

# A Knowledge-Rich Approach to Measuring the Similarity between Bulgarian and Russian Words

Svetlin Nakov

Faculty of Mathematics and Informatics  
Sofia University "St. Kliment Ohridski"  
5 James Boucher Blvd, Sofia, Bulgaria  
nakov @ fmi.uni-sofia.bg

Elena Paskaleva

Linguistic Modeling Laboratory  
Bulgarian Academy of Sciences  
25A Acad. G. Bontchev St, Sofia, Bulgaria  
hellen @ lml.bas.bg

Preslav Nakov

Department of Computer Science  
National University of Singapore  
13 Computing Drive, Singapore  
nakov @ comp.nus.edu.sg

## Abstract

We propose a novel knowledge-rich approach to measuring the similarity between a pair of words. The algorithm is tailored to Bulgarian and Russian and takes into account the orthographic and the phonetic correspondences between the two Slavic languages: it combines lemmatization, hand-crafted transformation rules, and weighted Levenshtein distance. The experimental results show an 11-pt interpolated average precision of 90.58%, which represents a sizeable improvement over two classic rivaling approaches.

## Keywords

Orthographic similarity, phonetic similarity, cross-lingual transformation.

## 1. Introduction

We propose an algorithm that measures the extent to which a Bulgarian and a Russian word are perceived as similar by a person who is fluent in both languages. Leaving aside the full orthographical identity, we assume that words with different orthography can be also perceived as similar when they have the same or a similar stem and inflections, as in the Bulgarian word *афектирахме* and the Russian *аффе́ктировались* (both meaning ‘we were affected’).

Bulgarian and Russian are closely related Slavonic languages with rich morphology, which motivates us to study the typical orthographical correspondences between their lexical entries (conditioned phonetically and morphologically), which we use to formulate and apply transformation rules for bringing a Russian word close to Bulgarian reading and vice versa. Our algorithm for measuring the similarity between Bulgarian and Russian words first reduces the Russian word to an intermediate Bulgarian-sounding form and then compares it orthographically to the Bulgarian word. The algorithm starts by transliterating the Russian word with the Bulgarian alphabet, and then transforms some typical Russian morphemes and word parts (e.g., prefixes, suffixes, endings, etc.) to their Bulgarian counter-parts. Since both Bulgarian and Russian are highly-inflectional languages, lemmatization is used to convert the wordforms to their lemmata in order to reduce the differences at the morphological level. Finally, the orthographic similarity is measured using a modified Levenshtein distance with letter-specific substitution weights.

## 2. Method

The normalization of the Bulgarian and the Russian words into corresponding intermediate forms has phonetic and morphological motivation and is performed as a sequence of steps, which will be described below.

### 2.1. Transliteration from Cyrillic to Cyrillic

In a strict linguistic sense, *transcription* is the process of transition from sounds to letters, i.e., from speech to text; it is carried out generally in a monolingual context. In a bilingual context, the notion of *transliteration* is used to denote the transition of sounds and their letter correspondences in one language to letters in another language. The term *transliteration* is commonly used for the transition of letters when the two languages use different alphabets. In this paper, we deal with *transliteration* since we work with written texts.

The linguistic objective of our investigation is to introduce more formal criteria to the investigation of possible *cognates* between Russian and Bulgarian. By *cognates* we mean words with equal or close orthography denoting the same meaning; words with equal/close orthography but different meaning are *false cognates/friends*. For their further investigation in multilingual research, we need to define the exact expression of that identity/closeness by particular metrics and procedures.

For a pair of languages from different families, the source of cognates is borrowing between them or from a third language. Besides borrowing, an essential source of cognates in related languages is their common protolanguage. However, in the historical development of both languages, three factors lead to different grapheme shape for fully identical words: (1) language-specific phonetic laws and resulting changes, (2) settings of the spelling systems regulating the *sound-letter* transition, and (3) divergence in the grammatical systems and the grammatical formatives.

#### 2.1.1 Full coincidence (equality) of letters

Both Russian and Bulgarian use the Cyrillic alphabet in their writing systems, but Russian uses two letters not present in Bulgarian: *ы* and *э*. Most other letters generally show a full coincidence with some exceptions to be listed in the following subsections. The list below presents the

full identity of Cyrillic letters in both languages in the cognates: *азбука – азбука*, *буква – буква*, *воля – воля*, *гипс – гипс*, *дух – дух*, *езда – езда*, *жена – жена*, *закон – закон*, *истина – истина*, *йод – йод*, *кипарис – кипарис*, *лак – лак*, *монета – монета*, *нож – нож*, *опора – опора*, *пост – пост*, *река – река*, *сом – сом*, *том – том*, *ум – ум*, *факт – факт*, *химия – химия*, *царь – цар*, *чай – чай*, *шум – шум*, *щит – щит*, *юг – юг*, *яхта – яхта*

As the above list shows, the full identity of the grapheme shape of cognates is manifested mainly when the transformed letter is in initial position.

### 2.1.2 Regular letter transitions

*Replacing Russian letters that are missing in the Bulgarian alphabet.* The transitions discussed here stem from historic differences in the phonetic and the spelling systems of the two languages. Bulgarian and Russian differ in their contemporary phonetic system mainly at the level of pronunciation; in the distinction of soft and hard consonants. The Russian-specific letters *ы* and *э* serve to denote the variant of a ‘hard consonant+u/e’ while in Bulgarian all consonants preceding *u* and *e* are soft. This basic difference of the phonetic systems gives us the regular correspondence *ы-и* and *э-е* in all Russian-Bulgarian cognates containing these two letters, e.g., *рыба – риба*, *поэт- поет*.

*Removing a Russian letter.* Another regular phonetic difference between the two languages, which is also related to the opposition *soft/hard*, is the allowed softness of a consonant preceding another consonant (*пальто*) or in final position (*шесть*). Such phonetic combinations are not allowed in Bulgarian: see the corresponding *палто* and *шест*. This regularity allows us to remove all Russian *ь* in these positions in the initial stage of the process of cognate comparison.

*Partial regularity of the letter transitions.* In non-initial positions, other not so regular but repeated letter correspondences can be observed, e.g., *е-я* in *хлеб-хляб*, *е-ъ* in *серп-сърп*, *о-ъ* in *сон-сън*, *у-ъ* in *муж-мъж*, etc. The iterativity of such transitions is due to the specific development of the spelling systems in the two languages. One such example is the disappearance of some Old Slavic letters and their regular replacement with different letters in Russian and Bulgarian. The above-mentioned change *у-ъ* is due to the disappearance of an Old Slavic letter called ‘big yus’ and its regular replacement by different vowels in all contemporary Slavic languages. The transition is only partially regular since not all occurrences of the letter have the same etymological origin.

### 2.1.3 Transformations of n-grams

The *sound-letter* transition legitimated by the spelling rules of the two languages is specific as well; its specificity is observed at the level of the grapheme composition of the full cognates, i.e., those that are borrowed from third languages or that are identical morphologically.

*Transformations originating from spelling.*

A fundamental difference between Russian and Bulgarian spellings is the treatment of *double consonants*. Russian allows them in every part of the word structure, while in Bulgarian they are only possible at the morpheme boundary. Thus, all words borrowed from third languages keep their double consonants in Russian, but lose them in Bulgarian, e.g., *процесс – процес*, *аффект – афект*, etc. In this way, a regular transition *ll-l* can be formulated for all double consonants with the following stipulation of grammatical origin.

In words of Slavic origin, consonant doubling occurs mainly at the morpheme boundary, but in Russian the phenomenon is more frequent since Russian spelling rules are more ‘phonetic’. For example, they reflect the change *voiced-voiceless* for all prefixes ending with *з* and preceding the initial *с* of the next morpheme. Bulgarian spelling is more ‘morphological’ and conservative; it keeps the *з* in writing, although it is voiceless in pronunciation, e.g., *рассуждение – разсъждение*, *бессмертный – безсмъртен*, etc. This transformation of *hard-soft* consonants in the final prefix position is only valid for the couple *з-с*. Thus, the Bulgarian-Russian transition *зс-сс* can be formulated as regular for prefixes only and cannot be viewed as a universal for other parts of the word, e.g., *кавказский – кавказки*.

Next, the following general question in treating double consonant correspondences arises: if we want to stay in the domain of uni- and bigram transformations, removing the second consonant in Russian can be ambiguous *поддержать – поддържам*, but *буддист – будист*, *вводить – въвеждам*, *раввин – равин*. The legal consonant doublings in Bulgarian can be only outlined in a larger context – a window of up to five letters, containing the prefix and the next consonant, as in *предд*, *надд*, *подд*, *изз*, *разз*, etc., where the second consonant should be preserved. Note that these exceptions from the rule are only valid for double *д*, *з* and *в* – final letter of prefixes, and for *н* – first letter of the affix *н*, e.g., *непременно – непременно*, but *аннотация – анотация*.

*Transformations of morphological origin.*

In addition to the divergent development of phonetic and spelling systems, the two languages develop different grammatical systems, both at a systemic and at a morphemic level – different categories with different graphemic expressions. That divergence leads to different grapheme shapes for words that are lexically conceived as cognates, e.g., *жены – жената*, and the difference is manifested in the ending part of the word, consisting of affixes, and ending and related to grammatical forms.

The transformations are made in two directions and for both languages. They can consist of removal of a letter sequence or its transformation.

#### 1. Removing agglutinative morphemes.

Each of the two languages has one agglutinative mechanism of word formation (but for different parts of speech) – the reflexive morpheme *ся* and *сь* in Russian verb conjugation and the postpositioned article in Bulgari-

an in nominal inflections (for nouns and adjectives). The corresponding grammatical meanings are expressed in the twin language by other means (the article is totally missing in Russian and the reflexivity of verbs is expressed by a lexical element in Bulgarian – the particle *се*). Thus, removing these morphemes is the first step in the process of conversion to an intermediate form, e.g., *веселиться – веселить*, *квадратът – квадрат*. Note that the Russian agglutinative morpheme *ся/сь* at the end of the word are non-ambiguous: all 212,000 wordforms with the ending *ся* in our Russian grammatical dictionary are reflexive verb forms. This is not the case with the Bulgarian article, where only removing the morpheme *ът* for masculin is non-ambiguous, while removing *та*, *ят* and other article morpheme can trim the stem, e.g., *жена-та*, but *квадрат-а*. We intentionally do not derive a transformation rule from the last correspondence.

Removing Bulgarian articles depends on the accepted conception about the place of lemmatization in the algorithm – should we set the orthographic similarity for all four members of the language pair – lemmata and wordforms – or should we measure the similarity at the lexical level only – the lemmata. In the latter case, no removal is necessary (see 1.3)

## 2. Transforming ending strings.

There is a big group of adjectives in the two languages derived from other parts of speech and formed with the suffix *н* and an adjectival ending, e.g., *шум – шумный*, *шум – шумен*. When the adjective is derived from a noun ending with *н*, we get a doubled *н* in the Russian lemma and in the Bulgarian wordforms, e.g., *гарнизон-гарнизонный* and *гарнизон – гарнизонни*. Another regular correspondence is manifested in the word derivation with the suffix *ск*. All these combinations of *н/нн/ск* and different adjectival endings give the correspondences shown in Table 1.

Russian Ending	Bulgarian Ending	Examples
-нный	-нен	<i>военный</i> → <i>военен</i>
-ный	-ен	<i>вечный</i> – <i>вечен</i>
-нный	-нен	<i>ранний</i> → <i>ранен</i>
-ний	-ен	<i>вечерний</i> → <i>вечерен</i>
-ский	-ски	<i>вражеский</i> → <i>вражески</i>
-ый	-и	<i>стрелковый</i> – <i>стрелкови</i>
-ной	-нен	<i>стенной</i> – <i>стенен</i>
-ной	-ен	<i>родной</i> – <i>роден</i>
-ой	-и	<i>деловой</i> – <i>делови</i>

Table 1: Transforming Russian adjectives to Bulgarian.

For verbs, there are some regularities in the correspondences of the endings of the Russian infinitive and the Bulgarian verb's main form in first person singular. Table 2 below shows some examples.

Russian Ending	Bulgarian Ending	Examples
-овать	-ам	<i>декорировать</i> → <i>декорирам</i>
-ить, -ять	-я	<i>бродить</i> → <i>бродя</i> <i>блещать</i> → <i>блещя</i>
-ать	-ам	<i>давать</i> → <i>давам</i>
-уть	-а	<i>гаснуть</i> – <i>гасна</i>
-еть	-ея	<i>белеть</i> → <i>белея</i>

Table 2 – Transformation of Russian verbs to Bulgarian.

Concerning the transformation of endings, it is important to note that two linguistic problems are interrelated here: (1) the formal revelation of the morpheme boundary, and (2) the correct correspondence with the Bulgarian ending. The existing ambiguity in resolving these two problems requires serious statistical investigations before the rules can be formulated.

With ambiguity not taken into account, the proposed transformation rules for Russian word endings could sometimes generate the wrong Bulgarian wordform, e.g., *висеть* could become *висея*, while the correct Bulgarian form is *вися*. In order to limit the negative impact of that, we measure the similarity (1) *with* and (2) *without* applying rules for lemmatization; we then return the higher value of the two.

## 2.2. Lemmatization

Bulgarian and Russian are highly-inflectional languages, i.e., they use variety of endings to express the different forms of the same word. When measuring orthographic similarity, endings could cause major problems since they can make two otherwise very similar words appear somewhat different. For example, the Bulgarian word *отправената* ('the directed', a feminine adjective with a definite article) and the Russian word *отправленному* ('the directed', a masculine adjective in dative case) exhibit only about 50% letter overlap, but, if we ignore the endings, the similarity between them becomes much bigger. Thus, if our algorithm could safely ignore word endings when comparing words, it might perform better.

If we could remove the ending, the similarity would be measured using the stem, which is the invariable part of the word. Unfortunately, both the ending as a letter sequence and the location of the morpheme boundary are quite ambiguous in both languages. Thus, we need to lemmatize the text, i.e., convert the word to its main form, the lemma. If every member of the pair of candidate cognates from L1 and L2 is represented by a wordform (WF) and its lemma (L), then we could compare: L1 with L2, WF1 with WF2, L1 with WF2 and WF1 with L2. Considering these four options, we can get a better estimation for the similarity not only between close wordforms like the Bulgarian *отправената* and the Russian *отправленному*, which look different orthographically, but have very close lemmata, but also between such

very different words like the Bulgarian *кънейки* ('bathing', a gerund) and the Russian *копейки* ('copper', plural feminine noun).

The lemmatization of the Bulgarian and the Russian words can be done using specialized dictionaries. In the present work, we will use two large grammatical dictionaries that contain words, their lemmata, and some grammatical information.

### 2.3. Transformation Weights

Let us now come back to the transliteration rules and to the next steps in our algorithm. There are orthographical correspondences between candidate cognates that are not as undisputable as the general rules, but are still observed in the development of the languages, at least for ones with a proven etymological basis. As was shown above, the regular correspondences between the languages can be due to phonetic and spelling reasons. Besides the unconditional letters transitions described above, not so regular ones occur in several cases, and their existence can be taken into account when constructing the weight scale for measuring similarity.

A general principle when building a weight scale is that the correspondences between letters denoting consonants and vowels (hereinafter 'vowels' and 'consonants' only) should be measured separately. The maximal orthographic distance between different letters is 1 (as for  $a-u$ ) and the maximal similarity has weight 0 (as for  $a-a$ ). All weight values between 0 and 1 are assigned to letter correspondences that exist in a non-regular way in some cognates (the above-mentioned correspondence  $y-\text{ь}$  was due to etymological reasons). Another general admission is that consonants and vowels with similar sequences of distinctive phonetic features (differing only in the place of articulation or in the presence/absence of voice, e.g.,  $\bar{b}-\text{в}$ ,  $\bar{b}-n$ ) have lower weight distance. The same is valid for the pair of letters denoting a regular phonetic change, e.g., *reduction* (as in  $a-\text{в}$ ,  $o-y$ ) or *softening* of the preceding consonant (as in  $y-\text{ю}$ ,  $a-\text{я}$ ). Regular correspondences observed in a limited lexical sector (e.g., borrowed from Latin and Greek) such as  $z-x$  also have a lower distance.

Table 3 shows the letter transformation weights, which can be used to measure the orthographic similarity after the Bulgarian and Russian words have been transliterated to a subset of the Cyrillic alphabet.

The weights  $w(a, b)$  are used to transform the letter  $a$  into the letter  $b$  and vice versa. This weight function  $w$  is symmetric by definition, i.e.,  $w(a, b) = w(b, a)$ . All other weights not given in Table 3 are equal to 1.

In order to write the Russian words in the modified Bulgarian alphabet used in Table 3, we make the following preliminary transformations for all Russian words:

$\text{э} \rightarrow e$ ;  $\text{ы} \rightarrow u$ ;  $\text{ь} \rightarrow$  (empty letter);  $\text{ѣ} \rightarrow$  (empty letter)

Table 3 shapes the match between letters and the sounds they denote in Bulgarian and Russian. It further correlates weights for letter transformation that have been phonetically justified.

$a$	$w(a, e)=0.7$ ; $w(a, u)=0.8$ ; $w(a, o)=0.7$ ; $w(a, y)=0.6$ ; $w(a, \text{ь})=0.5$ ; $w(a, \text{ю})=0.8$ ; $w(a, \text{я})=0.5$
$\bar{b}$	$w(\bar{b}, \text{в})=0.8$ ; $w(\bar{b}, n)=0.6$
$\text{в}$	$w(\text{в}, \bar{b})=0.6$
$z$	$w(z, x)=0.5$
$\text{д}$	$w(\text{д}, m)=0.6$
$e$	$w(e, u)=0.6$ ; $w(e, o)=0.7$ ; $w(e, y)=0.8$ ; $w(e, \text{ь})=0.5$ ; $w(e, \text{ю})=0.8$ ; $w(e, \text{я})=0.5$
$\text{ж}$	$w(\text{ж}, \text{з})=0.8$ ; $w(\text{ж}, u)=0.6$
$з$	$w(з, \text{ж})=0.5$
$u$	$w(u, \bar{u})=0.6$ ; $w(u, o)=0.8$ ; $w(u, y)=0.8$ ; $w(u, \text{ь})=0.8$ ; $w(u, \text{ю})=0.7$ ; $w(u, \text{я})=0.7$
$\bar{u}$	$w(\bar{u}, \text{ю})=0.7$ ; $w(\bar{u}, \text{я})=0.7$
$\kappa$	$w(\kappa, m)=0.8$ ; $w(\kappa, x)=0.6$
$m$	$w(m, \kappa)=0.7$
$o$	$w(o, y)=0.6$ ; $w(o, \text{ь})=0.8$ ; $w(o, \text{ю})=0.7$ ; $w(o, \text{я})=0.8$
$n$	$w(n, \bar{b})=0.8$ ; $w(n, x)=0.9$
$c$	$w(c, \text{ч})=0.6$ ; $w(c, u)=0.9$
$m$	$w(m, \bar{b})=0.8$ ; $w(m, x)=0.9$ ; $w(m, \text{ч})=0.9$
$y$	$w(y, \text{ь})=0.5$ ; $w(y, \text{ю})=0.6$ ; $w(y, \text{я})=0.8$
$\bar{\phi}$	$w(\bar{\phi}, \text{ч})=0.8$
$x$	$w(x, z)=0.9$
$\text{ч}$	$w(\text{ч}, \bar{\phi})=0.8$
$\text{ч}$	$w(\text{ч}, u)=0.9$
$\text{ь}$	$w(\text{ь}, \text{ю})=0.8$ ; $w(\text{ь}, \text{я})=0.8$
$\text{ю}$	$w(\text{ю}, \text{я})=0.8$

Table 3– Letter substitution weights.

### 3. The MMEDR Algorithm

The MMEDR algorithm (*modified minimum edit distance ratio*) measures the orthographic similarity between a pair of Bulgarian and Russian words using some general phonetic and morphologically conditioned correspondences between the letters of the two languages in order to estimate the extent to which the two words would be perceived as similar by people fluent in both languages. It returns a value between 0 and 1, where values close to 1 express very high similarity, while 0 is returned for completely dissimilar words. The algorithm has been tailored for Bulgarian and Russian and thus is not directly applicable to other pairs of languages. However, the general approach can be easily adapted to other languages: all that has to be changed are the rules describing the phonetic and the morphological correspondences.

### The MMEDR algorithm in steps:

1. Lemmatize the Bulgarian word.
2. Lemmatize the Russian word.
3. Transform the Russian word's ending.
4. Transliterate the Russian word.
5. Remove some double consonants in the Russian word.
6. Calculate the modified Levenshtein distance using suitable weights for letter substitutions.
7. Normalize and calculate the MMEDR value.

The algorithm first tries to rewrite the Russian word following Bulgarian letter constructions. As a result, both words are transformed into a special intermediate form and then are compared orthographically using Levenshtein distance with suitable weights for individual letter substitutions. The above general algorithm is run in eight variants with each of steps 1, 2 and 3 being included or excluded, and the largest of the eight resulting values is returned. A description of each step follows below.

### 3.1. Lemmatizing Bulgarian and Russian words

The Bulgarian word is lemmatized using a grammatical dictionary of Bulgarian as described in Section 1.3. If the dictionary contains no lemmata for the target word, the original word is returned; if it contains more than one lemma, we try using each of them in turn and we choose the one yielding the highest value in the MMEDR algorithm. The Russian word is lemmatized in the same way, using a grammatical dictionary of Russian.

### 3.2. Transforming the Russian Ending

At this step, we transform the endings of the Russian word according to Tables 1 and 2 and we remove the agglutinative suffix *ся*:

*нный* → *нен*; *ный* → *ен*; *нный* → *нен*; *ний* → *ен*; *ий* → *и*; *ый* → *и*; *ной* → *нен*; *ной* → *-ен*; *ой* → *и*; *ский* - *ски*; *ья* → *ь*; *овать* → *ам*; *ить* → *я*; *ять* → *я*; *ать* → *ам*; *уть* → *а*; *еть* → *ея*

The substitutions rules are applied only if the left hand-side letter sequences are at the end of the word. Rules are applied in the given order; multiple rule applications are allowed. Note that we do not have rules for all possible endings in Russian, but only for the typical ones – object transformation for adjectives and verbs.

Since all words have been already lemmatized in the previous step (if applied), verbs are assumed to be in infinitive and adjectives in singular masculine form. Adjective endings are transformed to their respective Bulgarian counter-parts, and reflexive verbs are turned into non-reflexive. Nouns are not considered since they generally have the same endings in the two languages

(after having been lemmatized) and thus need no additional transformations.

Of course, there are many exceptions for the above rules, but our experiments show that using each of them has more positive than negative effect. Initially, we tried using few more additional rules, which were subsequently removed since they were found to be harmful.

### 3.3. Removing double consonants

According to 1.1.3, the following substitution rules are applied for the Russian word:

*бб* → *б*; *жжж* → *жж*; *кк* → *к*; *лл* → *л*; *мм* → *м*; *нн* → *н*; *пп* → *п*; *сс* → *с*; *мм* → *м*; *фφ* → *φ*

### 3.4. Calculating the Modified Levenshtein Distance with Weights for Letter Substitution

Given two words, the Levenshtein distance [Levenshtein, 1965], also known as the *minimum edit distance* (MED), is defined as the minimum total number of single-letter substitutions, deletions and/or insertions necessary to convert the first word into the second one. We use a modification, which we call *modified minimum edit distance* (MMED), where the weights of all insertions and deletions are fixed to 1, and the weights for single-letter substitution are as given in Table 3.

### 3.5. Calculating MMEDR

At this step, we calculate MMEDR value by normalizing MMED – we divide it by the length of the longer word (the length is calculated after all transformations have been made in the previous steps). We use the following formula:

$$MMEDR(w_{bg}, w_{ru}) = 1 - \frac{MMED(w_{bg}, w_{ru})}{\max(|w_{bg}|, |w_{ru}|)}$$

### 3.6. Calculating the final result

The final result is given by the maximum of the obtained values for all eight variants of the MMEDR algorithm – with/without lemmatization of the Bulgarian word, with/without lemmatization of the Russian word, and with/without transformation of the Russian word ending. Note also, that lemmatization steps might result in calculating additional values for MMEDR – one for each possible lemma of the Russian/Bulgarian word.

### 3.7. Example

As we will see below, the proposed MMEDR algorithm yields significant improvements over classic orthographic similarity measures like LCSR (*longest common subsequence ratio*, defined as the longest common letter subsequence, normalized by the length of the longer word [Melamed, 1999]) and MEDR (*minimum edit distance*

*ratio*, defined as the Levenshtein distance with all weights set to 1, normalized by the length of the longer word, also known as *normalized edit distance /NED/* [Marzal & Vidal, 1993]). This is due to the above-described steps which turn the Russian word into a Bulgarian-sounding one and the application of letter substitution weights that reflect the closeness of the corresponding phonemes.

Let us consider for example the Bulgarian word *афектирахме* and the Russian word *аффектировались*. Using the classic Levenshtein distance, we obtain the following:  $MED(афектирахме, аффектировались) = 7$ . And after normalization:  $MEDR = 1 - (7/15) = 8/15 \approx 53\%$ . In contrast, with the MMEDR algorithm, we first lemmatize the two words, thus obtaining *афектирам* and *аффектировать* respectively. We then replace the double Russian consonant *-фф-* by *-ф-* and the Russian ending *-овать* by the first singular Bulgarian verb ending *-ам*. We thus obtain the intermediate forms *афектирам* and *афектирам*, which are identical, and  $MMEDR = 100\%$ . Note that some pairs of words like *афектирахме* and *аффектировались* could be neither orthographically nor phonetically close but could be perceived as similar due to cross-lingual correspondences that are obvious to people speaking both languages.

Let us take another example – with the Bulgarian word *избягам* and the Russian word *отбежать* (both meaning ‘to run out’), which sound similarly. Using Levenshtein distance:  $MED(избягам, отбежать) = 5$  and thus  $MEDR = 1 - (5/8) = 3/8 = 37.5\%$ . In contrast, with the MMEDR algorithm, we first transform *отбежать* to its intermediate form *отбегам* and we then calculate  $MMED(избягам, отбегам) = 0.8 + 1 + 0.5 = 2.3$  and  $MMEDR = 1 - (2.3/7) = 47/70 \approx 67\%$ , which is a much better reflection of the similarity between the two words.

Thus, we can conclude that, at least in the above two examples, the traditional MEDR does not work well for the highly inflectional Bulgarian and Russian. MEDR is based on the classic Levenshtein distance, which uses the same weight for all letter substitution, and thus cannot distinguish small phonetic changes like replacing *я* with *е* (two phonetically very close vowels) from more significant differences like replacing *я* with *з* (a vowel and a consonant that are quite different).

## 4. Experiments and Evaluation

We performed several experiments in order to assess the accuracy of the proposed MMEDR algorithm for measuring the similarity between Bulgarian and Russian words in a literary text.

### 4.1. Textual resources

We used the Russian novel *The Lord of the World* (*Властелин мира*) by Alexander Belyayev [Belayayev, 1940a] and its Bulgarian translation by Assen Trayanov [Belayayev, 1940b] as our test data. We extracted the first 200 different Bulgarian words and the first 200 different Russian words that occur in the novel, and we measured the similarity between them.

#	Bulgarian word	Russian word	MMEDR	Sim	Precision	Recall
1	беляев	беляев	1.0000	Yes	100.00%	0.68%
2	на	на	1.0000	Yes	100.00%	1.37%
3	глава	глава	1.0000	Yes	100.00%	2.05%
4	кандидат	кандидат	1.0000	Yes	100.00%	2.74%
5	за	за	1.0000	Yes	100.00%	3.42%
6	наполеон	наполеоны	1.0000	Yes	100.00%	4.11%
7	не	не	1.0000	Yes	100.00%	4.79%
8	ми	нас	1.0000	No	87.50%	4.79%
9	ми	мой	1.0000	Yes	88.89%	5.48%
10	ми	мы	1.0000	Yes	90.00%	6.16%
...	...	...	...	...	...	...
93	четвъртият	четвертым	0.9375	Yes	94.57%	59.59%
94	оставят	остается	0.9286	Yes	94.62%	60.27%
...	...	...	...	...	...	...
39998	са	в	0.0000	No	0.37%	100%
39999	са	к	0.0000	No	0.37%	100%
40000	боядисвали	к	0.0000	No	0.37%	100%

Table 4 – Results of the MMEDR algorithm.

### 4.2. Grammatical Resources

We used two monolingual dictionaries for lemmatization:

- **A grammatical dictionary of Bulgarian**, created at the Linguistic Modeling Department, Institute for Parallel Processing, Bulgarian Academy of Sciences [Paskaleva, 2002]. This electronic dictionary contained 963,339 wordforms and 73,113 lemmata. Each dictionary entry consisted of a wordform, a corresponding lemma, and some morphological and grammatical information.
- **A grammatical dictionary of Russian**, created at the Institute of Russian language, Russian Academy of Sciences, based on the Grammatical Dictionary of A. Zaliznyak [Zaliznyak, 1977]. The dictionary consisted of 1,390,613 wordforms and 66,101 lemmata. Each dictionary entry consisted of a wordform, a corresponding lemma, and some morphological and grammatical information.

### 4.3. Experimental Setup

We measured the similarity between all  $200 \times 200 = 40,000$  Bulgarian-Russian pairs of words. Among them, 163 pairs were annotated as very similar by a linguist who was fluent in Russian and a native speaker of Bulgarian; the remaining 39,837 were considered unrelated.

We used the MMEDR algorithm to rank the 40,000 pairs of words in decreasing order according to the

calculated similarity values. Ideally, the 163 pairs designated by the linguist would be ranked at the top. We can determine how well the ranking produced by our algorithm does using standard measures from information retrieval, *e.g.* 11-point interpolated average precision [Manning et al., 2008].

We compared the MMEDR algorithm with two classic orthographic similarity measures: LCSR and MEDR. Unfortunately, we could not directly compare our results to those in other work, since there were no previous publications measuring orthographic or phonetic similarity between words in Bulgarian and Russian.

#### 4.4. Results

Table 4 shows part of the ranking produced by the MMEDR algorithm. The table shows an excerpt of the ranked pairs of words along with their similarity calculated by the MMEDR algorithm, the corresponding human annotation for similarity (the column "Sim"), as well as precision and recall calculated for all rows from the beginning to the current row.

Table 5 shows the 11-pt interpolated average precision for LCSR, MEDR and MMEDR. We can see that MMEDR outperforms the other two similarity measures by a large margin: 18-22% absolute difference.

Algorithm	11-pt interpolated average precision
LCSR	69.06%
MEDR	72.30%
MMEDR	<b>90.58%</b>

Table 5 – Comparison of the similarity measuring algorithms.

### 5. Discussion

As Tables 4 and 5 show, the MMEDR algorithm works quite well. Still, there is a lot of room for improvement:

- Bulgarian and Russian inflectional morphologies are quite complex, with many exceptions that are not captured by our rules. This is probably a limitation of the general approach rather than a deficiency of the particular rules used: if we are to capture all exceptions, we would need to manually specify them all, which would require a lot of extra manual work.
- The transformation rules between Bulgarian and Russian are sometimes imprecise as well, *e.g.*, for very short words or for words of foreign origin.
- While linguistically motivated, the letter-for-letter substitution weights we used are *ad hoc*, and could be improved. First, while we used symmetric letter substitution weight in Table 3, asymmetric weights might work better, *e.g.* the Bulgarian prefixes *pa3-* and *u3-* are spelled as *pac-* and *uc-* in Russian when followed by a voiceless consonant. Thus, the

substitution weight for  $з \rightarrow c$  should probably be higher than for  $c \rightarrow з$ . We could further extend the rules to take into account the local context, *e.g.*, changing *pa3-* to *pac-* could have a different weight than changing *-з* to *-c* in general.

- Another potential problem comes from us using only one linguist for the annotation, which might have yielded biased judgments. To assess the impact of the potential subjectivity, we would need judgments by at least one additional linguist.

### 6. Related Work

Many algorithms have been proposed in the literature for measuring the orthographic and the phonetic similarity between pairs of words from different languages.

The simplest ones considered as orthographically close words with identical prefixes [Simard & al., 1992].

Much more popular have been orthographic similarity measures based on normalized versions of the Levenshtein distance [Levenshtein, 1965], the longest common subsequence [Melamed, 1999], and the Dice coefficient [Brew and McKelvie, 1996].

Somewhat less common have been phonetic similarity measures, which compare sounds instead of letter sequences. Such an approach has been proposed for the first time by [Russel, 1918]. Guy [1994] described an algorithm for cognate identification in bilingual word lists based on statistics of common sound correspondences. Algorithms that learn the typical sound correspondences between two languages automatically have also been proposed: [Kondrak, 2000], [Kondrak, 2003] and [Kondrak & Dorr, 2004].

Instead of applying similarity measures for symbolic strings on the words directly, some researchers have first performed transformations that reflect the typical cross-lingual orthographic and phonetic correspondences between the target languages. This is especially important for language pairs where some letters in the source language are systematically substituted by other letters in the target language. The idea can be extended further with substitutions of whole syllables, prefixes and suffixes. For example, Koehn & Knight [2002] proposed manually constructed transformation rules from German to English (*e.g.*, the letters *k* and *z* are changed to *c*; and the ending *-tät* is changed to *-ty*) in order to expand lists of automatically extracted cognates.

Finally, orthographic measures like LCSR and MEDR have gradually evolved over the years, enriched by machine learning techniques that automatically identify templates for cross-lingual orthographic and phonetic correspondences. For example, Tiedemann [1999] learned spelling transformations from English to Swedish, while Mulloni & Pekar [2006] and Mitkov & al. [2007] learned transformation templates, which represent substitutions of letters sequences in one language with letter sequences in another language.

## 7. Conclusions and Future Work

We have described and tested a novel algorithm for measuring the similarity between pairs of words based on transformation rules between Bulgarian and Russian. The algorithm has shown very high precision and could be used to identify possible candidates for cognates or false friends in text corpora. It can also be used in machine translation systems working on related languages where it could help overcome the incompleteness of translation dictionaries used in the system.

There are many ways in which we could improve the proposed algorithm. For example, we could adapt the algorithms described in [Mitkov et al., 2007] and [Bergsma & Kondrak, 2007] to Bulgarian and Russian and try to learn cross-lingual transformation rules for morphemes and other sub-word sequences automatically. We could then try to combine MMEDR with such rules.

## Acknowledgments

The presented research is supported by the project FP7-REGPOT-2007-1 SISTER.

## 8. References

- [Belyayev, 1940a] Belyaev A. "Lord of the World" (1940), in Russian, publisher "Onyx 21 Century", 2005, ISBN 5-329-01356-9, <http://lib.ru/RUFANT/BELAEW/lordwrld.txt>
- [Belyayev, 1940b] Belyaev A. "Lord of the World" (1940), translation from Russian to Bulgarian by A. Trayanov, publisher "National Youth", 1977, <http://www.chitanka.info/lib/text/2130>
- [Bergsma & Kondrak, 2007] Bergsma S., Kondrak G. "Alignment-Based Discriminative String Similarity". *Proceedings of the 45th Annual Meeting of the ACL*, pages 656–663, Prague, Czech Republic, 2007
- [Brew and McKelvie, 1996] Brew C. and McKelvie D. "Word-Pair Extraction for Lexicography". *Proceedings of the 2nd International Conference on New Methods in Language Processing*, pages 45–55, Ankara, Turkey, 1996
- [Guy, 1994] Guy J. "An Algorithm for Identifying Cognates in Bilingual Wordlists and Its Applicability to Machine Translation", *Journal of Quantitative Linguistics*, Volume 1 (1), pages 35-42, 1994
- [Koehn & Knight, 2002] Koehn P., Knight K. "Learning a Translation Lexicon from Monolingual Corpora". In *Proceedings of the Workshop of the ACL Special Interest Group on the Lexicon (SIGLEX)*, pages 9-16, Philadelphia, PA, 2002.
- [Kondrak and Dorr, 2004] Kondrak G., Dorr B. "Identification of Confusable Drug Names: A New Approach and Evaluation Methodology". *Proceedings of COLING 2004*, pages 952–958, Geneva, Switzerland, 2004
- [Kondrak, 2000] Kondrak G. "A New Algorithm for the Alignment of Phonetic Sequences". *Proceedings of NAACL/ANLP 2000: 1st conference of the North American Chapter of the Association for Computational Linguistics and 6th Conference on Applied Natural Language Processing*, pages 288-295, Seattle, WA, USA, 2000
- [Kondrak, 2003] Kondrak G. "Identifying Complex Sound Correspondences in Bilingual Wordlists". *Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Text Processing (CICLING 2003)*, pages 432-443, Mexico City, Mexico, 2003
- [Levenshtein, 1965] Levenshtein V. "Binary Codes Capable of Correcting Deletions, Insertions and Reversals". *Doklady Akademii Nauk SSSR*, Volume 163 (4), pages 845-848, Moscow, Russia, 1965
- [Manning et al., 2008] Manning C., Prabhakar R. and Schütze H. "Introduction to Information Retrieval". *Cambridge University Press*, ISBN 0521865719, New York, USA, 2008
- [Marzal & Vidal, 1993] Marzal A., Vidal E. "Computation of Normalized Edit Distance and Applications". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Volume 15, Issue 9, pages 926-932, USA, 1993
- [Melamed, 1999] Melamed D. "Bitext Maps and Alignment via Pattern Recognition". *Computational Linguistics*, Volume 25 (1), pages 107-130, ISSN:0891-2017, 1999
- [Mitkov et al., 2007] Mitkov R., Pekar V., Blagoev D. and Mulloni A. "Methods for Extracting and Classifying Pairs of Cognates and False Friends". *Machine Translation*, Volume 21, Issue 1, pages 29-53, Springer Netherlands, 2007
- [Mulloni & Pekar, 2006] Mulloni A. and Pekar V. "Automatic Detection of Orthographic Cues for Cognate Recognition". *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-06)*, pages 2387–2390, Genoa, Italy, 2006.
- [Paskaleva, 2002] Paskaleva E. "Processing Bulgarian and Russian Resources in Unified Format". *Proceedings of the 8th International Scientific Symposium MAPRIAL*, pages 185-194, Veliko Tarnovo, Bulgaria, 2002.
- [Russel, 1918] Russel R. "U.S. Patent 1,261,167", Pittsburgh, PA, USA, 1918
- [Simard et al., 1992] Simard M., Foster G., Isabelle P. "Using Cognates to Align Sentences in Bilingual Corpora". *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 67–81, Montreal, Canada, 1992
- [Tiedemann, 1999] Tiedemann J. "Automatic Construction of Weighted String Similarity Measures". *SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 213-219, College Park, MD, USA, 1999
- [Zaliznyak, 1977] Zaliznyak A. "Grammatical Dictionary of the Russian Language", publisher "Russian Language", Moscow, Russia, 1977