Computer Science Journal of Moldova, vol.**24**, no.1(70), 2016

Optical Character Recognition Applied to Romanian Printed Texts of the 18th–20th Century*

Svetlana Cojocaru Alexandru Colesnicov
Ludmila Malahov Tudor Bumbu

Abstract

The paper discusses Optical Character Recognition (OCR) of historical texts of the 18th–20th century in the Romanian language using the Cyrillic script.

We differ three epochs (approximately, the 18th, 19th, and 20th centuries), with different usage of the Cyrillic alphabet in Romanian and, correspondingly, different approach to OCR.

We developed historical alphabets and sets of glyphs recognition templates specific for each epoch. The dictionaries in proper alphabets and orthographies were also created. In addition, virtual keyboards, fonts, transliteration utilities, etc. were developed.

The resulting technology and toolset permit successful recognition of historical Romanian texts in the Cyrillic script. After transliteration to the modern Latin script we obtain no-barrier access to historical documents.

1 Introduction

At present Internet is the most valuable deposit of information as it can be accessed and researched from any point. New information is prepared electronically and can be exposed effortlessly. If we want to expose historical documents, we are to digitize them.

*The results published in this article were presented on November 13, 2015 at the seminar dedicated to the memory of Prof. Iu. Rogojin

Sometimes we even can access graphical images of text pages but this form effectively restricts their availability. In particular, graphical presentation makes impossible full text search.

Full text search needs textual transcription of the historical source that can be got by OCR. It was statistically showed that full-text search and quick access to contents are very important for the users, because access to the original document becomes smoother [1].

Using OCR for historical documents started in early 1990-s and progressed in parallel with the advance of OCR tools. Since 2008 big OCR projects have started, like large-scale OCR of newspaper collections in the United Kingdom and Austria [1]. Modern projects referred to in [1] are IMPACT (**Imp**roving **Ac**cess to **T**ext) under FP7, and EOD under EU Culture 2007-2013 programme.

The conversion of historical documents from the paper to accessible and searchable electronic form meets two obstacles that are not fully cleared till now.

Nowadays state-of-the-art in OCR guarantees relatively good results only on modern texts. For historical typography, results are worse. There are several causes of it. Historical fonts vary even in one book, and are less readable. Old paper introduces speckles and distortions. Linguistic components and resources of modern systems don't often know the peculiarities of historical language variations. Each text yields its own specific mix of features and problems, which implies that the quality of OCR for historical documents may vary from perfect to almost unacceptable.

The second general problem is produced by the historical orthography and language changes. Most users of digital libraries don't have a good command of old language and desire to use the modern orthography at their search. Any word can have numerous variants in the historical documents because of language evolution and lack of orthography standardization. To get satisfactory replies at search, it is necessary to skip over the gap between modern and new orthography.

In different languages, availability of texts in original historical orthography differs. For example, Romanian Cyrillic script of the 18th century has glyphs that are not supported by most OCR programs.

There are such variants in accessibility of lexical resources at the search in the historical documents. Very subtle details should be taken into account because of the alphabet evolution; for example, the Romanian language in the middle of the 19th centure used more than 17 alphabet modifications.

This situation is usual for many languages and for many cases when scientists, students, publicists, writers, statesmen, etc. want to learn from original historical documents without intermediate interpretations. Therefore, national systems for no-barrier access to historical documents are necessary, being supported by historical lexical resources, proper OCR tools and tools for quick interpretation of new unknown texts. Such systems should become available for interested users of these cultural data.

The OCR of manuscripts is a specific challenge, and we will not discuss it here.

In the paper, we would discuss the factors defining the reliability of the OCR result, and the techniques permitting to enhance it by the example of printed historical Romanian texts of the 18th–20th century in the Cyrillic script. The following epochs were preliminary distinguished in the Cyrillic scripts for Romanian, using the principle "since the present and back centuries" (see details in [2]):

Epoch 1: the 2nd half of the 20th century, Moldavian SSR, Russian-based Cyrillic script.

Epoch 2: 1830–1860, the so-called transitional alphabets, mix of Romanian Cyrillic and Latin script.

Epoch 3: the early 19th century and back, the Romanian Cyrillic script.

For epoch 1, the problem seems to be almost solved, and we shortly discuss our achievements in Sec. 4. We concentrate our discussion mainly on the 2nd epoch (Sec. 5). We research also epoch 3; our results are presented in Sec. 6.

2 Production process

The following four stages form the process of producing the textual transcription of a printed historical document.

1. Digitization (scan) and image preprocessing.
2. OCR.
3. Text post-processing.
4. Quality evaluation.

For **scanning,** we recommend specialized book scanners, for example, Plustek OpticBook [3], and scan with at least 600 DPI resolution. The worst case is when we get already scanned source from some collection and cannot regulate its properties.

There are several freely available programs for **image preprocessing** like line straightening, image cleaning, converting to black-and-white. One of such programs is ScanTailor. A big collection of such tools is presented at [4]. In particular, Agora is an interesting tool that analyses blocks of text and images on pages.

OCR is a most complicated and error-prone stage. We tested several OCR systems and selected ABBYY FineReader (AFR) [2],[5]. The latest AFR versions include some image preprocessing but we recommend separate tools as more powerful and versatile. The OCR program performs segmentation of image to characters, and produces text comparing characters with patterns. Then the dictionaries for supposed languages can be used to check the spelling of resulted text and correct it. Training mode can be proposed when the user manually corrects text segmentation to glyphs and pattern-to-glyph mapping.

Post-processing of the text mainly includes manual correction of the OCR outcome, and extracting words to replenish the dictionary used at OCR. AFR permits some manual corrections in its output window before storing the resulting text. Allocation of textual blocks may also need correction, depending on purpose. For example, it is not necessary for full text search. The post-processing may continue up to full restoration of physical text appearance.

Quality assurance also depends on purpose of text processing. It can be done at several levels: the scan and dataset level, institutional

and the project consortium level. It is recommended to perform thorough post-evaluation and error spotting over the first produced text samples to ensure consistency in further production.

3 Factors affecting scan and OCR quality

The recognition quality depends on: the scan quality; the alphabet selection; the OCR engine training over specific texts; the availability of dictionary corresponding to the proper historical period. In its turn, scan quality is influenced by factors like: black-letter typefaces; irregular spacing between letters and words; changing font sizes; poor paper; inconsistent inking; speckles; distortion and other geometric deformations of text, non-straight lines; text strike-through. In the worst case, these may imply the manual correction of each page image, e.g., despeckling.

The case of color and negative (white letters on black or dark background) printing is also very difficult. AFR splits the image of each page to blocks that can be attributed as text, table, or image. This splitting is not always perfect; the manual correction may be necessary.

OCR quality may be affected by: alphabet diversity; mix of scripts; use of special characters, digraphs, ligatures; use of accents; use of historical vocabulary; poor vocabulary recognition.

The task of dictionary creation seems to be a true vicious circle as it supposes studying a lot of potential hardly accessible sources, and extracting data through language and script barriers.

4 Recognition of Moldavian Cyrillic script

Moldavian Cyrillic script was used in Moldavian ASSR and Moldavian SSR. It was based on the Russian alphabet with one additional letter "ӂ". The typography of that period permits to obtain good scans. The dictionary was produced from recognized books themselves using manual correction of words; it can be expanded from new books. Details are discussed in [2], [5]. The purpose of recognition was mainly

re-editing valuable books in modern Latin script; for this purpose, a transliteration utility was developed [5].

5 Recognition of transitional alphabets

Transitional alphabets were used in the Romanian typography since 1830 and until 1860–1870 [6]. They can be characterized by regular many-to-one mapping of old Romanian Cyrillic letters to the mix of Latin and Cyrillic letters. This mapping could be expanded further to modern Latin Romanian script; slightly different orthography makes an obstacle. The existence of such mapping distinguishes the old Romanian Cyrillic and transitional scripts from Moldavian Cyrillic script that cannot be ([5]) regularly mapped to the modern Latin script.

There were many different transitional scripts. Our impression is that different typographers used them depending on the existed stock of letterpunches, progressively replacing the Cyrillic letterpunches with the Latin ones whenever the former were worn. We can see different alphabets at the same year. Book [6, p. 115] shows a "record" example of 1840 where the title page was printed in four different scripts simultaneously (old Cyrillic, simplified Cyrillic, transitional, and Latin). Sources count up to 17 variants of the transitional scripts. This diversity makes a main problem at OCR of these documents.

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 (Fig. 1, p. 112). It produces 7% of erroneous words.

The second approach was tested to solve the problem of alphabet variation. We rejected the principle of the exact text reconstruction after OCR. AFR permits to output the result in original glyphs or substitute them by any sequence of letters from the selected alphabet of recognition. This is called "ligatures" in AFR documentation. For AFR output, we invented a Latinized version of the 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".

(a)

ЛȢЇ ШЕКСПІР. 5

алції 'л аȣ апробат кȣ ꙟтфокаре, фъръ ꙟndoiалъ,
пентрȣ къ bedea ꙟнтр'ꙟнсȣл ȣn прілеж d'a аръта
сімціментȣл поетік ал ероȣлȣї лор, поате ꙟнкъ
de кꙟnd се афла ꙟn леагън. Къчї Аȣбреі історі-
сеще къ тꙟпърȣл Віліam, непȣтꙟnd съ се сȣпȣе
кȣ плъчере daторііlор крȣнтȣлȣї съȣ стат, къȣта
а'л ꙟпълца пріn експресііле ȣnȣї новіл сімцімꙟnт,
пронȣnцꙟnd кꙟте ȣn помпос dіскȣрс de кꙟте орї
тъiа вре о вітъ.

(b)

Л8Ї ШЕКСПІР 5

алції 'л а8 апробат к8 ⩚mфокаре, фъръ ⩚ndoiалъ,
пентр8 къ bedea ⩚нтр-⩚нс8л 8n прілеж d'a аръта
сімцімꙟнт8л поетік ал еро8л8ї лор, поате ⩚нкъ
de кꙟnd се афла ⩚n леагън. Къчі А8бреі історі-
сеще къ тꙟпър8л Віліam, неп8тꙟnd съ се с8бп8с
к8 плъчере daторііlор кр8нт8л8ї съі стат, къ8та
а'л ⩚пълца пріn експресііле 8n8ї новіл сімцімꙟnт,
прон8nцꙟnd кꙟте 8n помпос dіск8рс de кꙟте орї
тъiа вре о вітъ.

Figure 1. Romanian transitional script (1848) after OCR: (a) source; (b) text

Because of one-to-one letter mapping, the exact reconstruction of the text from a book is achieved applying a simple letter substitution selecting the desired variant of the transitional alphabet. We are developing the corresponding conversion utility.

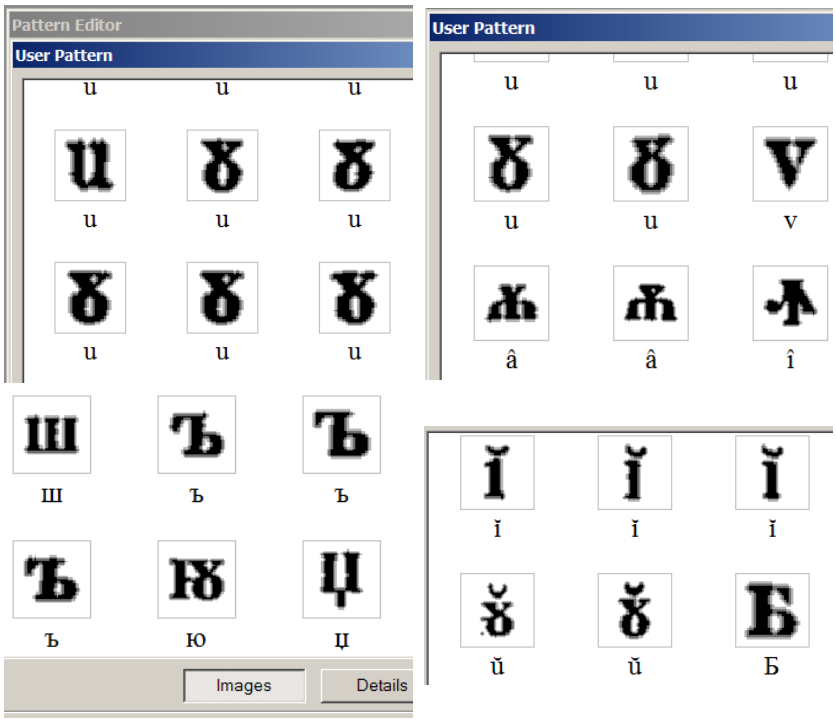


Figure 2. Part of AFR pattern collection for Romanian transitional alphabets with substitutions ("ligatures")

This approach also reduces drastically the volume of the dictionary. For example, "trekut" ("past", modern Latin script "trecut") in the recognition dictionary may check up to 16 variants obtaining by independently replacing t→т, r→p, k→к, u→ŭ).

This restriction of the recognition alphabet solves one small problem of interaction with AFR. AFR does not support arbitrary Unicode

glyphs in its dialogs and forms. Old Romanian letter "↑" was introduced in Unicode only after 2009. Standard system fonts do not contain some Romanian Cyrillic (and transitional) letters. As a result, we see in AFR empty boxes "□" instead of letters during training, alphabet formation, etc.

Work with ligatures also reduces errors to 4.8% (word level; see Sec. 6).

After training, we collected a set of glyph patterns for Romanian transitional alphabets. Part of this collection is shown in Fig. 2, p. 113.

Resuming, the OCR of Romanian transitional script should be performed as follows. Configure AFR with the corresponding "user language". Set the alphabet for this language from the corresponding string. Fill the recognition dictionary from the corresponding file. In the pattern editor of AFR, download recognition patterns from the corresponding file. After recognition, apply the utility and remap the result to the necessary variant of the transitional alphabet to restore the original glyphs.

You can also use the AFR output (before its remapping) to replenish the recognition dictionary. The recognition quality grows as the dictionary grows. We repeated recognition several times using the recognized text as new words source, with manual checking of the included words because of the absence of the historical lexicons.

6 Recognition of old Romanian Cyrillic script

AFR recognizes old Romanian Cyrillic Script. Small problems arose due to absence of necessary glyphs in system fonts, as it was already noted above. In fact, only three fonts in the whole world have old Romanian Cyrillic letters: Kliment, Unifont (bitmap font), and Everson Mono [7]–[9].

For example, the juridical text from 1786 was recognized with engine training and user supplied dictionary (Fig. 3, p. 117). This results in 4.5% errors (word level) with original glyphs and 3% errors with ligatures. We observed this effect with transitional scripts also.

This unexpected result is to be explained. The most likely reason

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

7 Conclusions

Digitization of historical texts includes their scanning and recognition; the latter was performed by ABBYY FineReader 12.

To use OCR for the Romanian Cyrillic script, we developed a set of historical alphabets and sets of glyphs templates, which are specific for each epoch. The spelling dictionaries in proper alphabets and orthographies were also created. Some auxiliary supporting tools like virtual keyboards, fonts, transliteration utilities, etc. were also developed.

Images were preprocessed with specific pre-OCR tools.

We have analyzed two approaches to recognition: using authentic glyphs, and using glyph substitution. The second approach solves the problem of diversity for transitional alphabet, and, due to some peculiarities of the AFR training mode, produces fewer errors.

OCR can dramatically increase the usability of digital libraries. The proposed solutions of the problems discussed in the paper can significantly impair the quality of the OCR outcomes. With it, full-text search and no-barrier access to digitized historical documents become possible.

References

[1] Historical Lexicon of Slovene:
Available: `http://www.digitization.eu/tools-resources/language-resources/1322-2/`

[2] E.Boian, C.Ciubotaru, S.Cojocaru, A.Colesnicov, L.Malahov. Digitization, recognition and conservation of the cultural and historical heritage. Academos, Nr. 1(32), 2014, pp. 61–68. ISSN 1857–0461. (In Romanian)

[3] Available: http://plustek.com/uk/products/opticbook-series/

[4] Available: http://www.digitisation.eu/tools-resources/tools-for-text-digitisation/

[5] C.Ciubotaru, S.Cojocaru, A.Colesnicov, V.Demidov, L.Malahov. Regeneration of Cultural Heritage: Problems Related to Moldavian Cyrillic Alphabet. Presented at The 11th International Conference "Linguistic Resources and Tools for Processing the Romanian Language". 26–27 November 2015. Eds: D.Gîfu, D.Trandabăţ, D.Cristea, D.Tufiş. pp. 177–184. ISSN 1843–911X. Available: http://consilr.info.uaic.ro/2015/Consilr_2015.pdf.

[6] S.Cazimir. The transitional alphabet. – Bucharest: Humanitas, 2006. ISSN 973–50–1401–7. (In Romanian)

[7] Available: http://kodeks.uni-bamberg.de/aksl/schrift/KlimentStd.htm

[8] Available: http://www.unifoundry.com/unifont.html

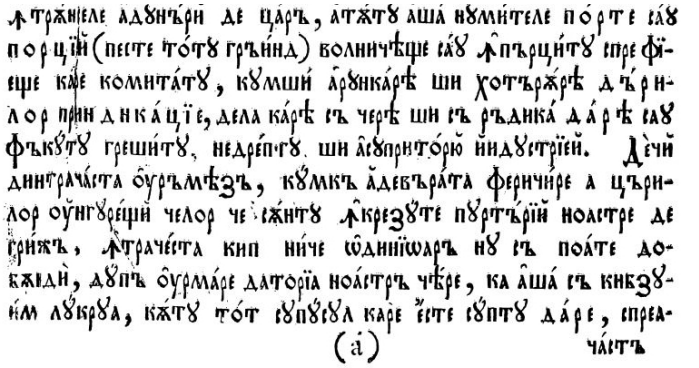
[9] Available: http://www.evertype.com/emono/

Svetlana Cojocaru, Alexandru Colesnicov,
Ludmila Malahov, Tudor Bumbu

Received March 14, 2016

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@gmail.com

(a)



(b)

Дечи диитрачаста оурѣмѣзъ, кѹмкъ адевърата феричире а църилор оугѹреции челор че сѫнтѹ ѧкрезѹте пѹртърїи ноастре де грижъ, ѧтрачеста кип ниче ѡдинїѡаръ нѹ съ поате добѫнди, дѹпъ оурмаре даторїа ноастръ чѣре , ка аша съ кнбзѹим лѹкрѹл, кѫтѹ тот сѹпѹсѹл каре есте сѹптѹ даре, спреа

(a) частъ

(c)

Дечи диитрачаста оурѣмѣзъ, кѹмкъ адевърата феричире а църилор оунгѹре^^^ челор че сѫнтѹ їкрезѹте пѹртърїи ноастре де грижъ, їтрачеста кип ниче ѡдинїѡаръ нѹ съ поате добѫнди , дѹпъ оурмаре даторїа ноастръ чѣре , ка аша съ кибзѹим лѹкрѹл, кѫтѹ тот сѹпѹсѹл каре есте сѹптѹ даре, спреа-

(a) частъ

Figure 3. Recognition of the juridical text of the 18th century in the Romanian Cyrillic script: (a) source; (b) original glyphs; (c) "ligatures"