On Technology for Digitization of Romanian Historical Heritage Printed in the Cyrillic Script

Svetlana Cojocaru, Lyudmila Burtseva, Constantin Ciubotaru, Alexandru Colesnicov, Valentina Demidova, Ludmila Malahov, Mircea Petic, Tudor Bumbu, Ştefan Ungur

Abstract

This paper describes the elaboration of the technology for digitization of the Romanian historical heritage printed in the Cyrillic script in the 17th–20th centuries.

The attention is focused to transliteration of recognized Cyrillic texts to the modern Latin script, to difficulties with older alphabets that are not fully supported by modern OCR engines, and to other concomitant problems.

We proposed solutions for these problems and integrated them into a corresponding technology and a tool pack that includes: alphabets, dictionaries, glyph patterns, transliteration and glyph restoration utilities, virtual keyboards, fonts, and user's manual.

Keywords: historical Romanian texts, OCR of Romanian Cyrillic scripts, 17th-20th century, software tools for OCR, transliteration utility.

1 Introduction

Problem of digitization and conservation of historic, literary, and cultural treasures represents a domain of priority in the Digital Agenda for Europe. The EU admits the necessity of coordinated efforts in this domain and manifests vast actions to activate this process. These actions include development of the European Digital Library *Europeana*, supported by the European Parliament resolution on the 5th of May, 2010 and the adopted EU Programs for Culture. Diverse aspects of this problem were treated in many European research projects [1]. In particular, the problem

of creation of linguistic resources, digitization and recognition of historic and literary heritage is attended in many European countries [6]–[13]. Regrettably, scientific centers of the Republic of Moldova aren't involved in these actions.

Massive usage of information technologies and communications (ITC) strongly stimulates development of the modern society and substantially contributes to the conception of information society. The Digital Agenda presented by the European Commission forms one of the seven pillars of the Europe 2020 Strategy. It adds dynamics and optimizes ITC benefits for economic growth, creation of new jobs, increase peoples' quality of life.

The Decision of the Government of the Republic of Moldova No. 857 of the 31st of October 2013 approves the National Strategy for the development of information society "Digital Moldova 2020", and the Plan of Actions for implementation of this Strategy. The Program "Creation, development and evaluation of the digital content in the RM in 2016–2020" is in the process of approbation.

Digitization and conservation of the cultural historic and linguistic heritage that includes old literature, archive documents, folklore records, etc., represent one of key domains affected by the Digital Agenda. This process will be related to the heritage preservation while its placement in the Internet will considerably simplify its usage, will extend area and possibilities for research, including in humanitarian domains, through modifications in international media of communication. In addition, execution of the planned works will permit unification, homogenization, and integration of national and cultural media in the international information society, and will confirm status of the Romanian language as language of communication in the European continent.

Although the cultural heritage domain has been intently researched during last decades, today the research will more focus on its multilingual nature and specific for each culture features. Digital age arrival passed the problem of cultural heritage preservation from conservation laboratories to computers. The cultural heritage presented in text form has showed the most suitable and informative digital representations. Texts processing is a highly developed domain today. The research of historical texts has developed specific methods of text processing, mostly, tools for

representation of unusual today scripting. Following the distribution of the Unicode over operating systems, the problem of encoding was solved for any historical script. To materialize old text in electronic form, we need now only specialized fonts covering the corresponding code points. Let's note that several Romanian Cyrillic letters (e.g., ѧ) were included in the Unicode only since 2009. But, being appropriate for preservation of textual cultural heritage, unusual fonts are difficult for perception even for linguistics professionals. Therefore, solving the problem of textual cultural heritage dissemination supposes the development of tools for transliteration or just reading in common script.

Solving these problems for the Republic of Moldova confronts difficulties and specific aspects: the number of existing resources is relatively small but they are kept in many book deposits; they were printed in a lot of diverse alphabets. Thus, old manuscripts and books in Moldova and Romania were produced, as a rule, in the old Romanian Cyrillic script (RC) [2], that differs from standard Church-Slavonic or Russian ones. The definitive formation of RC is dated back to the 17th century. The first Romanian grammar was printed by D. Evstatievici in 1757. Since 1830 until the official adoption of the Latin alphabet (RL) in 1862 several transitional alphabets (TR) were used; they were based on the Simplified Romanian Cyrillic script (SRC) but some letters were Latin [5]. The modern Romanian Latin alphabet (MRL) was adopted in 1904; with small variations, it is used till present. Variants of the Cyrillic alphabet that were used in the Moldavian ASSR in 1924–1940 and in the Moldavian SSR in 1940–1989 (the Moldavian Cyrillic script, MC) were an integral and irregular application of the Russian alphabet for the Romanian language.

Electronic sources exist mostly for old Romanian books printed in the Latin script, while those for the Cyrillic script practically don't exist except as scanned images. That's why the problem appears to create electronic Romanian resources of manuscripts and old books in the Cyrillic script. To create electronic resources of the cultural heritage printed in the corresponding periods we could use, for example, catalogs [3] and [4] from the old book repository at the "A. Lupan" central scientific library of the Academy of Sciences of Moldova (ASM).

This paper describes a technology for digitization and recognition of the historic and linguistic Romanian heritage printed in the Cyrillic script in the 17th–20th centuries. The technology is supported by a pack of the following tools and utilities:

- Alphabets for ABBYY FineReader (AFR).
- Dictionaries (word lists) for AFR.
- Recognition patterns as trained under AFR.
- Utility of transliteration from Cyrillic to Latin and vice versa for MC.
- Conversion utility for TR.
- Conversion utility for RC.
- Font that covers rare glyphs from RC and TR.
- Virtual keyboards.

Algorithm of verification of resulting text in the Latin script and semi-automated word recognition could use the Romanian spellchecker RomSP [15] and reusable linguistic resources [14].

We will concentrate on details of the transliteration and conversion rules.

2 Recognition of the Romanian Cyrillic Script

We began our work as we desired to re-publish some books printed in this alphabet. Under the USSR, the editorial activity produced many useful and interesting texts, but they are of no use to contemporary Romanian audience being printed in the Cyrillic script.

In the period of our interest (1951-1989) the printing quality was quite satisfactory, and scanning goes smoothly. The Moldavian Cyrillic script (MC) was used. It is the Russian alphabet without letters **ё**, **щ**, **ъ** and, since 1967, with one additional letter **ӂ**. At OCR, we were to add letter **ӂ** to the Russian alphabet and provide the dictionary. The dictionary was extracted from recognized texts themselves with manual corrections; then we repeated OCR. See details in [17].

The second referred variant is the Romanian Cyrillic alphabet of 1830–1860. The script was transitional from Cyrillic to Latin (TR). It was Cyrillic in its base with some letters replaced progressively by Latin ones.

We used two approaches to OCR of Romanian transitional scripts. The first approach is to reproduce the scanned text after OCR in its

original glyphs. It is possible with the corresponding AFR configuring and training, and by providing the proper dictionary. It produced up to 7% of erroneous words.

The second approach was invented to solve the problem of alphabet variation. AFR permits to output the result in original glyphs, or replace any glyph by a sequence of letters from the selected alphabet of recognition. AFR proposes this mode for ligatures but it may be used more generally for arbitrary substitution. For TR, we formed a version of the AFR output alphabet that can be set in one-to-one mapping with any transitional alphabet. For example, both **т** (Cyrillic) and **t** (Latin) will be recognized as **t**.

Another problem common for all variants of TR and RC is the absence of their glyphs in the usual system fonts. As the result, we do not see them in AFR dialogs during alphabet preparation, training, manual text correction, etc. The use of glyph substitution solves this problem also [18].

The third period of specific Romanian Cyrillic script usage is since the mid 18th century till 1830 (referred for simplicity as the 18th century). The Romanian typography practices of the 18th century had had two substantial differences from that of the older time, with the same RC of up to 47 letters. First, the usual Arabic number system is used. Second, upper accents had become rare and may be ignored. Therefore, the recognition doesn't imply sophisticated training.

AFR recognizes RC of the corresponding period. Small problems arose due to absence of necessary glyphs in system fonts, as it was already noted

The recognition of texts of the 18th century resulted in 4.5% of erroneous words with original glyphs, and only 3% of erroneous words with ligatures. We observed this effect with transitional scripts also.

The most plausible explanation is that, in the training mode, AFR skips some glyphs that are supposed to be recognized properly. With original glyphs, AFR skips more glyphs, while, at the glyph substitution, AFR should train substituted glyphs and performs more scrupulous training.

A special utility was developed that restores the original glyphs after recognition with substitutions for the texts of the 18th century.

The fourth period covers the 17th century and the 1st half of the 18th century when the Romanian typographies had strictly adhered the previous manual writing practices. This means that the numbers were encoded by letters with special ascending strokes, and accents over the line were substantial. Some words were traditionally printed with abbreviations and were also marked over them. Skipped letters were frequently set over the precedent letter, also with a special marker.

The recognition of such printing implies very subtle and thorough training. For example, each pair of a letter and another letter over it should be trained as a ligature.

Numbers (one or several letters with a marker) should also be trained as ligatures. This increases the number of recognition patterns, but, without ligatures, OCR for RC of the 17th century produces errors in more than 50% words, while with trained ligatures only in 6%.

3 Transliteration of the Recognized Text

3.1 Older Cyrillic Scripts in Unicode

The first problem is presentation of recognized Cyrillic text in computer, especially for TR and RC. In fact, only three fonts in the whole world have old Romanian Cyrillic letters: Kliment STD (ѳѳѢѣѧѧ), Unifont (ѳѵѢѣѧѧ), and Everson Mono (ѳѵѢѣѧѧ), and only since 2009. That's why we are developing for our tool pack our own font covering all necessary Unicode points. MC poses no such problems. In the period of our interest (1951–1989) the difference with the Russian alphabet was made by a single letter ѫ that is presented in commonly used fonts.

Some accented or combined letters are meanwhile missing and should be specially treated, for example, ӑ (ă) or ӗ̆ (ĭă) in TR. To present them in Unicode, it is necessary to use combining accents, and we can't fully reproduce subtle details of the graphical presentation of the original text.

Table 1. Correspondence of Some RC Specific Letters to MRL and Unicode.

Ѣ	Ea	0462	К	C, Ch (before **е**, **и**)	041A
ѣ	ea	0463	к	c, ch (before **е**, **и**)	043A
Ю̆	Ia	0465	Ĭ	Ĭ	012C
ю̆	ia	0464	ĭ	ĭ	012D
Ѧ	Â	0466	Ъ	Ă	042A
ѧ	â	0767	ъ	ă	044A
Ѧ̣	Î, Îm, În	A64E	Щ	Ș̌t	0429
ѧ̣	î, îm, în	A65F	щ	ș̌t	0449
Ꙉ	U	A64A	Ц	G	049F
ꙉ	u	A64B	**ц**	g	044F

3.2 MC: Bidirectional Transliteration

The transliteration MC→MRL was discussed in details in [17]. There are three groups of rules. Most letters (26 of 31) can be mapped one-to-one as shown in Table 2.

Table 2. MC→MRL: one-to-one letter mapping.

MC→MRL		**MC→MRL**		**MC→MRL**		**MC→MRL**	
а	a	з	z	п	p	ц	ț
б	b	и	i	р	r	ш	ș
в	v	й	i	с	s	ь	i
д	d	л	l	т	t	э	ă
е	e	м	m	у	u	ю	iu
ж	j	н	n	ф	f		
ӂ	g	о	o	х	h		

Context rules exist for three letters as shown in Table 3.

Table 3. MC→MRL: Context Rules in the Order of Application.

MC→MRL		Context
г	**gh**	before **е**, **и**, **ь**, **ю**, **я**
г	**g**	otherwise
кс	**x**	exceptions: **eczema** and derivatives, **Alecsandri**
к	**k**	as exception, examples: **kilogram**, **Kogălniceanu**, etc.
к	**ch**	before **е**, **и**, **ь**, **ю**, **я**
к	**c**	otherwise
ч	**c**	before **е**, **и**, **ь**, **я**
ч	**ce**	before **а**
ч	**ci**	otherwise

The letter **ы→â**, **î**, where **î** is written at the beginning or end of words, while **â** inside words. The difficulty is that **î** is kept after prefix, for example, **ne**+**însoți**t = **neînsoți**t (**unaccompanied**).

The letter **я→ea**, **ia**, **a** presents the biggest problem that can't be fully solved without access to dictionaries. Rules are mostly heuristic and statistical, and more than 20 rules do not cover all cases. This situation exists because MC was not thoroughly designed but is an irregular mapping of Romanian sounds to the Russian letters.

There are words that can't be transliterated according to these rules: foreign proper nouns and words of foreign origin that keep their writing in MRL. We use the exception dictionary for them.

The inverse transliteration MRL→MC (1967–1989) was mainly necessary to produce word list in MC from existing word list in MRL. This task equally meets difficulties, mainly with letter **i**. In particular, at the word ends **i** may be omitted, or converted to **и**, **й**, **ь**. Examples: **arici→арич** (**hedgehog**, singular), **arici→аричь** (**hedgehogs**, plural), [**а**] **cheltui→**[**а**] **келтуи** ([**to**] **count**; stress on **i**), [**eu**] **cheltui→**[**еу**] **келтуй** (**I count**: stress on **u**). Analogous problems appear at the transliteration of diphthongs and triphthongs. For example, diphthong **ia→я**, **ия**, **иа**: **soia→соя** (**soybean**), **caucazian→кауказиян** (**Caucasian**), **cartezian→ картезиан** (**Cartesian**).

Some of these problems could be solved by consulting Morpho-Syntactic Data (MSD), which were proposed in the framework of the project *MULTEXT-East* [19]. In the remaining cases, context analysis or even manual intervention could be performed.

The whole transliteration process is implemented as a set of filters each modelling a separate situation. The filters are:

- prefix filters;
- suffix filters;
- diphthong and triphthong filters;
- final filters (**letter→letter**).

Prefix filters are created separately for words that begin with the same letters (**creast*→кряст***, **crea*→креа***, **paie*→пае***, etc.). At transliteration, these filters are applied first.

Suffix filters are common for all words in the lexicon. They can be divided in two classes: conditional depending of the MSD value, and unconditional.

Diphthong and triphthong filters aim to transliteration of some letter combinations like: **ie**, **io**, **eio**, **chio**, etc. They are applied independently of position and context.

Final filters transliterate all letters that remain after application of other filters. For example: **d→д**, **c→к**, **ş→ш**.

Some filters can use rezults from the previous filters. Such filters may look like: **celуй→челуй**, **ieпь→ень**, combining Latin and Cyrillic letters.

If the situation is ambiguous, and expert's intervention (maual selection) is necessary, alternatives can be generated, for example: **кафен**[**иул**][**юл**] with the result **→кафениул** (**brownish**, **coffee color**; with the definite article); **ча**[**иул**][**юл**]**→чаюл** (**the tea**; with the definite article).

This algorithm was applied to the lexicon elaborated at the Al. I. Cuza University in Iași [20]. Automation rate of transliteration was approx. 90%.

3.3 Transitional Alphabets

Sources count approx. 17 versions of TR. In this paper we deal with 36 Cyrillic letters (from 43) that were found in the analyzed texts. Meanwhile, our algorithm permits simple addition of new letters would they be found during the future text analysis. The problem is much

simpler than with MC. Two types of rules are used, simple one-to-one mapping, and context rules.

Transliteration of 32 letters is performed under simple rules (Table 4).

Table 4. TR→MRL: one-to-one letter mapping.

TR→MRL		**TR→MRL**		**TR→MRL**		**TR→MRL**	
а	a	й	i*	т	t	ю	iu
б	b	л	l	ф	f	ѣ	ea
в	v	м	m	х	h	є	ia
д	d	н	n	ц	ţ	ѧ	â
е	e	о	o	ш	ş	ȣ	u*
ж	j	п	p	щ	şt	ĭ	i*
з	g	р	r	ь	i*	ъ	ă
и	i	с	s	э	ă	ѫ	g**

* At linguists' request, rules **й**,**ь**,**ĭ**→ **ĭ** and **ȣ**→**ŭ** may be applied.
Before **e, **i** only.

The four remaining letters are transliterated under context rules (Table 5).

Table 5. TR→MRL: Context Rules in the Order of Application.

TR→MRL		**Context**
г	**gh**	before **е**, **и**, **ю**
г	**g**	otherwise
кс	**x**	exceptions: Table 3
к	**ch**	before **е**, **и**, **ю**
к	**c**	otherwise
ч	**c**	before **е**, **и**
ч	**ce**	before **а**
ч	**ci**	otherwise
ѧ	**î***	before **m**, **n**
ѧ	**îm**	before **b**, **p**
ѧ	**în**	otherwise

*In some texts, always **ѧ**→**î** (simple rule).

3.4 Glyphs and Transliteration Rules for RC

Conform *Gramatica românească* (*The Romanian Grammar*) of 1797 by Radu Tempea, RC contains 43 letters: **Аа Бб Вв Гг Дд Єє Жж Ѕѕ Зз Ии Її Кк Лл Мм Нн Оо Пп Рр Сс Тт Ѹѹ Оуоу Фф Хх Ѡѡ Цц Чч Шш Щщ Ъъ Ыы Ьь Ѣѣ Жж Юю Ꙗꙗ Ѧѧ Θθ Ѱѱ Ѯѯ Vv Ѧ̱ѧ̱ Ц̨ц̨**.

The mid line of letters **Нн** and **Ии** in old scripts is inclined from horizontal only slightly, so both may look very like to **Нн**.

Glyphs like **й** and **ѹ** weren't treated as separate letters in RC, but as **и** and **ѹ** with diacritic sign.

Most letters (37) are transliterated under simple rules (Table 6).

Table 6. RC→MRL: one-to-one letter mapping.

RC→MRL		RC→MRL		RC→MRL		RC→MRL	
а	a	л	l	ф	f	ю	iu
б	b	м	m	х	h	ꙗ	ia
в	v	н	n	ѡ	o	θ	t
д	d	о	o	ц	ţ	ѱ	ps
е	e	п	p	ш	ş	ѯ	x
ж	j	р	r	щ	şt	v	i
ѕ	dz	с	s	ъ	ă	ѧ	g*
з	z	т	t	ы	î		
и	i	ѹ	u	ь	i		
ї	i	оу	u	ж	î		

*Before **е**, **і** only.

The remaining 6 letters need context rules (Table 7).

Table 7. RC→MRL: Context Rules in the Order of Application.

RC→MRL		Context
г	**gh**	before **е**, **и**, **ї**, **ю**
г	**g**	otherwise
кс	**x**	
к	**ch**	before **е**, **и**, **ї**, **ю**
к	**c**	otherwise
ч	**c**	before **е**, **и**, **ѣ**
ч	**ce**	before **а**
ч	**ci**	otherwise

ѣ	**e**	after ч; exception **чѣ→cea**
ѣ	**ea**	otherwise
ѧ	**a**	at the beginning of word; after **ї**, **ц**
ѧ	**e**	after **ч**
ѧ	**ea**	after another consonant; at the end of word
ѧ	**ia**	otherwise
ѧ̱	**îm**	before **b**, **p**
ѧ̱	**în**	otherwise

3.5 Examples of Transliteration

Figure 1 presents an example from "William Shakespeare Biography" book of 1849, which illustrates the complexity of the problem. In addition, this example demonstrates how useful the proposed instrument can be for specialists, especially for those who are not familiar with Cyrillic writing. Text of the 18th century is presented in Figure 2.

D. ѵititорї dе амвеле сексе.

Прiїмiцї дѵедстъ традѫкцiе а меа шi ѵiтiцї о кѫ сiнсерiтате, жѫдекжнd деспре джнса къ нѫ аре де скоп а амѫза нiчї а дикжита пе ѵine-ва, ѵi нѫмаї а моралiза. О сокотеск, дѫпе пъререа меа, ка ѫнiкъ дп фелѫл еї; къчї де шi афаръ дъ водевiлѫрї шi дестѫле комедiї ѵе сжит ла лѫмiнъ дп лiмба патрiеi, маї сжит шi оаре-каре трацедiї;' дар сжит преа сiгѫр къ din нiчї ѫна нѫ вецї пѫтеа траце маї мѫлт фолос ка dintr' аѵесте кап d'опере а ле ѵелебрѫлї Шекспiр, дитжiѫл шi неiмiтабiлѫл поет драматiк пжнъ дп секолѫл актѫал. Въ рекоmand дикъ шi ѵiтiреа вiецiї дѵестѫї ценѫ din каре пѫтецi траце ѫи фолос нѫ маї пѫцin прin моралѫл ѵе ea кѫprinde, мѫлцѫmindѫ-въ тот о датъ шi кѫрiосiтатеа прin пътрѫndереа вiртѫцiлор шi вiцiiлор ѵе карактерiсеаzъ пе Фаїмосѫл шi етернiсатѫл Енглезiлор поет. Iар еѫ дndестѫлжndѫ-мъ де зелѫл Domnii- водстре, мъ воiѫ сiлi маї мѫлт спре а въ мѫлцѫmi.

1847, Феврѫарiе 25.

Тома А. Бардат.

D. cititorї de ambele sexe.

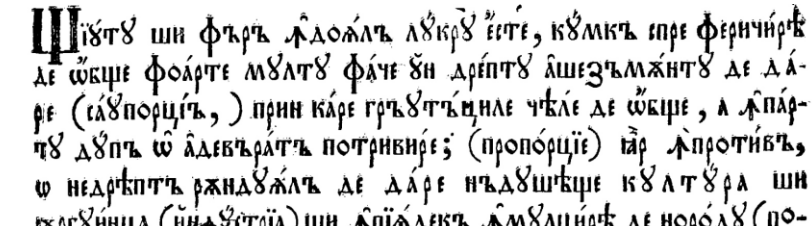
Priїmiţї această traducţie a mea şi citiţї o cu sinceritate, judecând despre dânsa că nu are de scop a amuza nicїї a încânta pe cine-va, ci numaї a moraliza. O socotesc, dupe părerea mea, ca unică în felul eї; căcї de şi afară dă vodeviluрї şi destule comediї ce sânt la lumină în limba patriei, maї sânt şi oare-care traghediї; dar sânt prea sigur că din nicї una nu veţї putea trage maї mult folos ca dintr' aceste cap d'opere a le celebruluї Ṣexpir, întâiul şi neimitabilul poet dramatic până în secolul actual. Vă recomand încă şi citirea vieţiї acestuї geniu din care puteţi trage un folos nu maї puţin prin moralul ce ea cuprinde, mulţumindu-vă tot o dată şi curiositatea prin părtrunderea virtuţilor şi viţiilor ce caracterisează pe Faïmosul şi eternisatul Englezilor poet. Iar eu îndestulându-mă de zelul Domnii- voastre, mă voiu sili maї mult spre a vă mulţumi.

1847, Fevruarie 25.

Toma A. Bagdat.

Figure 1. Translator's Introduction to the Book of 1849 (Biography of Shakespeare, Romeo and Juliette, Othello). In TR.

a)



b)

Щїѹтѹ ши фъръ ѧдоѧлъ лѹкрѹ есте, кѹмкъ спре феричирѣ де ѡбще фоарте мѹлтѹ фаче ѹн дрептѹ ашезъмжнтѹ де даре (саѹпорцїѣ,) прин каре гръѹтъциле чѣле де ѡбще, а ѧпартѹ дѹпъ ѡ адевъратъ потривире;(пропорцїе) ар ѧпротивъ, ѡ недрѣптъ ржндѹалъ де даре нъдѹшѣще кѹлтѹра ши сжргѹинца (индѹстрїа) ши ѧмпїадекъ ѧмѹлцирѣ де нородѹ (попѹлацїа).

c)

Ştiutu şi fără îndoială lucru este, cumcă spre fericirea de obşte foarte multu face un dreptu aşezămîntu de dare (sauporţiă,) prin care grăutăţile cele de obşte, a împartu după o adevărată potrivire; (proporţie) iar împrotivă, o nedreaptă rînduială de dare năduşeşte cultura şi sîrguinţa (industria) şi împiadecă înmulţirea de norodu(populaţia).

Figure 2. Political Text of the 18th Century: a) Image; b) OCR; c) Transliteration (MRL).

3.6 Transliteration utility (Cyrillic→Latin and vice versa)

Formally speaking, transliteration is a system of parametrized rules that are applied to each i-th character x_i of a Romanian word-form X in the Cyrillic script. The result $y_i = \mathrm{Trans}(x_i, \mathrm{Pos}(i, X))$ is a sequence of

characters whose concatenation produces the converted word-form Y in the Latin script.

To avoid ambiguity, the exception dictionary with foreign words, proper nouns, and difficult variants is used.

The accuracy of conversion is up to 95% for MC, up to 96% for TR, and up to 98% for RC. We should conclude that the old Romanian Cyrillic script reflected the word composition mostly accurate.

The utility is written in Java that fully supports Unicode. If the font is properly registered in the operating system, Java programming tools for the interface solve the visualization problem by a simple addressing to this font.

The transliteration utility has a user friendly interface. Files can be opened through menu or by drag-and-drop. The historical period can be selected by user or auto-detected. Supported file formats are TXT, RTF, DOC, DOCX.

The inverse transliteration (MRL→MC) is also provided as an experimental option.

3.7 Comparative Analysis of the Transliteration Process for Cyrillic Script of Different Periods

At this section we present the comparative analysis of transliteration process for historical Romanian Cyrillic scripts of different periods.

Comparing transliteration of 1830–1860 and 1945–1989 Cyrillic scripts we will mention the following important aspects. For letters that are identical in both scripting and are transliterated applying elementary rules, the process is exactly the same. There are some letters (**г**, **к ч**, **ц**) the transliteration rules for which are not so elementary, but are also identical for any Cyrillic scripting.

Transliteration of 1830–1860 Cyrillic script gives, nevertheless, better results than processing of 1945–1989 Cyrillic script. Transliteration of transitional alphabets was successful for 99% of words while for MC this fraction was 91%. TR of 1830–1860 has no problems with letters **ы** and **я**. The rule for **ы** that should be converted to **â** or **î** has some fuzziness, namely, keeping of **î** after prefixes. Transliteration of **я** from MC creates a number of ambiguous situations and strongly depends on the context. The most complicated case is the occurrence of **я** inside words. Three variants are **я→ea**, **я→ia**, **я→a**. We use some heuristically and

statistically motivated rules but most cases imply addressing external dictionaries. In 1830–1860 TR did not provoke such issues because letter **я** was not used at every phonetically suitable situation. RC contains specific letters, for example, **ѣ→еа**; **ѥ→іа**.

4 Conclusion

The proposed technology simultaneously contributes to heritage preservation, simplifies considerably its usage, extends domains and possibilities for research including humanitarian domains and enriches international communication media.

Execution of the planned works will permit unification, homogenization, and integration of national and cultural media in the international information society, and will confirm status of the Romanian language as language of communication in the European continent.

The proposed technology could be used for completion of reusable linguistic resources with new words extracted from digitized texts and attested by linguists-experts. It could also be used in creation of e-learning platforms using these texts as didactic material at learning.

It is possible to apply the described technology for another language.

The technology will automate processing of texts printed in different variants of the Romanian Cyrillic script used in 17th–20th centuries, and give unlimited access to them.

References

[1] http://www.digitisation.eu/community/map-of-the-digitisation-landscape/

[2] Bărbulescu, I. Fonetica alfabetului cirilic în textele române din vécul XVI şi XVII. Bucureşti, 1904.

[3] Картя Молдовей. Сек. XVII – ынч. сек. XX. Едиций векь. Сек. XVII – ынч. сек. XIX. Каталог женерал. Кишинэу: «Штиинца», 1990.

[4] Cartea Moldovei (sec. XVII – înc. sec. XX), Ediţii cu caractere chirilice (sec. XIX – înc. sec. XX). Catalog general. Chisinău: «Ştiinţa», 1992.

[5] Cazimir, Ş. Alfabetul de tranziţie. – Bucureşti: Humanitas, 2006.

[6] Tufiş, D.; Diaconu, L.; Barbu, A.M.; Diaconu, C., Morfologia limbii române, o resursă lingvistică reversibilă şi reutilizabilă. Limbaj şi Tehnologie, Editura Academiei Române, Bucureşti, 1996, pp. 59–65.

[7] Moruz, M.; Iftene, A.; Moruz, A.; Cristea, D. Semi-automatic alignment of old Romanian words using lexicons. In: Proceedings of the 8th International Conference „Linguistic resources and tools for processing ofthe Romanian language", Iași, Editura Universității „A.I. Cuza", 2012, p. 119–125.

[8] Karlsruher Virtueller Katalog. http://www.ubka.unikarlsruhe.de/kvk.html

[9] Corpus Cyrillo-Methodianum Helsingiense. http://www.helsinki.fi/slaavilaiset/ccmh/

[10] Vitas, D.; Krstev, C.; Obradović, I.; Popović, L.; Pavlović-Lažetić, G. An Overview of Resources and Basic Tools for the Processing of Serbian Written Texts. http://poincare.matf.bg.ac.rs/~cvetana/biblio/Solun03MATF.pdf

[11] Pavlov, R.; Bogdanova, G.; Paneva-Marinova, D.; Todorov, T.; Rangochev, K. Digital archive and multimedia library for bulgarian traditional culture and folklore. International Journal "Information Theories and Applications", Vol. **18**, Number 3, 2011, pp. 276–288.

[12] Indermühle, E.; Liwicki, M.; Bunke, H. Recognition of Handwritten Historical Documents: HMM-Adaptation vs. Writer Specific Training. www.dfki.de/~liwicki/pdf/InLiBu08-01.pdf

[13] Корниенко, С.; Айдаров, Ю.; Гагарина, Д.; Черепанов, Ф.; Ясницкий, Л. Программный комплекс для распознавания рукописных и старопечатных текстов. «Информационные Ресурсы России» №1, 2011.

[14] Resurse lingvistice reutilizabile pentru limba română. http://www.math.md/elrr/

[15] Colesnicov, A. The Romanian spelling checker ROMSP: the project overview. Computer Science Journal of Moldova, vol. **3**, nr. 1(7), 1995, pp.40–54.

[16] Boian, E.; Ciubotaru, C.; Cojocaru, S.; Colesnicov, A; Malahov, L; Petic, M. Electronic linguistic resources for historical standard Romanian. The Proceedings of the conference "Linguistic resources and tools for processing the Romanian language", 16–17 May, 2013, Iasi, pp. 35–50.

[17] Boian, E.; Ciubotaru, C.; Cojocaru, S.; Colesnicov, A; Malahov, L. Digitizarea, recunoașterea si conservarea patrimoniului cultural-istoric. Akademos, Nr. 1(32), 2014, pp. 61–68.

[18] Cojocaru, S.; Colesnicov, A.; Malahov, L.; Bumbu, T. Optical Character Recognition Applied to Romanian Printed Texts of the 18th–20th Century. Computer Science Journal of Moldova, v. 24, Nr. 1(70), 2016, p. 106–117. ISSN 1561–4042.

[19] Erjavec, T. MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora.In Proc. of the Fourth Intl. Conf. on Language Resources and Evaluation, LREC'2004, ELRA, 2004. http://nl.ijs.si/ME/Vault/CD/docs/mte-d11f/

[20] Simionescu, R. Hybrid POS Tagger. In: Proceedings of "Language Resources and Tools with Industrial Applications", Workshop Eurolan 2011 summerschool, 2011.

S. Cojocaru[1,2], L. Burtseva[1,3], C. Ciubotaru[1,4], A. Colesnicov[1,5], V. Demidova[1,6], L. Malahov[1,7], M. Petic[1,8], T. Bumbu[1,9], Ş. Ungur[1,10]

[1]Institute of Mathematics and Computer Science
5 Academiei str., MD-2028, Chişinău
Republic of Moldova
[2]E-mail: svetlana.cojocaru@math.md
[3]E-mail: luburtseva@gmail.com
[4]E-mail: constantin.ciubotaru@math.md
[5]E-mail: acolesnicov@gmx.com
[6]E-mail: valentina.demidova@math.md
[7]E-mail: lmalahov@gmail.com
[8]E-mail: petic.mircea@gmail.com
[9]E-mail: bumbutudor10@gmail.com
[10]E-mail: ungur.stefan41@gmail.com