The translation algorithm from pre-reform spelling into modern spelling, taking into account the morphology of words

O Yu Kozhemyakina and E P Tagirova

Institute of Computational Technologies of SB RAS, Lavrentiev av. 6, Novosibirsk, 630090, Russian Federation

E-mail: olgakozhemyakina@mail.ru, e.tagirova@g.nsu.ru

Abstract. The paper presents the translation algorithm from pre-reform spelling into modern one. The regularities in the errors made by pymorphy analyzer were used in the adaptation of it to the work with pre-reform spelling. On the basis of the dependencies between pre-reform and modern spelling for words, united by morphological characteristics, the exact algorithm, for processing the documents related to the late 19th – early 20th century, is proposed.

1. Introduction
One of the most important problems for the library community is the digitization of printed publications stored in the collections of libraries. The pre-revolutionary publications, which are printed in accordance with the spelling accepted before 1918, are also should be digitized. However, the algorithms for digitized texts processing are based on the dictionaries in modern spelling. Thus, the development and implementation of software applications that translate texts from pre-reform spelling into modern ones becomes relevant.

The spelling is a set of rules that ensure the uniformity of words and their grammatical forms in the writing. The pre-revolutionary spelling (otherwise called old or pre-reform) is the spelling of the Russian language, which was used before its reform in 1918. There was no generally accepted unique norm of the old spelling, only the spelling of the last 50 years before the reform was more or less normalized [1] – these rules are used in this work.

Among the modern studies on the pre-reform spelling, we emphasize the work of P.I. Davydov [2], because the author has studied and analyzed a significant number of sources and literature, including textbooks, dictionaries and reference books, that were published in the old spelling until 1917.

Referring to the decree “About the implementation of the new spellin” from October 10, 1918, the author describes a number of changes that have occurred with spelling [2, p. 11–13], from which we can distinguish the main groups:

1. The spelling of morphemes:
   1.1. Prefixes;
   1.2. Case endings;

2. The usage of particular letters.

However, the spelling is a set of clear rules on the basis of which the translation algorithm can be developed.

At the moment, there are many online services for the text translation from the old spelling into the modern one, for example [3]. Their usage in the framework of a full-fledged software package

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd
for automatic text processing, is quite difficult. In addition, the algorithms used by such services are not described anywhere, what makes almost impossible the improvement and optimization of the automatic translation process.

So, the aim of this work is to develop and describe an algorithm for automatic text translation from Russian pre-reform spelling to the modern one, taking into account the morphology of words.

2. General concept of translation algorithm

The two main steps of the word translation can be defined:
1) the search and editing of the spelling of the morpheme;
2) the search and editing of the particular letters.

The second step involves a simple replacement of the obsolete letter to its modern equivalent. However, the spelling of morphemes, which are considered on the first step, is not so trivial: it is not always possible to get the correct translation by simple replacing of the letters (as on the second step). For example, consider the word синія. The obsolete letter і corresponds to the modern и. In the case of a simple replacement, we get – синия. However, the correct translation for the original word is the adjective синие. It is necessary to select one or several morphological features of the word for its correct editing. The algorithm of morphological analysis of words can be used to solve this problem. More detailed description is provided below. Also, for the implementation of the first step of translation it is necessary to identify the patterns between the old and modern spelling rules, on the basis of which the formal rules of translation will be described. These rules will be presented further in the article.

Some modern equivalents of obsolete morphemes have a different sound, what can lead to the loss of rhyme after the implementation of the first step of translation. For example, the lines from the poem “Silentium!” by F.I. Tyutchev — “Пускай в душевной глубине // И всходят и зайдут они”, will be translated as “Пускай в душевной глубине // И всходят и зайдут они”. This situation is unacceptable when we are working with poetic texts. In the simplest case, to prevent rhyme loss, it is sufficient to exclude the editing of the last word ending of the line at the first step. Then at the second step there will be a simple replacement of the letter with its modern equivalent, which does not change the sound of the word when we read it, and therefore it does not break the rhyme. Thus, the lines from the above example will be translated in another way: “Пускай в душевной глубине // И всходят и зайдут они”, so the letter и will be replaced with its modern equivalent – е.

Since in modern spelling the letter ё is often abolished, it is possible to provide two modes of the algorithm: with the usage of ё and without it. To implement the variant without ё, it is enough to replace ё with e at the second step of the translation. As for the first variant, the second step remains unchanged. However, it should be considered that there are some words that were used with e before the reform, and after it began to be written with ё. For example, the pronoun она in the accusative case in the pre-reform spelling was written as ee, and in the modern — её. Since the number of such words is significantly limited, they can be represented in the form of a set of “key-value” pairs, where the key is the obsolete form of the word, and the value is its modern equivalent with ё or e depending on the chosen mode of the algorithm.

3. The testing and comparative analysis of morphological analyzers during the work with pre-reform spelling

As we already mentioned, for the implementation of a correct translation it is necessary to take into account the morphology of the translated words. The morphological analysis algorithms can be used to identify the morphological features of the word.

There are two algorithms of morphological analysis for Russian language which are the most popular: the software application MyStem from the company “Yandex”, the principle of which is described in [4], and the library for the Python language – pymorphy described in [5].

Three texts were prepared for testing the algorithms:
1) natural language text in pre-revolutionary spelling,
2) the same natural language text but in modern spelling,
a control set of words in pre-revolutionary spelling, the writing of their morphemes differs from the modern, and also the words that have the same spelling of “problematic” morphemes, but do not require the translation.

The following parameters are selected for evaluating the accuracy of the algorithms:
1) the presence of a correctly defined lemma,
2) the presence of a correctly defined part of speech,
3) the presence of a correctly defined pair “lemma-part of speech”,
4) the presence of all characteristics (tags) which are correctly defined at the same time,
5) the percentage of correct analysis results which are evaluated as first, second and etc. (they are combined).

During the testing it was found that in most cases the correct definition of morphological characteristics depends on the correctly established lemma and to what part of speech the word is belong to.

Since each of the analyzers produces several variants, it is necessary to determine what is the place of the correct variant. To do this, we use the 5th parameter.

The algorithms show similar results when we work with texts in Russian in modern spelling, and these results are close to those that were obtained in the work [6], in which the tests were carried out on the text corpus of 6 million tokens. However, when we work with pre-revolutionary spelling, the best results were shown by MyStem. The test results are presented in tables 1 and 2.

| Table 1. Test results of Mystem |
|-----------------------------------|
| Parametr, % | Text 1 | Text 2 | Text 3 |
| Lemma | 97,22 | 97,22 | 78,57 |
| Part of speech | 100 | 100 | 100 |
| Lemma-Part of speech | 97,22 | 97,22 | 78,57 |
| All tags | 97,22 | 97,22 | 78,57 |
| 1st place | 74,29 | 74,29 | 36,36 |
| 2nd place | 22,86 | 22,86 | 54,54 |
| Below 2nd place | 2,85 | 2,85 | 9,09 |

| Table 2. Test results of pymorphy |
|-----------------------------------|
| Parametr, % | Text 1 | Text 2 | Text 3 |
| Lemma | 63,89 | 97,22 | 28,57 |
| Part of speech | 77,78 | 97,22 | 78,57 |
| Lemma-Part of speech | 63,89 | 97,22 | 21,43 |
| All tags | 58,34 | 91,67 | 21,43 |
| 1st place | 71,43 | 69,69 | 100 |
| 2nd place | 23,81 | 27,27 | 0 |
| Below 2nd place | 4,76 | 3,03 | 0 |

To improve the accuracy of the analyzers when we are working with pre-revolutionary spelling, a temporary modification of words is performed, which consists in replacing of obsolete letters with modern equivalents. The test results are presented in tables 3 and 4.
Also, based on the test results, we can see that *pymorphy* shows much better results with modifying the input words than without it. It is explained by the fact that in some words it is enough to replace the letters to get the modern spelling, which can be easily analysed by *pymorphy*. The results of *MyStem* do not change after the modifications, because in its dictionaries (unlike *pymorphy*) there is a significant number of analysed words in pre-revolutionary spelling.

Despite the advantage, *MyStem* also makes significant number of errors when it works with the words in pre-revolutionary spelling, what significantly reduces its accuracy. Thus, we can conclude that there is a need for a morphological analyzer which is adapted for a work with old spelling. For example, it can be a new special algorithm. However, as the algorithms of the analyzers themselves differ insignificantly, and the accuracy of the analysis mainly depends on the used dictionaries (for instance, *MyStem*, which uses different dictionaries, including some from the National corpus of the Russian language, which contains the texts in pre-revolutionary spelling), it will be advisable to use one of the existing algorithms with some modifications: pre-processing of input words and/or post-processing of the performed analysis results.

As Yandex distributes its application only in the form of a console utility, what makes difficult to use it in the developed program, since the usage of the wrapper only slows down the speed of operation, the *pymorphy* library was chosen as an algorithm for morphological analysis.

### 4. Adaptation of *pymorphy* library for a work with pre-reform spelling

At the testing stage it was noticed that the replacement of obsolete letters with their modern equivalents leads to a significant improvement in the accuracy of morphological analysis. This modification is used for the adaptation *pymorphy* for the work with pre-reform spelling, however, it is not enough.

The original idea was to decide on the basis of morphological characteristics whether to make replacement (translation) or not. For example, in the word *просто* the ending will be replaced by -бы, because it is a noun of 3rd declension in ablative case. But for the noun *évolution* such replacement will not happen as it is of 1st declension in accusative case. This approach is not suitable, because the testing revealed errors in the work of *pymorphy*, consisting in the fact that for words whose case endings differ from modern spelling, the lemmatization is incorrectly performed, while the right definition of the lemma depends on the definition of morphological characteristics which are required for the translation. For instance, *pymorphy* correctly defines the lemma *красный* for the word *красного*.
But for the same adjective in pre-reform spelling – краснаго – the lemma is not determined correctly (possible options: красной, краснаго etc.)

The new idea, that was used in the development of the algorithm, is to make first a temporary replacement of the ending with its modern equivalent. In this case we will get the modern form of a word or a nonexistent or other (new) word. Let’s consider how these two options are realized by the example of the words простию and эволюцію:

1) the word is given as input data to the translation algorithm:
   a) простию,
   b) эволюцію;

2) the replacement of the obsolete morpheme in the word is performed (-ію):
   a) простью,
   b) эволюцію.

With successful replacement the morphological characteristics of the word correspond to the expected ones. That is, the word with replacement – простью – will be defined as a noun of 3rd declension in ablative case, so the translation is done. Otherwise, the replacement leads to a change in the word itself, and not to its translation (эволюцію), so the characteristics will differ from the expected. In this case, the word will be transformed to the original form (эволюцію).

The pymorphy analyzer returns all possible variants of morphological analysis. Each of them has an estimate of the probability that it is correct. The variant of analysis, the evaluation of which is the highest, really turns out to be true in about 79% of cases [8]. However, this is not enough, so the translation algorithm takes into account all possible variants.

5. The rules of translation

For each group of words which are united by morphological features, when they are translated from pre-revolutionary spelling into modern one, we determine the exact sequence of steps of the algorithm.

5.1. The prefixes

We consider the limited number of prefixes: -из, -воз/вз, -раз/роз, -низ, -без, -через/чрез. In the old spelling the letter з was kept before a letter с after prefix for the first four prefixes. In modern spelling in this case the letter з is replaced by с. The last two prefixes in the pre-reform spelling always had з at the end (besides the words exceptions), but in the modern one – з is replaced by с before the voiceless consonant.

1. The algorithm looks for the occurrence of the prefix in the word.
2. If it is a prefix of the first four on the list, the algorithm checks the next letter.
   2.1. If the letter is с, then the algorithm replaces з with с.
   2.2. Else — no changes.
3. Else the algorithm determines whether the subsequent consonant is voiceless or voiced.
   3.1. If consonant is voiceless, then the replacement on е is done.
   3.2. Else — no changes.

Example: 1) разбудить – no changes, since б is not voiceless; 2) разсказать – replace by с, sine the condition 2 is satisfied.

5.2. Case endings

The rules for case endings take into account the results of testing the pymorphy library used for morphological analysis of words, during which the patterns of errors and the principle of their correction (that was described above) were derived.

A noun, ablative case, III declension

In the old spelling the words of this type had a variative ending -ію (main), sometimes – -ью (additional). In the modern spelling only the additional form has left.

1. In the case of -ью the algorithm leaves the word unchanged.
2. Else:
2.1. The algorithm looks for the -ію from the end and replaces it with -вю.
2.2. The word after replacement is given to analyzer input.
2.3. If a noun is of III declension (i.e. feminine and its lemma ends on -ѣ) and the ablative case, then the algorithm keeps the replacement of the ending, otherwise – returns -ію.

Example: 1) революцію – the algorithm returns -ію, since революцію is not of III declension; 2) простію – the algorithm replaces the ending with -вю, since the condition 2.3 is satisfied.

An adjective, feminine/neuter, plural

Such adjectives, which depend on nouns of feminine and neuter in the plural, in the old spelling had the endings -ыя, -ія. In the modern version -ыє, -іє are respectively used. Such endings can be found not only in adjectives, but also in nouns, so for the correct translation it is necessary to exclude the replacement of the endings in nouns.

1. The algorithm looks for -ым/-ім from the end.
2. The algorithm replaces -ым/-ім with -ым/-іє, respectively.
3. The word after replacement is given to analyzer input.
4. The algorithm chooses a variant of the analysis, in which the part of speech is defined as an adjective, and the number is plural. If such variant is not found, then the algorithm returns -ым/-ім. Otherwise – it goes to step 5.
5. The algorithm gives a lemma (from chosen variant) to the analyzer input. If the word is an adjective and really is a lemma (i.e. masculine gender, singular number and nominative case), then the algorithm keeps the replacement, otherwise – the algorithm returns -ым/-ім.

Example: 1) революція – the algorithm returns -іє, since this word is not an adjective; 2) красныя – the algorithm performs the replacement with -ыє, since the word красныя and its lemma are the adjectives.

The nouns, prepositional case, neuter, II declension (with the ending -ѣ)

Such words could have the ending -ы. The ending -ы can be found only in words whose lemma ends on -ѣ (II declension). So it is enough to perform the following steps:

1. The algorithm looks for the -ы from the end.
2. The algorithm performs the replacement with -ѣ.

In poetic texts, the -ы form is acceptable for nouns whose lemma ends on -іє, but not for the words with lemma on -ѣ. Therefore, if the poetry mode is selected, the following sequence of actions will be performed:

1. The algorithm looks for -ы from the end.
2. The replacement with -ѣ is done.
3. The word after replacement is given to analyzer input.
4. If the lemma ends on -ѣ, then the replacement is kept, else (lemma ends on -іє) -ы is returned.

Example: 1) в платье – translated to в платье; 2) в забвенье – -ы is kept, if poetry mode is on, since its lemma is забвенье. Otherwise the replacement with -ѣ is done.

An adjective/pronoun/participle/number, genitive/accusative, masculine/neuter (with the endings -ым, -ім)

Such words had the endings -аго, -аго. In modern spelling they correspond to the endings -озо, -его. The ending -озо is used in the case of soft consonant and all (including solid ones) hissings before it (exactly, in the ending a vowel itself softens or leaves a hard-consonant sound before it). The ending -озо is used in other cases. The ending -аго was used in the case of a soft consonant before it, so the original ending can be replaced with -го. The ending -аго was used in the case of a hard consonant or hissing (any – both hard and soft) before it, so after the hissing – the ending is replaced with -го, otherwise – with -озо. Also the endings -аго/-аго can be found in some nouns and should be conserved.

1. The algorithm looks for -аго/-аго from the end.
2. If the word is found in the analyzer dictionary, then the algorithm skips this word. Else the step 3 is performed.
3. If the ending is -аго:
3.1. if the previous letter is hissing, then the ending is changed to -его;
3.2. else – with -ого;
4. If the ending is -яго, then it is replaced with -его.
5. The word after replacement is given to the analyzer input.
6. The algorithm looks for the variant of analysis in which the same features (part of speech, case, gender) are found. If such variant was found, then the lemma from it is given to the analyzer input and the step 7 is performed. Else -аго/-яго are returned.
7. If the lemma is derived correctly, so the word is in its initial form and its part of speech is defined also right, then the replacement is kept, else -аго/-яго are returned.

Example: 1) лучшаго – лучшего, since the condition on the step 2 is not satisfied; 2) люмбаго – no changes, since the condition on the step 2 is satisfied.

Онѣ, однѣ, однѣхъ, однѣмъ, однѣми
Such words were always used only with the feminine gender, so the replacement of the letter is not suitable (because not one but они). These words can be stored in some dictionary with its equivalents (with the letter у), and when a dictionary word is found, it will be replaced with the equivalent.

Ея (нея) – pronoun, genitive/possessive pronoun
This word will be replaced with its equivalent from the dictionary (as in the case above). If the mode without ё is on, then the equivalent is ee (нее), else её (неё).

Ее (нее) – pronoun, accusative
This word will be replaced with its equivalent from the dictionary (as in the case above). If the mode without ё is on, then the replacement is cancelled, else её (неё).

6. Test results of the translation algorithm
During the testing, the accuracy of translation was calculated in different modes with different texts. The test results are shown in the table 5.

| Text appearance | The poetic mode | The mode of support of the letter ё | Result, % |
|-----------------|-----------------|-----------------------------------|-----------|
| Poetic text №1  | +               | -                                 | 100       |
| Prosaic text    | -               | -                                 | 100       |
| Poetic text №2  | +               | +                                 | 100       |
| Non-fiction     | -               | +                                 | 99.07     |

During the processing of the non-fiction text, some cases were identified for which the conversion rules are not formalized and are not taken into account in the algorithm. For example, the word концемь. The ъ was removed in the translation, however, the ending has not changed, what does not comply with modern spelling rules, according to which the word has the form of концом. Such writing is due to the fact that the sound [?] in the Russian language was soft sometime [8, p. 152].

7. Conclusion
In this paper we have studied and tested the basic algorithms of morphological analysis of words. The evaluation of the test results revealed the patterns in the errors made by morphological analyzer pymorphy during the work with pre-revolutionary spelling. The obtained regularities were used in the adaptation of this analyzer to the work with pre-reform spelling. On the basis of the revealed dependencies between the rules of old and modern spelling the formal rules of translation are described. Thus, on the basis of the above results, the algorithm has been developed to translate the texts in Russian from pre-reform spelling to modern one, taking into account the morphology of words. It should be noted that due to the insufficiently strict (in comparison with modern one), codification of the pre-reform spelling, as well as earlier, the more complex algorithms will be required to create the programs that allow with minimal distortion to translate texts of the middle of 19th century and earlier into modern
spelling, possibly including machine learning methods. However, for the processing large volumes of documents related to the Russian history of the late 19th – early 20th century, the proposed algorithms are sufficient.

Acknowledgments
The work was supported by Russian Science Foundation grant № 19-18-00466.

References
[1] Tseytlin R M 1972 Grot Yakov Karlovich Big Soviet Encyclopedia 3rd edition (Moscow: Soviet Encyclopedia) vol 7 p 353 (in Russ.)
[2] Davydov P I 2013 The refernece of the Russian old spelling (Moscow: Kitony) (in Russ.)
[3] Slavenica http://slavenica.com
[4] Segalovich I 2003 A fast morphological algorithm with unknown words guessing induced by a dictionary for a web search engine Proceedings of the International Conference on Machine Learning; Models, Technologies and Applications pp 273–280 (in Eng.)
[5] Korobov M 2015 Morphological analyzer and generator for Russian and Ukrainian languages International Conference on Analysis of Images, Social Networks and Texts pp 320–332 (in Eng.)
[6] Dereza O V, Kayutenko D A and Fenogenova A S 2016 Automatic morphological analysis for Russian: A comparative study Computational Linguistics and Intellectual Technologies; Proceedings of the Annual International Conference “Dialogue” pp 1–13 (in Russ.)
[7] User Manual – Morphological analyzer pymorphy2. http://pymorphy2.readthedocs.io/en/latest/user/guide.html#select-correct
[8] Barkovskiy V I and Kuznetsov P S 2016 Historical Grammar of Russian language (Moscow: LENAND) (in Russ.)