed by way of the Viterbi algorithm (Viterbi, 1967). Initially, we begin with no known correspondences, i.e. place each character of each word into a separate rule.

Usage in our scenario: It is easy to see that two-part MDL naturally guards against overfitting by weighing each rule's complexity against its utility. However, if the desired correspondences are not purely statistical in nature, as is the case here, then we may benefit from abusing the MDL formalism slightly. Using the model as a *ranking* mechanism rather than an exact prescription of the nature of rules, we can evolve all of our word pairs starting from zero known correspondences, up to the point where all word pairs are analyzed with a single word-to-word rule. Thus, we can observe which rules are found, in which order they are found, and from which previous rules each new rule is constructed. This allows to identify correspondences at the different linguistic levels and provides insight into the statistical importance of both finely-grained and coarse correspondences.

Example evolution paths: The model proposes a set of statistically important correspondences, which we illustrate by way of example in Figure 1. Correspondences are indicated by the boxes spanning different substrings, while the numbers at the lower right corners of the boxes indicate the step in which they were found.

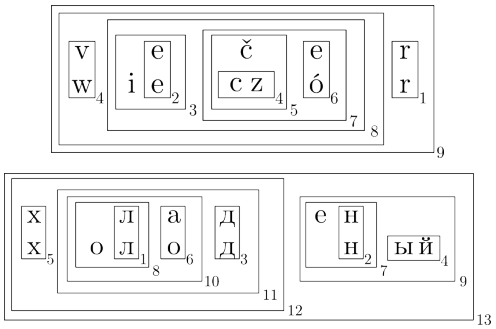

Figure 1: Examples for evolved correspondences.
Top: CS-PL *večer – wieczór* ('evening').
Bottom: BG-RU *хладен – холодный (chladen – cholodnyj)* ('cold').

In the CS-PL cognate pair *večer – wieczór 'evening'*, the model suggests the correspondences *(r:r), (e:e), (e:ie), (v:w), (:cz), (č:cz), (e:ó), (če:czó), (eče:ieczó),* and *(veče:wieczó)* (in this order) before placing both words into

one correspondence rule. In the BG-RU cognate *хладен – холодный (chladen – cholodnyj) 'cold'*, the model proposes, in order, the correspondences *(л:л) [l:l]*, *(н:н) [n:n]*, *(д:д) [d:d]*, *(:ый) [:yj]*, *(х:х) [ch:ch]*, *(а:о) [a:o]*, *(ен:н) [en:n]*, *(л:ол) [l:ol]*, *(ен:ный) [en:nyj]*, *(ла:оло) [la:olo]*, *(лад:олод) [lad:olod]*, and *(хлад:холод) [chlad:cholod]*. The order in which correspondences are discovered reflects their importance for the data. With this, the model allows expert linguists to choose from many different rules.

Comparison to linguistic rules: The question arises to what extent the model is able to replicate the diachronically-based rules. To answer it, we compare the lists of statistically discoverable rules with our hand-crafted linguistic rules. The resulting statistics are listed in Table 2. The details of this analysis can be found in the language resource accompanying this document.

language pair	CS-PL	RU-BG
# ling. rules	103	77
of those applicable	96	70
of these discovered	85 (88.5%)	53 (75.7%)

Table 2: Comparison statistical/hand-crafted rules.

While the model does not replicate all of the diachronically-based rules, not all of them are necessarily truly applicable. We check for applicability in the most general way possible: by simply seeing whether the strings associated by a rule are present in any of the word pairs. However, many of the applicable-but-not-found rules correspond to quite deep linguistic processes. As an example, the rule *(б:бл) [b:bl]* – intended to capture a historical phonetic correspondence that does not occur in the data – is counted as applicable due to the presence of BG-RU cognates such as *близък – близкий (blizăk – blizkij) 'close'*. Similar observations can be made for most of the other non-discovered rules.

After having outlined our model, we next turn our attention to the linguistic significance of learned correspondences. We are particularly interested to see to what extent the statistical rules can help to complete the hand-crafted linguistic rule sets.

4. Linguistic Utility of Statistical Results

We next discuss the linguistic aspects the correspondences discovered by our statistical model. We divide our discussion into two parts and treat orthography and morphology separately.

Our discussion does not claim to be exhaustive; rather, we intend to give the interested reader an impression of the extent to which rules found by our model may correspond to linguistic concepts.

For most of the discovered correspondences, we give the iteration in which they were found. The iteration in which the model found each correspondence reflects the statistical relevance of the correspondence. Generally, the earlier a rule was first found, the more relevant it is. In our data, we add '%' and '# ' symbols as start and end markers, respectively. Not surprisingly, the first two iterations always result in correspondences between (%:%) and (# :#). This effectively leads to an artificial offset of iteration numbers for all other rules.

In total, the model performed 3010 steps for CS-PL, and 2696 steps for RU-BG, corresponding to the discovery of 3010 and 2696 potential new rules, respectively.

4.1. Orthographic Correspondences

In our previous experiment, we originally formulated 103 unique diachronically-based orthographic correspondences for CS-PL and 77 unique correspondences for BG-RU, including equal-to-equal correspondences (e.g., *(y:y), (u:u), (m:m), (re:re)*; *(б:б) [b:b], (г:г) [g:g], (к:к) [k:k], (п:п) [p:p], (т:ть) [t:t'], (б:бл) [b:bl], (в:вл) [v:vl], (жд:жс) [žd:ž], (м:мл) [m:ml], (п:пл) [p:pl], (а:а) [a:a], (е:е) [e:e], (ъ:у) [ă:u], (и:ы) [i:y], (я:е) [ja:e], (ла:оло) [la:olo]* etc.). In the first transformation experiment (Fischer et al., 2015) we used only those correspondences which represented orthographic mismatches between target and source language units (e.g., *(e:ie), (é:a), (d:dz), (šť:szcz), (lou:łu); (т:ть) [t:t'], (б:бл) [b:bl], (в:вл) [v:vl], (жд:жс) [žd:ž], (м:мл) [m:ml], (п:пл) [p:pl], (ъ:у) [ă:u], (и:ы) [i:y], (я:е) [ja:e], (ла:оло) [la:olo]* etc.). Thus only 48 correspondences were applied on parallel word lists for the BG-RU mapping, and 81 for the CS-PL mapping.

We next discuss newly-discovered orthographic correspondences.

Czech and Polish both use the Latin alphabet, but with different systems of diacritical signs. Firstly, Czech has a repertoire of letters with diacritics for which Polish often uses digraphs. Secondly, the languages use different diacritical signs. Furthermore, there are sound correspondences that are represented differently in orthography, which were not accounted for by the previous rule set.

Czech uses two basic diacritical signs: ´ for marking a long vowel (plus the circle in *ů* as an alternation of *ú* which is used only at stem onset) and ˇ (*háček*) in the consonants *č, ň, š, ř, ž* and its alternation ’ (*klička*) in *ď* and *ť*. The *háček* on top of *ě* palatalizes the preceding consonant.

Polish has four different diacritical signs: ´ (*kreska*), used in the vowel *ó* and performing a similar function to Czech *háček* in *ń, ć, ś*, and *ź*; the overdot (*kropka*) used only in *ż*; the *ogonek* ˛ used in *ą* and *ę*; and the stroke used in *ł*.

The Czech letters *á, č, ď, é, ě, ch, í, ň, ř, š, ť, ú, ů, ý, ž* as well as *q, v,* and *x* are not part of the Polish alphabet, and the Polish letters *ą, ć, ę, ł, ń, ś, w, ż* and *ź* do not exist in Czech, although *w* appears in Czech in foreign named entities and loan words.

The Czech characters *č* or *š* can correspond to the digraphs cz (e.g. CS-PL *tečka – teczka 'dot'*; *(č:cz)* suggested in iteration 53) and *sz* (e.g. in CS-PL *veš – wesz 'louse'*; *(š:sz)* suggested in iteration 54) – both were part of the original rule set. However, *sz* can also correspond to *ś* in Polish: *(š:ś)* was found in iteration 212.

A general tolerance of diacritical signs is reasonable in a reading intercomprehension scenario, where readers could simply delete unknown elements around graphemes that they are otherwise familiar with.[5] This applies accordingly

[5]This fact can be modelled by distinguishing BASE and DIA-

for tolerance of diacritical signs in rules found for pairs such as *jazyk – język 'language/tongue'* (*(a:ę)* is found in iteration 151), *zvíře – zwierzę 'beast'* (*(e:ę)* is found in iteration 416), *široký – szeroki 'broad'* (*(i:e)* is found in iteration 231). These pairs were previously categorized as untransformed because the set of correspondences allowed only transformations of CS-PL *(á:ę)*, *(ě:ię), (e:e)* and *(í:e)*.

Throughout all three word lists, there were no striking differences in the rates of non-transformable cognates (min. 43.40% in the Swadesh list; max. 46.37% in the Pan-Slavic list) in our previous experiments. When analyzing the non-transformable 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 the rule set needs to be extended to account for correspondences involving characters with or without diacritics in both transformational directions: For example, the original set of correspondences allowed a transformation of the CZ *é* to the PL *a* or *ie* only. However, e.g., the pairs CS-PL *plést – pleść 'to knit'* or *déšť – deszcz 'rain'* are instances that demand a similar rule tolerating the absence of the diacritical sign above the grapheme *e*. These correspondences are found by the model in iteration 94 (*(e:a)*) and 26 (*(e:ie)*). Another correspondence that becomes apparent in those two lists is CS-PL *(k:g)* (found in iteration 150). In CS-PL *kde – gdzie 'where'*, 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. This can be explained by the fact that in this case Czech retains the original *k* whereas Polish prefers the phonetic rendering *g* (Kellner, 1936).

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)* (iteration 40), *(s:ś)* (iteration 50), e.g., in the pairs *živý – żywy 'alive'*, *světlý – światły 'bright'*. Previously formulated correspondences allowed only for *(i:i), (í:i), (sť:szcz)* (here, tolerating diacritics would be necessary again), and *(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 rendered in speaking and writing in the two languages, with adaptations relying more or less heavily on the original internationalism. As examples, consider CS-PL *mač – mecz 'match'*, *leasing – lis 'leasing'*, *apartmá – apartament 'appartment'*. Polish uses *ks* instead of *x*: CS-PL *maximum – maksymum*, *export – eksport* (*(x:ks)* is identified in iteration 190). Furthermore, there are no exceptions for internationalisms in Polish orthography, in contrast to Czech. This becomes apparent when comparing the pairs CS-PL *legitimace – legitymacja*, *kredit – kredyt*, *praktika – praktyka*, *medicína – medycyna* (*(i:y)* from iteration 40). 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 (*(ti:ty)* is found in iteration 247; *(di:dy)* in iteration 1461. This rule occurs in only one pair). The phonetic principle seems to be obeyed more strongly in Polish orthography than in Czech orthography when looking at internationalisms.

Bulgarian and Russian use the Cyrillic alphabet. Three letters of the Russian alphabet do not occur in Bulgarian: *ы, э, ё*. 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*.

The diachronically-based orthographic correspondences that were applied lack many of the possible orthographic correlates (e.g., the RU-BG internationalism list exhibited the lowest rate with 5.36% of correctly transformed words based on the original set of correspondences). The internationalism list based on EuroComSlav requires many additional BG-RU correspondences. Examples for statistical rules found by our model and deemed linguistically meaningful are: *(ъо:ё) [ʹo:ё]* (e.g. in *актьор – актёр (aktʹor – aktër) 'actor'* or in *партньор – партнёр (partnʹor – partnër) 'partner'*, found in iteration 166); *(е:э) [e:ė]* (e.g. in *економия – экономия (ekonomija – ėkonomija) 'economy'*, or *експорт – экспорт (eksport – ėksport) 'export'*, *енергия – энергия (energija – ėnergija) 'energy'* etc., found in iteration 146); *(n:nn) [p:pp]* (e.g. in *апарат – аппарат (aparat – apparat) 'administration, mechanism'*, *апетит – аппетит (apetit – appetit) 'appetite'*, found in iteration 240); *(c:cc) [s:ss]* (e.g. in *бос – босс (bos – boss) 'boss'* or *дискусия – дискуссия (diskusija – diskussija) 'discussion'*, found in iteration 141).

However, the model did not recognize the following regular orthographic correspondences between Bulgarian and Russian (Gribble, 1987; Valgina et al., 2002; Ivanova et al., 2011): *(л:лл) [l:ll]* (*алигатор – аллигатор (aligator – alligator) 'alligator'*, *колега – коллега (kolega – kollega) 'colleague'*); *(p:pp) [r:rr]* (*перон – перрон (peron – perron) 'platform'* etc.); *(н:нн) [n:nn]* (*тунел – тоннель (tunel – tonnel') 'tunnel'* etc.).

Taking into consideration systematic phonological and morphological aspects, e.g., Bulgarian /ə/ will most often correspond to /u/ in Russian (both from the back nasal vowel of Common Slavic */ǫ/): BG *зъб (zăb), път (păt), ръка (răka)* – RU *зуб (zub), путь (putʹ), рука (ruka) 'tooth', 'road', 'hand'/'arm'*. In suffixes, and rarely in roots, when *ъ (ă)* is or was a mobile vowel it will correspond to *о (o)* in Russian (Gribble, 1987): BG *зъл (zăl), зла (zla)* – RU *зол (zol), зла (zla) '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 (zăl)* – RU *злой (zloj) 'wicked'*. The lack of some BG-RU orthographic correlates, e.g., *(ъ:ø) [ă:ø]*, and differing morphological features could explain the amount of non-transformable adjectives in the Pan-Slavic and Swadesh lists.

From the Common Slavic *ь in the so-called strong position we get *е (e)* in both Bulgarian and Russian: BG *отец (otec), ден (den)* – RU *отец (otec), день (denʹ) 'fa-*

CRITIC parts in characters with diacritics, with the hypothesis that differences in BASE would be stronger than differences in DIACRITIC. The effect of differences in diacritics will also be tested in web-based reading intercomprehension experiments.

ther', 'day'. However, in a few words, Bulgarian ъ *(ă)* may correspond to Russian *e (e)* or *ë (ë)* (Gribble, 1987): BG *пъстър (păstăr)* – RU *пёстрый (pëstryj) 'colorful'*; BG *тъмно (tămno)* – RU *темно (temno) 'dark'*. All these orthographic correspondences were found by our MDL-based model: *(ъ:ø) [ă:ø]* (*гладък – гладкий (gladăk – gladkij)* 'smooth'); *(ъ:е) [ă:е]* (*лъв – лев (lăv – lev) 'lion'*); *(ъ:ё) (тъмен – тёмный (tămen – tëmnyj) 'dark'*).

4.2. Morphological Correspondences

Our original rule set did not account for morphological correlates. In the output of our model, however, we can also find morphological correspondence rules. In linguistics, it is generally accepted to distinguish between derivational and inflectional morphology. In both cases, one deals with certain ensembles of units (Akhmanova, 1971). However, derivational and inflectional aspects may interfere with each other on the surface of words, and even orthography may play a role in the exact manifestation of morphological elements. Thus, oftentimes, inflectional and derivational aspects have to be considered jointly in order to formulate correspondences based on morphological features.
Nonetheless, our model revealed some systematic correspondences belonging to the realm of morphology. We begin with discussing inflectional morphology.

4.2.1. Inflectional Morphology

Our cognate lists contain nouns only in nominative singular form, adjectives only in their masculine singular forms and verbs only in their 3rd person singular forms. Due to this, the underlying lists do not allow for a comparison of the complete inflectional morphological systems of the languages. The focus here thus lies rather on the extent to which the correspondences between the inflectional endings of the words in the lists can be described by the statistical alignment, in particular for segmentation of a word form into stem and inflection.
Do note that since we added beginning markers ('%') and end markers ('# ') to words, our rules allow to explicitly distinguish between initial and final positions within the words, which lends itself well to the inflectional schemes of the Slavic languages under consideration.

Czech and Polish: The statistical model presents some correspondences in inflectional endings that were not covered by the diachronically-based orthographic correspondence rules. Some examples for newly gained correspondences are *(e#:a#)* (iteration 35) for feminine nouns, *(ý#:i#)* (iteration 76) for masculine adjective endings. Most of the internationalisms that were categorized as untransformed in our previous test differ in their endings, often because of being of different gender in the two languages, e.g., CS-PL *univerzita – uniwersytet 'university'*, *teritorium*[6] *– terytoria 'territory'* (*(um#:a#)* in iteration 250), *recept – recepta 'recipe'*, *sál – sala 'hall'*, *salát – sałata 'salad'* (*(#:a#)* in iteration 20), but sometimes having different endings despite having the same gender, such as in *penze – pensja 'pension'* (*(e#:a#)* in iteration 35, even before *(e:a)* discovered in iteration 94).
Although the model correctly suggested the *(#:a#)* correspondence in *recept – recepta*, it did not suggest *(a#:#)* in the example *univerzita – uniwersytet*. In the latter case it aligned the two separate correspondences *(a:e)* and *(#:t#)* that consequently must be statistically more meaningful. This suggests that it is more often the case that CZ nouns end with an *–a* (feminine nouns and some masculine nouns of the *předseda* paradigm) where there is no ending in Polish than vice versa. The model also discovered the 3rd person verb endings *(e#:ie#)*, in iteration 79. Besides these, the model was successful in discovering identical inflectional correlates.

Bulgarian and Russian: The most characteristic feature of Bulgarian inflectional morphology is its loss of case except for vocative forms and remnants (nominative, accusative, dative) in the pronoun system (Gribble, 1987; Townsend and Janda, 1996). Bulgarian and Russian nouns are divided into three genders: masculine, feminine, and neuter. These distinctions are usually reflected by different endings. In most cases Bulgarian nouns have the same gender division as in Russian (Gribble, 1987), but there are some differences. Masculines can end in *-а (-a)*, as in Russian, but they may also end in *-о (-o)*, as in *чичо (čičo) 'paternal uncle'*. Bulgarian nouns referring to persons may be neuter and end in -е (-e): *момче (momče) 'boy' момиче (momiče) 'girl'* in contrast to Russian. Among the BG-RU correspondences that the model suggested are the following correspondences of noun endings, sorted by frequency upon discovery[7] (iteration in brackets): e.g., *(а#:а#)* 149 (12) (for feminine); *(я#:я#)* 40 (38) (for feminine); *(о#:о#)* 36 (45) (for neuter) etc. However, the last ending is ambiguous and may also be an adverb ending, for example, BG-RU:*много – много (mnogo – mnogo) 'a lot of'*. There are 6 examples of BG-RU adverbs with the inflectional correspondence *(о#:о#)*.
Most Bulgarian adjectives have the zero ending *-ø* for the masculine forms with some exceptions with the suffix *-ск- (-sk-)*. These adjectives have the ending -и (-i). Russian adjectives have the following endings for the masculine form: *-ый (-yj), -ий (-ij)* or *-ой (-oj)*. The following masculine adjective endings were suggested, sorted by frequency (iteration numbers in brackets): *(#:ый#)* 57 (28); *(#:ий#)* 17 (82); *(#:ой#)* 15 (87).
Unlike Russian, Bulgarian has no infinitives, so we used the 3rd person singular present tense verb forms to compare these languages. There are three regular conjugations in Bulgarian in contrast to the two regular conjugations in Russian. The Russian second conjugation (with *-и-*) corresponds to the Bulgarian second, but the Russian first conjugation splits up into the first and third in Bulgarian (Gribble, 1987). However, we assume that Russian verbs of the

[6] Although *–um* is a Latin suffix, it is replaced by regular inflectional endings in declension (except in the 1st (nominative), 4th (accusative) and 5th (vocative) case singular), cf. Ústav pro jazyk český Akademie věd České republiky Internetová jazyková příručka. `http://prirucka.ujc.cas.cz/?id=263`, accessed 03-09-2016.

[7] As usage of a correspondence may change as new correspondences are discovered in other parts of the data, we here give only the frequency at introduction as frequency measure.

first or the second regular conjugations may correspond to Bulgarian verbs of the first, the second or the third regular conjugations. The following BG-RU inflectional correspondences of the verb forms were suggested by the model, sorted by frequency (iteration in brackets): *(u#:um#)* 27 (56); *(#:em#)* 25 (61); *(e#:ëm#)* 21 (71); *(e#:em#)* 14 (91); *(u#:em#)* 4 (266) *(u#:ëm#)* 2 (360). However, the model did not recognize the following ending correspondence: *(ø:-um) [ø:-it]* in *дыша – дышит (diša – dišit) 'breathes'*.

4.2.2. Derivational Morphology

Due to the aforementioned intermingling of different levels, analysis of the extent to which the model captures derivational morphology is much more complex than analyses regarding orthography or inflectional morphology. We therefore present only a very preliminary analysis.

Unsurprisingly, derivational morphological correspondences tend to be discovered later than inflectional and orthographical ones. This is as linguistically expected, as derivational processes are observable on the stems, i.e. only after segmentation into stem and inflection has occurred.

Czech and Polish: Additional correspondences suggested by the model that reveal correlates of affixes are for instance *(st:ść)* (iteration 301), respectively *(ost#:ość#)* (iteration 303) as feminine suffixes in nouns such as e.g. *mladost – młodość 'youth'*, the identical masculine noun suffix *(ek#:ek#)* (iteration 314), but also mismatching suffixes such as *(ek#:#)* (iteration 550), *(ec#:#)* (iteration 645). The correspondence between the feminine suffixes *–c(e)* and *–cj(a)*, such as in *legitimace – legitymacja 'legitimation'* were discovered as *(ce#:cja#)* in iteration 115, and *(i(e)#:i(a)#)* was discovered in iteration 154. For adjectives, *(ný#:ny#)* (iteration 77), *(ký#:ki#)* (iteration 104) and *(lý#:ly#)* (iteration 117), and *(vý#:wy#)* (iteration 172) are amongst the earliest suggested.

Bulgarian and Russian: Feminine abstract nouns in Bulgarian and Russian are formed from adjectival bases with the productive derivational suffixes: *-ост (-ost)* for BG nouns vs. *-остъ (-ost')* for RU nouns. This suffix correspondence was discovered by our model in iteration 2110 of 2696. It was used in multiple words: BG-RU *радост – радостъ (radost – radost') 'joy'*; *младост – молодостъ (mladost – molodost') 'youth'*; *старост – старостъ (starost – starost'), 'old age'* .

5. Conclusion

We studied systematic orthographic and morphological correspondences between two pairs of related Slavic languages: Czech-Polish and Russian-Bulgarian. We analyzed both hand-crafted diachronically-based and statistically-inferred correspondence rules. For statistical inference, we used a model based on the Minimum Description Length principle. With the help of this model, we were able to replicate the linguistic rules to a very large extent and discovered many additional non-trivial correspondences, which cover orthography and inflectional morphology well.

The combination of our statistical model and expert knowledge is very promising for future work in comparative linguistics. In growing our rules, we proceeded without any interference from an expert linguist. Shaped into a tool, our model will greatly facilitate the formulation of correspondence rules. For this, the model would simply propose candidate rules, the user would select those to be kept, and the model would then propose new sets of rules while keeping the ones the user selected. In this way, all the relevant correspondences can be presented and chosen from easily.

In future work, we will focus on linguistically refining the model such that it is able to capture processes and measure linguistic distances at the different levels concurrently. We will place particular focus on derivational morphology and provide further, detailed linguistic analyses of our models.

6. Acknowledgements

This work has been funded by Deutsche Forschungsgemeinschaft (DFG) under grant SFB 1102: Information Density and Linguistic Encoding.

7. Bibliographical References

Akhmanova, O. (1971). *Slavic Historical Phonology in Tabular Form*. Mouton, The Hague, Paris.

Angelov, A. (2004). EuroComSlav Basiskurs - der panslavische Wortschatz. `http://www.eurocomslav.de/BIN/inhalt.htm`, accessed 2016-02-17.

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

Covington, M. A. (1996). An algorithm to align words for historical comparison. *Comput. Linguist.*, 22(4):481--496, December.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, 39(1):1--38.

Fischer, A. K., Jágrová, K., Stenger, I., Avgustinova, T., Klakow, D., and Marti, R. (2015). An orthography transformation experiment with Czech-Polish and Bulgarian-Russian parallel word sets. In B. Sharp, et al., editors, *NLPCS*, pages 115--127, Krakow, Poland. CaFoscarina Editrice, Venezia.

Gribble, C. E. (1987). *Reading Bulgarian through Russian*. Slavica Publishers, Inc., Columbus, Ohio.

Grünwald, P. D. (2007). *The minimum description length principle*. Adaptive computation and machine learning. Cambridge, Mass. MIT Press.

Ivanova, E. J., Šanova, Z. K., and Dimitrova, D. (2011). *Bolgarskij jazyk*. Karo, St. Petersburg.

Kay, M. (1964). The logic of cognate recognition in historical linguistics. Research memorandum, CA: RAND Corporation, Santa Monica, July.

Kellner, A. (1936). *Revise polského pravopisu. Slovo a slovesnost. Ústav pro jazyk český*. Akademie věd České republiky.

Levenshtein, V. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707.

Likomanova, I. (2004). EuroComSlav Basiskurs - der panslavische Wortschatz. `http://www.eurocomslav.de/kurs/iwslav.htm`, accessed 2016-02-17.

Lowe, J. B. and Mazaudon, M. (1994). The Reconstruction Engine: A Computer Implementation of the Comparative Method. *Computational Linguistics*, 20(3):381--417. http://www.aclweb.org/anthology-new/J/J94/J94-3004.pdf.

Shannon, C. E. (2001). A mathematical theory of communication. *SIGMOBILE Mob. Comput. Commun. Rev.*, 5(1):3--55, January.

Swadesh, M. (1952). Lexico-statistic dating of prehistoric ethnic contacts. *Proceedings of the American Philosophical Society*, 96(4):452--463.

Townsend, C. E. and Janda, L. A. (1996). *Common and Comparative Slavic: Phonology and Inflection with special attention to Russian, Polish, Czech, Serbo-Croatian, Bulgarian*. Slavica Publishers, Inc., Columbus, Ohio.

Valgina, N. S., Rosental', D. Ė., and Fomina, M. I. (2002). *Sovremennyj russkij jazyk*. Logos, Moscow.

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

Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, IT-13(2):260--269, April.

A. F. Žuravlev, editor. (1974-2012). *Etimologičeskij slovar' slavjanskich jazykov. Praslavjanskij leksičeskij fond.* Nauka, Moscow.