Proceedings of the Fifth Conference of Mathematical Society of Moldova
IMCS-55, September 28 - October 1, 2019, Chisinau, Republic of Moldova

Generation of the Romanian Cyrillic lexicon for the period 1967 – 1989

Constantin Ciubotaru, Valentina Demidova, Tudor Bumbu

Abstract

This paper is dedicated to the issue of generation of the Romanian Cyrillic lexicon used between 1967 and 1989. The rules for transliteration of words in the modern Romanian lexicon in their equivalents written in Cyrillic and vice versa are specified and argued. The respective algorithms have been developed and implemented, which is an expert tool in the lexicon generation process. The activity of the expert is reduced to the verification of the transliterated variants and the modification of the transliteration rules.

Keywords: lexicon, transliteration, Moldovan Cyrillic, cyrillization, romanization of Cyrillic, morpho-syntactic tagsets.

1 Introduction

The problem of digitizing and preserving the linguistic historical heritage is a priority domain of the digital agenda for Europe. The digitization process requires solving a series of problems related to the recognition, editing, translation, and interpretation of printed texts. The solution of these problems has to do with specific difficulties: a large number of periods in the evolution of the language, a small volume of resources widely distributed, a large diversity of alphabets used in their printing.

The paper addresses the issues related to the digitization and transliteration of the historical linguistic heritage printed with Cyrillic alphabet in the period 1967–1989 on the territory of the Moldovan Soviet Socialist Republic following the linguistic norms of the modern Romanian language. The alphabet used in this period (the Moldovan Cyrillic alphabet, AlfCYR) is, in fact, the alphabet of the Russian language from which the letters "ё", "щ" and "ъ" were excluded and extended in 1967 by the introduction of the letter "ӂ". Specific difficulties: lack of resources in electronic format, fragmentary grammatical descriptions that allow ambiguous interpretations.

According to the definition of dexonline [1] *transliteration* is "the transcription of a text from one alphabet to another, rendering the letters by their equivalents, without regard to the phonetic value of the signs".

The process of transliterating the Romanian words in their written equivalents with characters of the AlfCYR alphabet we will call *cyrillization*. For instance, *puiului*⇒ *пуюлуй* (chicken), *fiului*⇒ *фиулуй* (son), *cenușiu*⇒ *ченушиу* (greyish), *viermi*⇒ *вермь* (worms), *vierii*⇒ *виерий* (boars).

The inverse procedure for cyrillization will be called *romanization of Cyrillic*, e.g. *пуюлуй* ⇒ *puiului* (chicken), *бъет* ⇒ *biet* (poor), *боер*⇒ *boier* (boyar), *пепт*⇒ *piept* (chest).

Taking into account the lack of Romanian Cyrillic resources for this period, the following lexicon generation algorithm is proposed. Sets of rules are defined after which the algorithms for cyrillization and romanization of Cyrillic are constructed. The cyrillization algorithm is applied to the Romanian lexicon (we used the lexicon developed at the "Al. I. Cuza", Iasi [2]), noted by LexROM. There will be obtained a variant of the Cyrillic lexicon which can be further processed by the algorithm for the romanization of Cyrillic, thus obtaining a new variant for the Romanian lexicon. The ideal situation would be for these two lexicons to coincide. If mismatches occur, the expert(s) intervenes, who can modify the rules of cyrillization \romanization of Cyrillic, repeat the whole process or intervene with corrections on the constructed

Cyrillic lexicon. The diagram of this process is inserted in Figure 1. The accuracy of the obtained Cyrillic lexicon largely depends on the qualification of the expert(s).

2 Cyrillization

If the problem of digitizing and recognizing the printed text is solved relatively simply, then the problem of cyrillization is more difficult. The rules of transliteration according to cyrillization can be divided into two categories: general rules and context-sensitive rules. The general rules establish the situations when the transliteration directly substitutes letter with a letter(s) ignoring the contextual dependencies. Thus, the letter *b* will be always transliterated into *б*, we will note this through $b \Rightarrow$ *б*. The same action will be performed for the following pairs: $a \Rightarrow$ *а*, $ă \Rightarrow$ *э*, $â \Rightarrow$ *ы*, $c \Rightarrow$ *к*, $d \Rightarrow$ *д*, $e \Rightarrow$ *е*, $f \Rightarrow$ *ф*, $g \Rightarrow$ *г*, $h \Rightarrow$ *х*, $i \Rightarrow$ *и*, $î \Rightarrow$ *ы*, $j \Rightarrow$ *ж*, $k \Rightarrow$ *к*, $l \Rightarrow$ *л*, $m \Rightarrow$ *м*, $n \Rightarrow$ *н*, $o \Rightarrow$ *о*, $p \Rightarrow$ *п*, $r \Rightarrow$ *р*, $s \Rightarrow$ *с*, $ș \Rightarrow$ *ш*, $t \Rightarrow$ *т*, $ț \Rightarrow$ *ц*, $u \Rightarrow$ *у*, $v \Rightarrow$ *в*, $x \Rightarrow$ *кс*, $z \Rightarrow$ *з*.

For instance, *chirilizarea* $\Rightarrow$ *кхирилизареа* (cyrillization). Of course, that *chirilizarea* $\Rightarrow$ *кирилизаря* would be correct. In this case, the contextual dependencies intervene. Because such dependencies are many and quite complicated, we will only expose a few. More information on this topic can be found in [4]. From the example above we notice that the transliterations *chi* $\Rightarrow$ *ки* and *ea* $\Rightarrow$ *я* are correct. A few more contextual rules: *gi* $\Rightarrow$ *жи*, *ghi* $\Rightarrow$ *ги*, *ci* $\Rightarrow$ *чи*, *ci* $\Rightarrow$ *ч*, *iu* $\Rightarrow$ *ю*, *iu* $\Rightarrow$ *иу* (the examples in paragraph 1).

Many difficulties arise when transliterating the letter *i* at the end of the word. All the possible transliterations: $i \Rightarrow$ *и* (*citi* $\Rightarrow$ *читы* (read)), $i \Rightarrow$ *й* (*pui* $\Rightarrow$ *пуй* (chicken)), $i \Rightarrow$ *ь* (*arici* $\Rightarrow$ *аричь* (hedgehog), plural), $i \Rightarrow$ (*arici* $\Rightarrow$ *арич*, singular). To make the correct decisions, contextual rules can sometimes be supplemented with morpho-syntactic information. We use MSD (morpho-syntactic-description) tags present in the LexROM lexicon [2]. For example, for infinitive verbs the letter *i* at the end of the word will be translated into *и*, the masculine nouns in the

plural nominative-accusative case the articulated form will end in *ий*, and the transliteration $i \Rightarrow ь$ at the end of the word is characteristic for nouns in the plural dative-genitive case, and also for verbs, the second person present, past, more than perfect.

The order of the two categories of transliteration rules is very important. Contextual dependencies always take precedence over general rules.

CYRILLIZATION ALGORITHM

0. ***Start***

1. The lexicon of the modern Romanian language is given [2] (we will note it LexROM_1) and the rules of cyrillization (general and context sensitive).
* We will build the Romanian Cyrillic lexicon for the period 1967–1989 (we will note it by LexCYR) *\

2. Initially LexCYR $= \varnothing$

3. ***For all*** words *wrom* from LexROM_1:

3.1. Apply on *wrom* the context-sensitive rules for cyrillization. Note the result by $wcyr_1$.

3.2. Apply on $wcyr_1$ general rules for cyrillization. Note the result by *wcyr*.

3.3. Include *wcyr* in LexCYR.

4. ***Stop***

3 Romanization of Cyrillic

Romanization of Cyrillic is facing the same problems as cyrillization. The general and contextual rules for this procedure are also defined. The general rules are relatively simple, for example, $а \Rightarrow a$, $р \Rightarrow r$, $ю \Rightarrow iu$, $ь \Rightarrow i$. If only the general rules apply to transliteration, we obtain, for instance, *пуюлуй* $\Rightarrow$ *puiului*, *бъет* $\Rightarrow$ *biet*, *боер* $\Rightarrow$ *boer*, *пепт* $\Rightarrow$ *pept*. The last two transliterations are incorrect and their correct variants are: *боер* $\Rightarrow$ *boier*, *пепт* $\Rightarrow$ *piept*. In this case contextual

rules are also needed. E.g, *гω* ⇒ *ghω*, if *ω* ∈{*е,и,я,ю,ъ*} and *гω* ⇒ *gω*, if *ω* ∉{*е,и,я,ю,ъ*} (*георгинэ*⇒*gheorghинэ* (dahlia), *гогоашэ* ⇒ *gogоашэ* (donut)). Rules for letter *я*: *я* ⇒ *ia* (usually at the beginning of the word), *ия* ⇒ *ia* (usually at the end of the word), *я* ⇒ *ea*. It is very difficult to make the right decision for the presence of the letter *я* inside the word. In such situations the algorithm will use the rule *я* ⇒ [*ia*][*ea*], leaving the correct decision to the expert.

Another difficult problem to mention is the transliteration of the letter *ы*, which can be substituted either by î or by â. Our algorithm follows the recommendations of the Romanian Academy regarding this spelling.

ROMANIZATION OF CYRILLIC ALGORITHM

0. ***Start***

1. The Romanian Cyrillic Lexicon for the period 1967–1989 (LexCYR) and romanization of Cyrillic rules (general and context sensitive) are given
* We will build the modern Romanian lexicon (we will note it by LexROM_2) by applying the transliteration method *\

2. Initially $\text{LexROM}_2 = \varnothing$

3. ***For all*** words *wcyr* from LexCYR:

3.1. Apply on *wcyr* the context-sensitive rules for romanization of Cyrillic. Note the result by $wrom_1$.

3.2. Apply on $wrom_1$ general rules for romanization of Cyrillic. Note the result by $wrom_2$.

3.3. Apply on $wrom_2$ rules of transliteration for letter *ы*. Note the result by *wrom*.

3.4. Include *wrom* in LexROM_2.

4. ***Stop***

4 Lexicon generation technology

As mentioned above, the electronic resources for the period 1967–1989 are completely absent, complete exposure of the grammar used is missing, many of the interpretations of the transliterated words are ambiguous. Therefore, the expert plays a major role in the lexicon generation process. The proposed technology aims to automate this process. With the available cyrillization and romanization of Cyrillic algorithms, but also access to the formalized rules, the lexicon generation process can take a few iterations. At each iteration, the expert(s) intervene(s) to modify the set of rules and, possibly, directly, the Cyrillic lexicon. This scheme is described in detail in Figure 1.

5 Conclusion

The paper proposes a technology for the generation of the Romanian Cyrillic lexicon for the period 1967–1989 applying the transliteration method. Starting from the lexicon of the modern Romanian language elaborated at the University "Al. I. Cuza", Iasi [2] (1.096.674 words) the cyrillization and romanization of Cyrillic algorithms are applied consecutively. The intermediate results are available to the experts, who can modify or extend the set of rules applied to transliteration, but also directly correct the obtained Cyrillic lexicon. The final lexicon will be a result of performing several such iterations. The main problems that need to be solved by the expert(s) are the ambiguities that arise as a result of cyrillization\romanization of Cyrillic. For example, *iu* ⇒ [*ю*] [*uy*], *eie* ⇒ *е* [*e*] [*ĭe*] [*ue*] – at cyrillization, *я* ⇒ [*ia*] [*ea*] – at romanization of Cyrillic.

For all the words in LexROM starting with the letter "c", in total 171846 words, 6381 ambiguities were found at the first iteration, which represents 3.7 %. Two iterations were needed to overcome these ambiguities. Of course, the accuracy depends considerably on the qualification of the expert. Also here, at the first iteration comparing the lexicon LexROM_1 with LexROM_2 we obtained 8960 words that do not

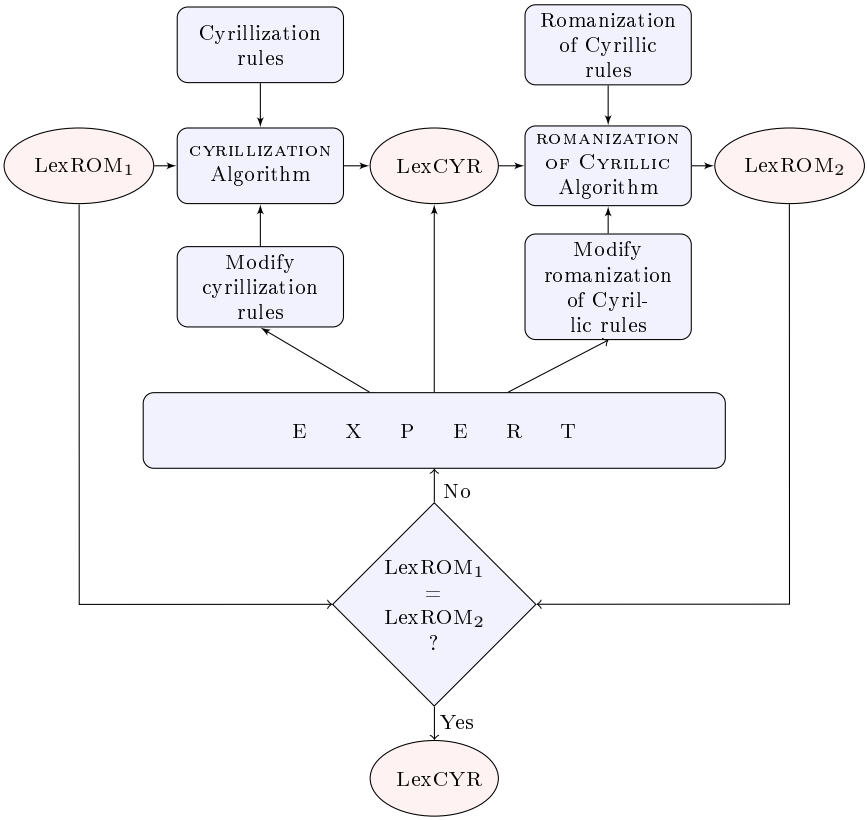

Figure. 1. The scheme for generating the Romanian Cyrillic lexicon

coincide, which represents about 5.21%. Most of these mismatches are incorrect general rules for the romanization of the Cyrillic. The proposed technology allows returning to the previous intermediate variants, thus revising the lexicon. To better understand the role of the expert and contextual dependencies, we applied on LexROM lexicon only the general rules (paragraph 2) of cyrillization. As a result, we got only

42.2% correct words.

Bibliography

[1] https://dexonline.ro/definitie/translitera%C8%9Bie.

[2] http://nlptools.info.uaic.ro/WebPosRo/resources/posDictRoDiacr.txt

[3] S. Cojocaru, E. Boian, C. Ciubotaru, A. Colesnicov, V. Demidova, L. Malahov. *Regeneration of printed cultural heritage: challenges ang technolologies.* Chişinău: The Third Conference of Matematical Society of the republic of Moldova, 19-23 August, 2014, pp. 481–489.

[4] V. Demidova. *Particular Aspects of the Cyrillization Problem.* Chişinău: The Third Conference of Matematical Society of the republic of Moldova, 19-23 August, 2014, pp. 493–498.

[5] V. Demidova. *The Peculiarities of the Romanization of Cyrillic.* Studia universitatis Moldaviae, 2015, no. 2(82), Seria "Științe exacte și economice", pp. 16–20 (in Romanian).

Constantin Ciubotaru[1], Valentina Demidova[1], Tudor Bumbu[1,2]

[1]Vladimir Andrunachievici Institute of Mathematics and Computer Science
[2]The State University "Dimitrie Cantemir"

E–mails: chebotar@gmail.com,demidova@math.md,bumbutudor10@gmail.com