Computer Science Journal of Moldova, vol.**25**, no.2(74), 2017

On Digitization of Romanian Cyrillic Printings of the 17th–18th Centuries

Svetlana Cojocaru Alexandru Colesnicov
Ludmila Malahov Tudor Bumbu Ştefan Ungur

Abstract

The paper describes in details recognition of Romanian texts of the 17th–18th centuries printed in the Cyrillic script, and their conversion to the modern Latin script. The challenges are discussed, and solutions of problems are proposed.

The elaborated technology and a tool pack include historical alphabets, sets of recognition patterns, and spelling dictionaries in the corresponding orthographies for ABBYY Finereader. In addition, virtual keyboards, fonts, a transliteration utility, and the user manual were developed.

This permits successful recognition of old Romanian texts in the Cyrillic script. Transliteration to the Latin script grants no-barrier access to historical documents.

1 Introduction

OCR of old books is a sophisticated task. The problems arose from peculiarities of historical typography, non-standardized spelling, and physical degradation of the documents due to their aging and usage.

This paper describes digitization of the Romanian texts of the 17th–18th centuries that were printed in the old Romanian Cyrillic script (RC). We need to OCR them, to present the recognized text in the fonts of the corresponding period, to transliterate then to the modern Romanian Latin script (MRL). Sometimes the reverse transliteration is also useful. We use the existing programs and develop our own pack of additional programs and data.

In the literature, we met the opinion that commercial software like ABBYY Finereader (AFR) is not fully suitable for OCR of old printings. It is so due to the big variability of old fonts, and because such software is internally trained with modern fonts and can't be fully retrained by users. A good introduction into the problem and further references can be found in [1, p. 2−4]. Authors of [1] prefer free "fully trainable" OCR tools like Tesseract or Ocropus.

Old Cyrillic fonts, especially of the selected epoch, are much less variable than Latin ones. The usage of the Cyrillic script is connected with the Slavonic liturgical language of the Orthodox church. For example, Orthodox Serbians use the Cyrillic alphabet while Catholic Croats speaking the same language use the Latin one. Therefore, the Cyrillic alphabet is used not so widely. Additionally, the Latin alphabet exists in blackletter and Antiqua typefaces that adds variability.

Our approach is based on the use of AFR. We produced sets of annexes to AFR containing templates for recognition collected after OCR training, alphabets, and spelling dictionaries (word lists).

We found that successful OCR supposes usage of many such sets corresponding not only to the epochs of Romanian Cyrillic printing but even to the specific typographies. It means that the said variability shows itself up but doesn't prevent obtaining good results with AFR.

It is possible that AFR work for old Cyrillic fonts better than for Latin ones due to some subtle similarity of old and modern Cyrillic fonts.

2 Digitization and its Problems

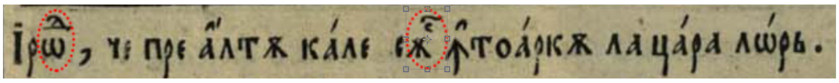
Romanian printing of the 16th, 17th and a big part of the 18th centuries used the 47-letter Romanian Cyrillic script and thoroughly imitated look and feel of the manuscripts.

We take as our study case the *New Testament of Belgrad* printed in 1648. Belgrad was a typography site; now this Romanian city is named Alba-Iulia.

This book of 682 pages is the very first edition of the full text of New Testament in the Romanian language. Till now we processed the

Gospel according to Matthew (78 pages) and continue the work.

One of the most common peculiarities of these printings is the positioning of some letters above the precedent letter (Fig. 1; a: scan; b: text in RC; c: the same in MRL; d: currently used Romanian variant; e: English, King James version). In manuscripts, this was made regularly to save place. Almost each consonant could be placed overline. At word end, the following ь was omitted; sometimes, the following vowel was omitted. This manner of writing was adopted from the Church Slavonic manuscripts [2].

a 

b ... Ірѡд, че пре алтѫ кале сѫ сѫ ꙟтоаркѫ ла цара лѡрь.
c ... Irod, ce pre altă cale să să întoarcă la țara lor.
d ... Irod, pe altă cale s-au dus în țara lor.
e ... Herod, they departed into their own country another way.

Figure 1. Overline letters in RC in 1648 (part of Matthew 2:12)

Another peculiarity is the usage of standard abbreviations for selected words like І҃с Х҃с for Іисꙋсъ Христосъ (**Jesus Christ**). These abbreviations are composed from several letters and overline signs.

The third peculiarity is denoting numbers by letters with overline signs, like є҃ for 5.

The fourth is the stress signs. They were different for vowels at the word end as in а̑фла̀ (**afla**, Eng. **learn**) and in another position as in то́ате (**toate**, **all**). A special sign was set over each vowel at the word beginning as in the same а̑фла̀; this sign replaced the stress if it was the case as in а̑лте (**alte**, **other**).

To OCR these peculiarities, we need to train these combinations as ligatures, a feature available with AFR. Fig. 2 shows several trained ligatures in the form of AFR recognition templates (patterns). See [3] for further examples of recognition patterns.

There are more peculiarities, for example, writing several words

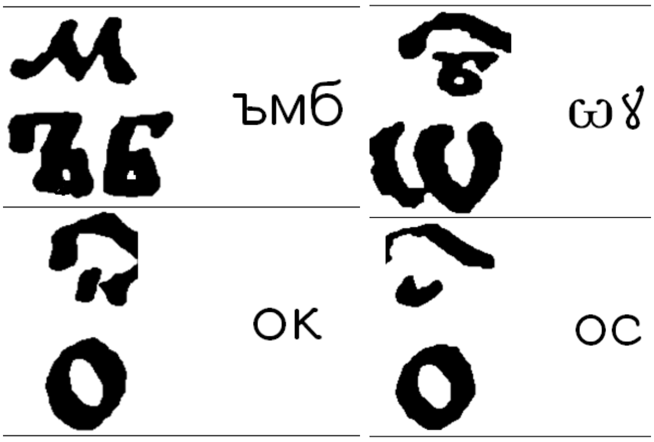


Figure 2. Examples of AFR patterns for ligatures of the 17th century

in one (mostly prepositions and articles with nouns, adjectives, and numerals), writing in lowercase only, etc.

A big problem is the absence of the spelling dictionary for the old Romanian language. It is different from the modern one, both due to the language development, and due to absence of fixed spelling rules; some words may be spelled differently in different epochs and in different places.

The importance of the proper dictionary can be illustrated by the following experiment. We took one page from a book of the late 18th century. Recognition with training but without dictionary resulted in 13% of erroneous words. Then we created a list of words from the page and repeated OCR with only 5% of erroneous words. More pages from the same book but with the dictionary restricted by this one page showed the rate of erroneous words of 8.5%.

We used a spelling dictionary created manually after OCR and partial correction of the result, of approx. 1.600 words for the 17th century at the moment.

To work with old Romanian texts more comfortably, a virtual keyboard was developed (Fig. 3).

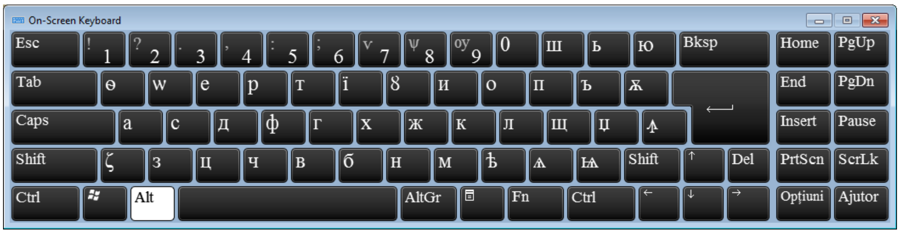


Figure 3. Virtual keyboard for RC

We implemented a tool pack collecting the tools necessary to digitize the Romanian Cyrillic printings [4]. It includes: AFR add-ons stored from AFR during training with alphabets, OCR patterns, and spelling dictionaries; the AAConv transliteration utility; virtual keyboards; a shell for selection and uploading the proper add-on into AFR; font covering rare Cyrillic glyphs; user manual.

Developing this approach, we found that the OCR works with better accuracy when we train and use separate templates not only for different epochs of Romanian historical typography but for each typography or a group of typographies that had used, presumably, fonts from the same source.

3 Transliteration to the Latin Script

Once the scanned image was processed and the editable and intelligible Cyrillic text was obtained, the transliteration process takes place.

Here is an example of recognized Romanian text (17th century):

> картѣ дєꙟтъи алꙋи Самоил ,17,стих 35. нєчє нꙋмаи сълє стрѫжꙋꙗскѫ ши сълє пъꙁѣскѫ зꙋа ши ноаптѣ,кꙋмь пъꙁїѦ Ѩковь патриꙗрхꙋлъ ѡилє лꙋи Лаван. бытїє, 31 стих 40 нєчє сълє џїє ꙟкисє ꙟстаꙋль

As we got the editable text, we can apply conversion rules to get the text in MRL. The conversion rules can be “one-to-one” like а→**a** or **т→t**, and "one-to-many". Rules may be context dependent. Examples: **ꙟ→îm** (**ꙟпърат→împărat**); **ꙟ→în** (**ꙟдоѧлъ→îndoială**); **ꙟ→î** (**ꙟнсъ→însă**). As we see, there are three different cases of conversion for letter **ꙟ**, and just a single one for **а** and **т**.

See Tab. 1 for examples of context dependent rules.

Table 1. Examples of RC→MRL rules

Cyrillic	Latin
ꙟ	**îm** before **б**, **п**
ꙟ	**î** before **м**, **н**
ꙟ	**în** all other cases
ѧ	**a** at the beginning of word; after **ї**, **ц**
ѧ	**e** after **ч**
ѧ	**ea** after another consonant; at the end of word
ѣ	**e** after **ч**; exception **чѣ→cea**
ѣ	**ea** all other cases

All these and many other rules were inplemented in the AAConv utility. Once opened or inserted in AAConv, the text is automatically converted to the modern Romanian Latin script. Below we present the previous Romanian Cyrillic text transliterated to MRL:

> **cartea deîntăi alui Samoil ,17,stih 35. nece numai săle strâjuiască şi săle păzească zua şi noaptea,cum păzea Iacov patriarhul oile lui Lavan. bîtie, 31 stih 40 nece săle ţie închise**

The transliteration utility has many settings and features accessible by the user. One of the functionalities we developed gives the possibility to convert a Cyrillic text in two different modes:

Transliteration with actualization. In addition to the conversion to the Latin script, some words and archaic letters will be changed to the modern ones to allow the text to be more understandable. At the same time it takes away some specifics of the period. For example, the old word **nece** will be replaced by its modern version **nici**.

Transliteration without actualization. The text will be converted to the MRL preserving archaic words and syntactic structures, as we saw above.

4 Reverse transliteration

AAConv gives also the possibility of backward transliteration, namely, MRL→RC. The conversion rules for the backward transliteration are “one-to-one”, “many-to-one”, and “one-to-many”, and can be context dependent.

This feature was added because there are several old books that were manually transliterated in MRL. Applying the reverse transliteration, we can use them to replenish the dictionaries (word lists) for AFR, to evaluate OCR result, and to present the text in its original form permitting to fix text authenticity.

One-to-one rules are obvious, being mostly the reversed one-to-one rules of the direct transliteration. Examples: **u**→ȣ (**lucru**→лȣкрȣ); **o**→о (**proporţie**→пропорцїе), etc.

Some one-to-one rules are context dependent. There are also rules of the kind “many-to-one”, and (rarely) “one-to-many”. Examples of such rules are presented in Tab. 2.

We present below an example of MRL→RC transliteration. We do not use overline signs in the transliteration.

> **Bine au înţeles acestea cei Crai sfinţi de demult, că nui numai aceasta deregătoria lor, să poarte grije de oamenii ce sînt sub biruinţa lor numai trupeşte, ce mai vărtos să aibă şi săsă vestească că cuvăntul lui Dumnezău întru Ei, din casele să ştie voia lui Dumnezeu şi să înţeleagă lucrul spăseniei lor.**

> Бине аȣ ꙟцелес ачестѣ чеи Краи сфинци де демȣлт, къ нȣи нȣмаи ачѣста дерегъторїа лор, съ поарте гриже де оаменїи че

Table 2. Examples of MRL→RC context dependent rules

Latin	Cyrillic
z	ꙃ if preceded by **ă**
z	ꙁ̆ if surrounded by **u** from both sides
z	ꙅ if preceded by e and succeeded by **i** or **ă**
z	з in all other cases
i	нѩ if preceded by **r** and succeded by **a**
i	ѧ if preceded by **a** is at the end of the word
i	ѩ if preceded by **a** and succeeded by **r** or **c**
i	ї if succeeded by one of the vowels **o**, **e**, **ă**
i	и in all other cases

сънт сꙋпт бирꙋинца лор нꙋмаи трꙋпѣще, че маи въртос съ аибъ ши съсъ вестѣскѫ къ кꙋвънтꙋлъ лꙋи Дꙋмнезъꙋ ꙟтрꙋ Еи, дин каселе съ щїе воїа лꙋи Дꙋмнезѣꙋ ши съ ꙟцелѣгъ лꙋкрꙋл спъсенїеи лор.

5 Conclusion

We developed a tool pack containing, in particular, OCR templates and other additions to AFR to recognize Romanian Cyrillic printings of the 17th–18th centuries. We found that better results can be achieved when we use separate OCR templates for each typography. For further processing, we developed the transliteration utilities that convert the recognized text to the Latin script and vice versa. The more ambitious task for the future is the exact electronic reproduction of look and feel of the original text with all overline letters and signs.

With our Romanian colleagues, we use the recognized historical texts and the collected dictionaries in the development of several projects: a diachronic corpus; lexicon for a POS-tagger; PROIEL (Pragmatic Resources in Old Indo-European Languages), etc.

References

[1] U.Springmann, A.Lüdeling, "OCR of historical printings with an application to building diachronic corpora: A case study using the RIDGES herbal corpus," *arXiv:* 1608.02153v1 [cs.CL], 6 August 2016. [Online]. Available: `https://arxiv.org/pdf/1608.02153v1.pdf`.

[2] O.V.Tvorogov, "On overline letters in Russian manuscripts of the 15th–17th centuries". [Online]. Available: `http://www.ruslang.ru/doc/lingistoch/1966/11-tvorogov.pdf`. (in Russian)

[3] S.Cojocaru, A.Colesnicov, L.Malahov, T.Bumbu, "Optical Character Recognition Applied to Romanian Printed Texts of the 18th–20th Century," *Computer Science Journal of Moldova*, vol. 24, no. 1(70), pp. 106–117, 2016. ISSN: 1561–4042.

[4] S.Cojocaru, L.Burtseva, C.Ciubotaru, A.Colesnicov, V.Demidova, L.Malahov, M.Petic, T.Bumbu, Ș.Ungur, "On Technology for Digitization of Romanian Historical Heritage Printed in the Cyrillic Script," in *Proceedings of the Conference on Mathematical Foundations of Informatics MFOI-2016*, (Chisinau, Republic of Moldova), 2016, pp. 160–176. ISBN: 978–9975–4237–4–8.

Received May 17, 2017

Svetlana Cojocaru, Alexandru Colesnicov,
Ludmila Malahov, Tudor Bumbu, Ștefan Ungur

Institute of Mathematics and Computer Science
Str. Academiei 5,
Chişinău, MD-2028,
Moldova
Phone: +373 22 72 59 82
E–mail: {svetlana.cojocaru,kae,mal}@math.md
{bumbutudor10,ungur.stefan41}@gmail.com