PROCEEDINGS

International Workshop

17 September 2009
Borovets, Bulgaria

Multilingual resources, technologies and evaluation for Central and Eastern European languages

EDITED BY

Cristina Vertan
Stelios Piperidis
Elena Paskaleva
Milena Slavcheva

TABLE OF CONTENTS

[Nakov & колектив, 2009b] Nakov S., Paskaleva E., Nakov P. "A Knowledge-Rich Approach to Measuring the Similarity between Bulgarian and Russian Words", *Workshop on Multilingual Resources, Technologies and Evaluation for Central and Eastern European Languages held in conjuction with RANLP 2009*, Borovets, Bulgaria, 2009

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 Str., 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 words are perceived as similar by a person who is fluent in both languages. We assume that words with different orthography and phonetic composition can be perceived as similar when they have the same or a similar stem and inflections, as in the Bulgarian word *афектирахме* and the Russian *аффектировались* (both meaning '*we were affected*').

Bulgarian and Russian are highly related Slavonic languages with rich morphology and this motivates us to study the typical orthographical, phonetic and morphological correspondences between these languages and use them to formulate and apply transformation rules for bringing a Russian word to Bulgarian sounding and vice versa. Our algorithm for measuring the similarity between Bulgarian and Russian words first reduces the Russian word to an intermediate form with Bulgarian sounding, performs some transformations over the Bulgarian word to obtain corresponding intermediate form and finally compares orthographically the obtained intermediate forms. The algorithm starts by transcribing the Russian words with the Bulgarian alphabet, and then transforms some typical Russian morphemes and word parts (*e.g.*, prefixes, suffixes, endings, *etc.*) to their corresponding Bulgarian ones. As second step it transforms some Bulgarian letters and word parts to make the Bulgarian word sounding more like the intermediate form of its Russian correspondence. Since both Bulgarian and Russian are highly-inflectional languages, lemmatization is used to deal with some particular endings. Finally, the orthographic similarity is measured using a modified Levenshtein distance with letter-specific substitution weights.

The equalization of the Bulgarian and Russian words into their corresponding intermediate forms has phonetic and morphological motivation and is performed as sequence of steps described in details below.

1.1. Cyrillic Alphabet and Transcription

In a strict linguistic sense, *transcription* is the process of matching the sounds of human speech to special written symbols, using a set of exact rules, so that these sounds can be reproduced later. Both Russian and Bulgarian use the Cyrillic alphabet in their transcription, but some letters have different phonetic function in the two languages; moreover, Russian uses three letters that do not exist in the Bulgarian alphabet. Still, there are generally accepted transliterations between the letters used in the Russian and in the Bulgarian alphabet, and some rules for phonetic transcription could be easily derived. Table 1 presents the typical correspondences between Bulgarian and Russian letters:

Russian Letter	Bulgarian Letter	Examples Russian - Bulgarian
а	*а*	***а**збука – **а**збука*
б	*б*	***б**уква – **б**уква*
в	*в*	***в**оля – **в**оля*
г	*г*	***г**ипс – **г**ипс*
д	*д*	***д**иректор – **д**иректор*
е	е (sometimes *я*) (sometimes *ъ*) (sometimes *ьо*)	***е**жегодно – **е**жегодно* *хл**е**б – хл**я**б* *с**е**рп – с**ъ**рп* *акт**е**р – акт**ьо**р*
ё (letter used in textbooks only)	е	*м**ё**д – м**е**д*
э	е	***э**таж – **е**таж*
ж	*ж*	***ж**ена – **ж**ена*
з	*з*	***з**акон – **з**акон*
и	*и*	***и**стина – **и**стина*
й	*й*	***й**од – **й**од*
к	*к*	***к**ипарис – **к**ипарис*
л	*л*	***л**ак – **л**ак*
м	*м*	***м**онета – **м**онета*
н	*н*	***н**ож – **н**ож*

о	*о* (sometimes *ъ*)	*опера – опера* *сон – сън*
п	*п*	*палитра – палитра*
р	*р*	*река – река*
с	*с*	*сом – сом*
т	*т*	*танк – танк*
у	*у* (sometimes *ъ*)	*ум – ум* *дуб – дъб*
ф	*ф*	*факт – факт*
х	*х*	*химия – химия*
ц	*ц*	*цвят – цвет*
ч	*ч*	*черен – черный*
ш	*ш*	*шум – шум*
щ	*щ*	*щит – щит*
(missing letter)	*ъ*	*дно – дъно*
ъ	(missing letter)	*изъявление – изявление*
ь	(missing letter)	*день – ден*
ы	*и*	*рыба – риба*
ю	*ю*	*юноша – юноша*
я	*я*	*яхта – яхта*

Table 1 – Letter correspondences in Russian and Bulgarian.

Given a Bulgarian and a Russian word, it is sufficient to replace the Russian letters in the Russian word (or pairs of letters) with their Bulgarian counterparts from the above table in order to obtain a unified transcription, written with the Bulgarian alphabet. This transcription will be then used for measuring orthographic similarity.

Since our goal is to draw together the orthographical forms of the Bulgarian and Russian words in addition to these rules, we can also add some additional letter substitutions that transcribe the Bulgarian and Russian words into a unified intermediate form. We will return to this question in Section 2.3 below.

1.2. Double Consonants

Unlike Bulgarian, double consonants are abundant in Russian. Thus, we define transformation rules for double consonants in Russian in Table 2 in order to obtain intermediate form of the Russian words that is closer to the Bulgarian sounding.

Russian Form	Intermediate Form	Examples
-бб-	*-б-*	*суббота → субота*
-жж-	*-ж-*	*жужжать → жужать*
-кк-	*-к-*	*аккордеон → акордеон*
-лл-	*-л-*	*доллар → долар*
-мм-	*-м-*	*сумма → сума*
-пп-	*-п-*	*аппарат → апарат*
-рр-	*-р-*	*перрон – перон*
-сс-	*-с-*	*депрессия → депресия*
-тт-	*-т-*	*оперетта – оперета*
-фф-	*-ф-*	*эффект → эфект*

Table 2 – Transformation for double consonants in Russian.

Such transformations for double consonants are not needed for the Bulgarian words since in Bulgarian double consonants are much rarer and are usually preserved in the Russian corresponding words, *e.g.* *едновременно* → *одновременно* (‘*simultaneously*’) and *наддавам* → *наддавать* (‘*to outbid*’). We do not transform the Russian *-нн-* into *-н-* because *-нн-* is preserved between the Bulgarian and Russian words correspondences.

The only transformation rule for Bulgarian that we could apply is the one given in Table 3.

Bulgarian Form	Intermediate Form	Example
-зс-	*-с-*	*разстояние → растояние*

Table 3 – Transformation of consonants in Bulgarian.

The Bulgarian *-зс-* corresponds to the Russian *-сс-*, but the Russian *-сс-* is transformed into *-с-* (according to Table 2). In order to equalize the two forms, we have to transform both Bulgarian *-зс-* and Russian *-сс-* into *-с-*.

1.3. Lemmatization

Bulgarian and Russian are highly-inflectional languages, *i.e.*, they use variety of endings to express the different forms of the same word. For example, nouns, adjectives and pronouns decline in several cases in Russian and receive different endings, and, in Bulgarian, the nouns and the adjectives can be determined, which also gives them a specific ending (Bulgarian definite article is a suffix; there are no articles in Russian). The different inflected verb forms in Bulgarian and Russian are also formed by adding various endings.

When measuring orthographic similarity, endings could cause major problems since they can make two otherwise very similar words look somewhat different. For example, the Bulgarian word *отправен**ата*** (‘*the directed*, a feminine adjective with a definite article) and the Russian word *отправленн**ому*** (‘*the directed*, a masculine adjective in dative case) exhibit only about 50% letter overlap, but, if we ignore the endings, the similarity between them would be much greater. If our algorithm could safely ignore word endings when comparing words, it would perform better.

Lemmatization (or converting a word to its main form) is a common way to deal with the different inflected forms of the same word. Given the task to compare to what extent a Bulgarian and a Russian word are perceived as similar, we could first transcribe the Russian word with the Bulgarian alphabet. Then we can compare (a) the two words directly, or (b) their corresponding lemmata. We could also compare (c) the Bulgarian word to the lemma of the Russian word, or (d) the lemma of the Bulgarian word to the Russian word. Considering these four options, we can get a better estimate 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 *къпейки* ('*bathing*', a gerund) and the Russian *копейки* ('*copeck*', plural feminine noun).

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

1.4. Transformation of Russian Endings

The problem with the different endings is not entirely solved even after lemmatization. Indeed, the lemmata of adjectives and verbs have different endings in Bulgarian and Russian. For example, the Bulgarian lemma *декорирам* ('*I decorate*', a first person singular form, which is considered to be the lemma of a verb since verbs have no infinitive forms in Bulgarian) is too different from its corresponding Russian lemma *декорировать* ('*to decorate*', an infinitive). In order to overcome this problem, we decided to reduce the Russian endings to Bulgarian-sounding ones. Let us consider how such an alignment can be done. Table 4 and Table 5 show the typical transformation of Russian adjectival and verbal forms to the corresponding Bulgarian forms:

Russian Ending	Bulgarian Ending	Examples
-нный	*-нен*	*вое**нный** → вое**нен***
-ный	*-ен*	*веч**ный** – веч**ен***
-нний	*-нен*	*ра**нний** → ра**нен***
-ний	*-ен*	*вечер**ний** → вечер**ен***
-ий	*-и*	*вражеск**ий** → вражеск**и***
-ый	*-и*	*стрелков**ый** – стрелков**и***
-нной	*-нен*	*сте**нной** – сте**нен***
-ной	*-ен*	*род**ной** – род**ен***
-ой	*-и*	*делов**ой** – делов**и***

Table 4 – Transformation of Russian endings to Bulgarian ones (adjectives).

Russian Ending	Bulgarian Ending	Examples
-овать	*-ам*	*декориро**вать** → декорир**ам***
-ить, -ять	*-я*	*брод**ить** → брод**я*** *бле**ять** → бле**я***
-ать	*-ам*	*дав**ать** → дав**ам***
-уть	*-а*	*гас**нуть** – гас**на***
-еть	*-ея*	*бел**еть** → бел**ея***

Table 5 – Transformation of Russian endings to Bulgarian ones (verbs).

Before applying the transformations from Table 5, we need to introduce another rule – for the transformation of reflexive verbs in Russian (Table 6).

Russian Ending	Transformed Russian Ending	Examples
-ься	*-ь*	*весели**ться** → весели**ть***

Table 6 – Transformation of Russian reflexive verbs.

The reflexivity in both languages is expressed on different grammatical levels – we have a reflexive morpheme "*ся*" ("*сь*") in Russian and a reflexive lexeme-particle "*се*" in Bulgarian (e.g. Russian *веселиться* – Bulgarian *веселя се* '*I am having fun*'). Although the reflexive particle and the infinitive form of a verb differ semantically, we deliberately decided to equalize them. This is done in order to increase the orthographic similarity between a Russian verb and its Bulgarian counterpart (which excludes the reflexive particle "*се*").

Other typical difference is observed for the Bulgarian definite article, the morpheme *-та* (e.g. *жена**та*** '*the woman*') missing in the Russian grammatical system. We intentionally do not derive transformation rule from this correspondence because there are too much exceptions where the Bulgarian inflection *-та* is preserved in Russian, e.g. *анкета* and *анкета* ('*poll*').

It is important to apply the rules from Table 4, Table 6 and Table 5 in the proposed order since sometimes a more than one rule will be applicable for some words. For example, the Russian word *веселиться* will be first transformed to *веселить* and then to *веселя*, which will make it identical with the Bulgarian form of the same verb (ignoring the reflexive particle *се*).

Note that we perform transformation of Russian endings but we do not change the Bulgarian endings. This is because we want to turn all Russian words into an intermediate form which is closer to their Bulgarian correspondence. The Bulgarian endings are preserved because the intermediate form by design has Bulgarian sounding which is true for all Bulgarian words.

Of course, there are some exceptions, and the proposed transformation rules for Russian word endings cannot generate the correct Bulgarian wordform, *e.g.*, *вис**еть*** would become *вис**ея***, while the correct Bulgarian form is *вис**я***. In order to reduce the negative impact of that, we measure the similarity (1) with and (2) without applying these rules; we then return the higher value of the two.

1.5. Transformation Weights

Let us now return back to transcription. After a Russian word has been transcribed into Bulgarian alphabet, the different letters correspond more or less to different phonemes. Of course, some phonemes are very close, *e.g.*, the ones encoded by the vowels *о* and *у* and the consonants *д* and *т*, while others like *ц* and *а* are very different. This should be taken into account since we want to measure primarily whether two words sound similarly, and we do not care that much about whether they have similar spellings. The easiest way to achieve this is by assigning appropriate weights to letter substitutions so that similar phonemes have a lower weight than dissimilar ones.

Table 7 shows the letter transformation weights, which can be used to measure the orthographic similarity after the Bulgarian and Russian words have been transcribed to a subset of the Cyrillic alphabet.

а	*w*(*а*, *е*)=0.7; *w*(*а*, *и*)=0.8; *w*(*а*, *о*)=0.7; *w*(*а*, *у*)=0.6; *w*(*а*, *ъ*)=0.5; *w*(*а*, *ё*)=0.8; *w*(*а*, *ю*)=0.8; *w*(*а*, *я*)=0.5
б	*w*(*б*, *в*)=0.8; *w*(*б*, *п*)=0.6
в	*w*(*в*, *ф*)=0.6
г	*w*(*г*, *х*)=0.5
д	*w*(*д*, *т*)=0.6
е	*w*(*е*, *и*)=0.6; *w*(*е*, *о*)=0.7; *w*(*е*, *у*)=0.8; *w*(*е*, *ъ*)=0.5; *w*(*е*, *ё*)=0.3; *w*(*е*, *ю*)=0.8; *w*(*е*, *я*)=0.5
ж	*w*(*ж*, *з*)=0.8; *w*(*ж*, *ш*)=0.6
з	*w*(*з*, *с*)=0.5
и	*w*(*и*, *й*)=0.6; *w*(*и*, *о*)=0.8; *w*(*и*, *у*)=0.8; *w*(*и*, *ъ*)=0.8; *w*(*и*, *ё*)=0.7; *w*(*и*, *ю*)=0.7; *w*(*и*, *я*)=0.7
й	*w*(*й*, *ё*)=0.7; *w*(*й*, *ю*)=0.7; *w*(*й*, *я*)=0.7
к	*w*(*к*, *т*)=0.8; *w*(*к*, *х*)=0.6
л	*w*(*л*, *у*)=0.6
м	*w*(*м*, *н*)=0.7
о	*w*(*о*, *у*)=0.6; *w*(*о*, *ъ*)=0.8; *w*(*о*, *ё*)=0.6; *w*(*о*, *ю*)=0.7; *w*(*о*, *я*)=0.8
п	*w*(*п*, *ф*)=0.8; *w*(*п*, *х*)=0.9
с	*w*(*с*, *ц*)=0.6; *w*(*с*, *ш*)=0.9
т	*w*(*т*, *ф*)=0.8; *w*(*т*, *х*)=0.9; *w*(*т*, *ц*)=0.9
у	*w*(*у*, *ъ*)=0.5; *w*(*у*, *ё*)=0.8; *w*(*у*, *ю*)=0.6; *w*(*у*, *я*)=0.8
ф	*w*(*ф*, *ц*)=0.8
х	*w*(*х*, *ш*)=0.9
ц	*w*(*ц*, *ч*)=0.8
ч	*w*(*ч*, *ш*)=0.9
ъ	*w*(*ъ*, *ё*)=0.8; *w*(*ъ*, *ю*)=0.8; *w*(*ъ*, *я*)=0.8
ё	*w*(*ё*, *ю*)=0.6; *w*(*ё*, *я*)=0.7
ю	*w*(*ю*, *я*)=0.8

Table 7 – Letter substitution weights.

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 7 are equal to 1.

Note that in Table 7, we use the Bulgarian alphabet without the letters *щ* and *ь* and with the additional letter *ё*. This is because, as part of the phonetic transcription, we have preliminary made the following transformations in all Bulgarian words:

щ → *шт; ьо* → *ё; йо* → *ё*

In order to write the Russian words in the modified Bulgarian alphabet used in Table 7, we make the following preliminary transformations in all Russian words:

э → *е; щ* → *шт; ьо* → *ё; йо* → *ё; ы* → *и; ь* → *(missing letter); ъ* → *(missing letter)*

Since the letter *е* is used in Russian for two different sounds – *е* and *ё,* and since there is no clear criterion to tell which one is used in a particular case, we assign a low enough weight to the transformation between the letters *е* and *ё*, which will moderate the differences between the two possible sounds.

Table 7 shapes the match between letters and sounds in Bulgarian and Russian. It correlates phonetically justified weights for sound transformation; it also helps us account for phonetic characteristics when we measure orthographic similarity.

2. 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 morphological correspondences between 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. Transcribe the Bulgarian word.
5. Transcribe the Russian word.
6. Remove some double consonants in the Bulgarian and Russian words.
7. Calculate the modified Levenshtein distance using suitable weights for letter substitutions.
8. Normalize and calculate the MMEDR value.

The algorithm first tries to make the Russian word sound like a Bulgarian one and modifies the Bulgarian word to make it closer to its Russian correspondence. 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 detailed description of each step follows below.

2.1. Lemmatizing the 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.

2.2. Transforming the Russian Ending

At this step, we transform the endings of the Russian word according to Table 4, Table 6 and Table 5:

> *нный* → *нен*; *ный* → *ен*; *нний* → *нен*; *ний* → *ен*; *ий* → *и*; *ый* → *и*; *нной* → *нен; ной* → *-ен*; *ой* → *и*; *ься* → *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 for adjectives and verbs.

Since all words are 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 counterparts, reflexive verbs are turned into non-reflexive, and infinitives are transformed into the main form in Bulgarian, which is first person singular. Nouns are not considered since they generally have the same endings in the two languages (after having been lemmatized) and thus need no transformations.

Of course, there are many exceptions for the above rules, but our experiments show that using each of these rules 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.

2.3. Transcribing the Bulgarian and Russian Words

The Bulgarian word is transcribed using the following substitutions:

> *щ* → *шт*; *ьо* → *ё*; *йо* → *ё*

As a result, in the intermediate form each sound tends to correspond to one and only one letter, and thus, orthographic similarity implicitly approximates phonetic similarity.

The transcription of the Russian word is performed using the following substitutions:

> *э* → *е*; *щ* → *шт*; *ьо* → *ё*; *йо* → *ё*; *ы* → *и*; *ь* → *(empty letter)*; *ъ* → *(empty letter)*

As a result, we obtain a transcribed intermediate form of the Russian word, where each Russian sound is transformed into a Bulgarian one and tends to correspond to one and only one letter. Thus, measuring orthographic similarity reflects (to some extent) phonetic similarity as well.

2.4. Removing Some Double Consonants

According to Table 2, the following substitution rules are applied for the Russian word:

> *бб* → *б*; *жж* → *ж*; *кк* → *к*; *лл* → *л*; *мм* → *м*; *пп* → *п*; *рр* → *р*; *сс* → *с*; *тт* → *т*; *фф* → *ф*

According to Table 3, the following substitution rule is applied for the Bulgarian word:

> *зс* → *с*

2.5. 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 7.

2.6. 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:

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

2.7. 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.

2.8. 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 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 ≈ 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 *отбегать* (both meaning '*to run out*'), which sound similarly. Using Levenshtein distance, we obtain MED(*избягам*,*отбегать*) = 5 and thus MEDR = 1 – (5/8) = 3/8 = 37.5%. In contrast, with the MMEDR algorithm, we first transform *отбегать* 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 ≈ 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 *г* (a vowel and a consonant that are quite different).

3. 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.

3.1. Test 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 measured the similarity between them.

3.2. 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 contains 963,339 wordforms and 73,113 lemmata. Each dictionary entry consists of a wordform, a corresponding lemma, followed by 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 consists of 1,390,613 wordforms and 66,101 lemmata. Each dictionary entry consists of a wordform, a corresponding lemma, followed by some morphological and grammatical information.

3.3. Experimental Setup

We measured the similarity between all 200x200=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 cannot directly compare our results to those in other work, since there are no previous publications measuring orthographic or phonetic similarity between words in Bulgarian and Russian.

3.4. Results

Table 8 below shows part of the ranking produced by the MMEDR algorithm:

#	Bulga-rian	Rus-sian	MMEDR	Sim	Precision	Recall

	word	word				
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	канди-дат	кан-дидат	1.0000	Yes	100.00%	2.74%
5	за	за	1.0000	Yes	100.00%	3.42%
6	напо-леон	напо-леоны	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	четвър-тият	чет-вертым	0.9375	Yes	94.57%	59.59%
94	оставят	оста-ется	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	боядис-вали	к	0.0000	No	0.37%	100%

Table 8 – Results of 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 9 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	**90.58%**

Table 9 – Comparison of the similarity measuring algorithms.

4. Discussion

As Table 8 and Table 9 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 additional 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 7, asymmetric weights might work better, *e.g.* the Bulgarian prefixes *раз-* and *из-* are spelled as *рас-* and *ис-* in Russian when followed by a voiceless consonant. Thus, the substi-ution weight for *з* → *с* should probably be higher than for *с* → *з*. We could further extend the rules to take into account the local context, *e.g.*, changing *раз-* to *рас-* could have a different weight than changing *-з-* то *-с-* 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.

5. 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 approaches 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.

6. Conclusions and Future Work

We have described and tested a novel algorithm for measuring the similarity between pairs of words based on manual transformation rules between Bulgarian and Russian. The algorithm shows 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 to 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.

7. 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 (visited in April 2009)

[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 (visited in April 2009)

[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, USA, 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". *Proceedings of the 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