Revitalizing the historical Romanian texts with Cyrillic Scripts

Olesea Caftanatov and **Ludmila Malahov** and **Tudor Bumbu**
Moldova State University, Vladimir Andrunachievici Institute of Mathematics and Computer Science
olesea.caftanatov@math.md,ludmila.malahov@math.md,tudor.bumbu@math.md

Relevant UniDive working groups: WG2, WG3

1 Introduction

The aim of our work is revitalizing the historical Romanian texts with Cyrillic Scripts from the XVII – XX century. The revitalization process consists of a few main steps for instance: scanning, recognition, and transliteration from Cyrillic Scripts into Latin one. We researched various types of historical documents such as: manuscripts, religions books, dialectal text and others. The endeavor to address the linguistic heritage of Romanian history involves tackling several specific challenges, including: (1) dealing with a multitude of language evolution periods; (2) coping with the scarcity of widely available resources; (3) managing the diverse array of alphabets used in historical printings, including mixed Cyrillic-Latin “transition alphabets”; (4) overcoming the absence of reliable tools for accurately recognizing Cyrillic letters from various historical eras; (5) addressing the shortage of lexicons suitable for the time periods of these resources.

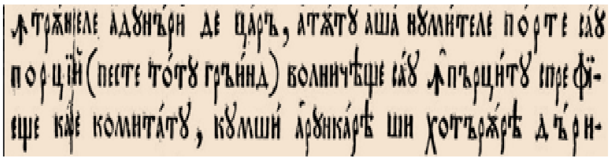
To surmount these obstacles, we have developed a comprehensive platform that seamlessly integrates a suite of software components for image processing, text recognition, and transliteration into contemporary Latin script. This platform has been finely tuned to handle text recognition and transliteration across different historical epochs and to accommodate the variations in alphabets used in Romanian language printings in both Romania and the Republic of Moldova.

The latest resource we collaborated on pertains to the dialectal texts published in Moldova in the years ‘60–‘80 of the past century are of particular importance for research in different fields of science like: linguistics (especially dialectal, the history of the Romanian language, stylistics), history, ethnography, folklore studies, sociology, psychology, ethnolinguistics, sociolinguistics, psycholinguistics.

2 Recognition process

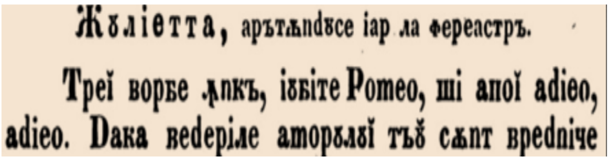
Considering our reliance on books as our primary resources, we found it imperative to employ an Optical Character Recognition (OCR) tool for the conversion of image text into editable content. To achieve this, we utilized Abbyy FineReader.

The OCR process for old books presented a complex challenge, stemming from the idiosyncrasies of historical typography, non-standardized spelling, and the physical deterioration of documents due to aging and use. To address this, we specifically targeted three epochs (roughly the 18th, 19th, and 20th centuries- see Figure 1-3), each characterized by distinct usage of the Cyrillic alphabet in Romanian and, consequently, necessitating a tailored approach to OCR.



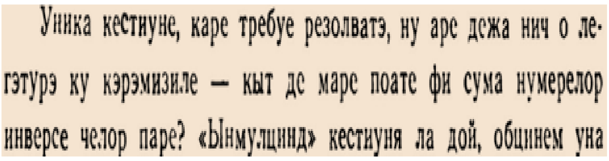
ꙟтрѫніеле адѹнъри де царъ, атѫтѹ аша нѹмітеле порте саѹ
порцїй (песте тотѹ гръннд) волнічеще саѹ ꙟпърцітѹ спре фі-
єще каре комітатѹ, кѹмші арѹнкарѣ ші хотърѫрѣ дъри-

Figure 1: XVIII century



Жѹліетта, аръткndѹсе іар ла фереастръ.
Треї ворбе ꙟнкъ, іѹбіте Ромео, ші апої adieo,
adieo. Dака вedepiле аморѹлѹї тъѹ сѫнт вpednіче

Figure 2: XIX century, mixed alphabet



Уника кестиуне, каре требуе резолватэ, ну арс дежа нич о ле-
гэтурэ ку кэрэмизиле — кыт дс марс поатс фи сума нумерелор
инверсе челор паре? «Ынмулцинд» кестиуня ла дой, обцинем уна

Figure 3: XX century

The technology supports virtual keyboards, fonts, a transliteration tool, and other possibilities. We created additional Python programs, for example, a tool to select the historical period and the geographical region, where the text was printed

(Iasi, Bucharest, Târgovişte, Bălgrad, Uniev, Sas Sebeş, Snagov, Buzău etc) [1-2]. In particular, for Bucharest, the system is trained in recognizing the fonts from the Royal Typography and these of the Bucharest Metropolitan Chair.

We meticulously developed historical alphabets and sets of glyph recognition templates unique to each epoch, accompanied by dictionaries featuring proper alphabets and orthographies. Additionally, we crafted virtual keyboards, fonts, transliteration utilities, and other tools. This comprehensive technology and toolset enabled the successful recognition of historical Romanian texts in the Cyrillic script [1].

Furthermore, we designed dictionaries for verifying the spelling of old Romanian texts spanning the 17th to 20th centuries. This technology was actively applied in the revitalization of cultural heritage embedded in printed historical documents. Examples include "Numbers and Ideals" by V. Andrunakievich, I. Kitoroaga (1969), the New Testament (1646), Amfilohie Hotiniul's "De obste gegrafie" (1795), dialect texts (vol. 1, p. 1, 1969 published), mathematical books, and more.

3 Transliteration process

Regarding the transliteration process we developed rules in collaboration with linguistics researchers and we created the transliteration programs using Python. Here is an example of recognized Romanian text (17th century):

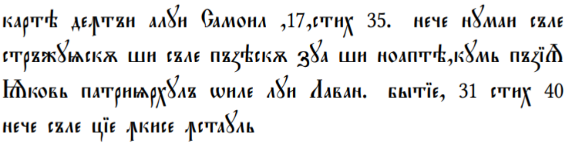

картѣ дєꙟтъи алꙋи Самоил ,17,стих 35. нєчє нꙋмаи сълє стръжꙋѩскѫ ши сълє пъꙁѣскѫ ꙁꙋа ши ноаптѣ,кꙋмь пъꙁїѧ Ѩковь патриѩрхꙋлъ ѡилє лꙋи Лаван. бытїє, 31 стих 40 нєчє сълє цїє ꙟкисє ꙟстаꙋль

Figure 4: Fragment of recognized text

As we got the editable text, we can apply conversion rules to get the text in modern Romanian Latin script (MLR). The conversion rules can be“one-to-one” and “one-to-many”. Rules may be context dependent. Below we present the previous Romanian Cyrillic text transliterated to MRL:

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 (neither). The text will be converted to the MRL preserving archaic word and syntactic structures, as we saw above. Transliteration of the dialectal text is challenging research because of phonetic script. The Cyrillic alphabet has approx. 80 letters with or without diacritics, 16 diacritics. Also, it has separators or even bonds. Thus, after recognition we got approx. 100 characters for transliteration. The character's position is very important, as a result we created 273 rules for transliteration. Our transliteration program runs 3 cycles depending on the context.

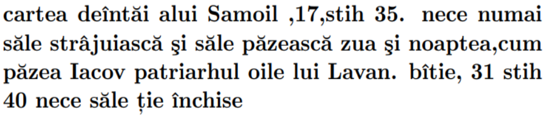

cartea deîntăi alui Samoil ,17,stih 35. nece numai săle strâjuiască şi săle păzească zua şi noaptea,cum păzea Iacov patriarhul oile lui Lavan. bîtie, 31 stih 40 nece săle ţie închise

Figure 5: Fragment of transliterated text

4 The Digitization Platform

We developed a Digitization Platform (DP) for the digitization process of Romanian Cyrillic historical documents. DP is a web application that features an interactive graphical interface and a set of APIs designed and managed through Django Rest Framework. This application offers seven consecutive steps to enable the recognition of Romanian Cyrillic documents from the XVII-XX centuries, transliteration of the texts into Latin script, editing of recognized/transliterated texts, and downloading or publishing the results [5].

The platform includes: (i)image processing engines, such as ScanTailor , AFR, and OpenCV; (ii)OCR models for Cyrillic characters, trained on datasets gathered from documents printed in the XVII-XX centuries; (iii)a transliteration tool from old Romanian Cyrillic to contemporary Latin script; (iv)virtual keyboards specific to the alphabets used in previously mentioned periods, such as the Romanian Cyrillic alphabet, the transition alphabet, and the Soviet Cyrillic alphabet. The DP architecture (see Figure 6) consists of a set of modules that are organized into 4 functional groups.

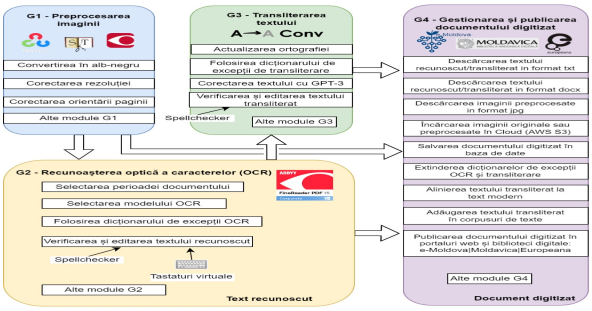

G1 - Preprocesarea imaginii
OpenCV
S T
Convertirea în alb-negru
Corectarea rezoluției
Corectarea orientării paginii
Alte module G1
G3 - Transliterarea textului
A→A Conv
Actualizarea ortografiei
Folosirea dicționarului de excepții de transliterare
Corectarea textului cu GPT-3
Verificarea și editarea textului transliterat
Spellchecker
Alte module G3
G2 - Recunoașterea optică a caracterelor (OCR)
ABBYY
FineReader PDF 15
Corporate
Selectarea perioadei documentului
Selectarea modelului OCR
Folosirea dicționarului de excepții OCR
Verificarea și editarea textului recunoscut
Spellchecker
Tastaturi virtuale
Alte module G2
Text recunoscut
G4 - Gestionarea și publicarea documentului digitizat
e-Moldova
MOLDAVICA
BIBLIOTECA NAȚIONALĂ DIGITALĂ
europeana
Descărcarea textului recunoscut/transliterat in format txt
Descărcarea textului recunoscut/transliterat in format docx
Descărcarea imaginii preprocesate in format jpg
Încărcarea imaginii originale sau preprocesate în Cloud (AWS S3)
Salvarea documentului digitizat în baza de date
Extinderea dicționarelor de excepții OCR și transliterare
Alinierea textului transliterat la text modern
Adăugarea textului transliterat în corpusuri de texte
Publicarea documentului digitizat în portaluri web și biblioteci digitale: e-Moldova|Moldavica|Europeana
Alte module G4
Document digitizat

Figure 6: The architecture of Digitization Platform (DP)

5 LEXICOGRAPHICAL DIACHRONIC ANALYSIS

Diachronic (across time) linguistics also known as historical linguistics investigates and describes the way in which languages change or maintain their structure over time. Hence, our goal is on the lexicographical diachronic analysis aspect of the language. The application supposes digitization of huge collections of historic texts and creation of the corresponding corpora. We selected approx. 15,000 words from folkloric book [3] and 7,000 words from native lyrical songs to experiment with diachronic analysis. The folkloric book was printed using Moldavian Cyrillic Script, so we had to recognize and transliterate it into Latin Script with the tools described above. The next step was to verify the orthography and to correct the mistaken characters. This process was manual and it took a long time. In the next step we got rid of stop words by using a dedicated java program with a list of stop words manually selected from text. Some examples from the stop word list are: “sunt” ((are (III, plural and I, singular)) very common verb), “de, ın, pe, de pe” (prepositions), “si, ca, sa, ci, dar, de, fie, daca, ori” (conjunctions), etc. After this step we obtained a clean text which was involved in the next processes.

In the next step we tried to extract the words in common and for this purpose we used latent Dirichlet allocation (LDA) which is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. We used a version of this model implemented in the Python framework, namely GraphLab [4]. In the configuration of the model, we set only to select words from 1 topic and to iterate 200 times (optimal iterations for our data, we experimented also with 300, 400 and 500 iterations with bad results). The final step was to measure the difference between terms from two different centuries and for this objective we used Levenshtein distance which is a string metric for measuring the difference between two sequences. Also, for the optimal approximation of the distance, Munkres algorithm was implemented in Python.

6 Conclusion

We developed a tool pack containing, in particular, OCR templates and other additions to AFR to recognize Romanian Cyrillic printings of the 17th–20th 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. We recognized and transliterated over 1000 pages of dialectal texts. The first volume of the dialectal book was prepared for publication, the other two books are used as resources for linguistic dialectal corpora. 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] S. Cojocaru, A. Colesnicov, L. Malahov, Digitization of Old Romanian Texts Printed in the Cyrillic Script, Second International Conference on Digital Access to Textual Cultural Heritage DATeCH-2017, Goettingen, June 1-2, 2017, 143–148.

[2] Bumbu T., Towards a Font Classification Model for Romanian Cyrillic Documents. In: Computer Science Journal of Moldova. 2021, nr. 3(87), pp. 291-298. ISSN 1561-4042.

[3] G. G. Botezatu, H.M. Baeu, E.V. Junghientu, M.G. Savina, E.V. Tolstenco, A.S. Hıncu,V.A. Cirimpei i I.D. Ciobanu, Folclor din prile Codrilor, Academia de Stiinte a RSS Moldovenesti, 1967.

[4] Joseph Gonzalez, Yucheng Low, Haijie Gu, Danny Bickson, Carlos Guestrin, PowerGraph: Distributed Graph-Parallel Computation on Natural Graphs, Proceedings of Operating Systems Design and Implementation (OSDI), 2012.

[5] Bumbu T. Tehnologii și resurse informaționale pentru digitizarea și procesarea textelor din patrimoniul istorico-cultural. Teza de doctor în informatică, 2023, State University of Moldova, pp. 95-118.