Medieval text processing: use case of Romanian uncial writing

Tudor Bumbu, Liudmila Burtseva, Svetlana Cojocaru, Alexandru Colesnicov, Ludmila Malahov

Information technologies have given rise to a new trend: the development of digital cultural heritage. The authors have persevered in the digitization and transliteration of Romanian texts printed in Cyrillic characters from the 17th to the 20th centuries, and their efforts have been crowned with the creation of the HeDy platform for processing historical texts [1].

This work aims to deepen the object of study over time by focusing on medieval manuscript texts written with uncial Cyrillic script. Documents with this type of writing are available between the thirteenth and the half of the nineteenth century in considerable numbers. Still, the possibility of studying them is severely limited by the lack of digital copies and the need to transliterate them into modern Latin script.

Even more difficulties for the contemporary reader arise for the following reasons:

- The mode of writing differs (sometimes substantially) from one person to another, presenting a much greater variety than in old printed texts. Moreover, even texts written by the same person can be different.
- Vocabulary and spelling are also different, with quite varied spellings of the same word.
- Along with the diversity of spellings, errors in copying are quite frequent, with faulty copies often serving as a source for their proliferation in the process of new copying.
- Handwritten text editing was done by writing over the line, writing in the margins, erasures, and writing on the erased place, which often led to damage to the medium.
- The massive use of abbreviations, often non-standardized, due to the desire to reduce laborious efforts, and save time and space.
- *Scripta continua*: continuous writing without spaces between words.

Of course, the list can be extended, as certain aspects are highlighted in the working process.

Despite these difficulties, several works demonstrate a good rate of accuracy in recognizing these texts (including the separation of scripta continua into words), with methods based on neural networks applied to Slavic ecclesiastical texts, as well as Latin and French ecclesiastical or secular texts [2, 3].

Approaches to recognition of medieval manuscript texts

As we are working with old Romanian texts that are not publicly available, we had to start by creating a collection of documents for testing software and technologies for word processing with uncial and semi-uncial writing.

Thanks to the collaboration with colleagues from the Institute of Theoretical Informatics (IIT), the branch of the Romanian Academy, Iași, several resources of texts with uncial writing have been identified, which serve as a basis for training the ABBY Fine Reader system. These resources were offered to us by the creators of the Deep Learning for Old Romanian

(DeLORo)[4] project, carried out at IIT, Romanian Academy, Iași Branch, a project in which the Romanian Academy Library in Bucharest was a partner and we take this opportunity to thank them.

We also got acquainted with the works describing systems for accessing historical handwritten documents. They can be classified as follows: automatic recognition systems based on neural networks with preliminary machine learning [5]; systems using the ABBY Fine Reader platform [6]; systems using other recognizers, for example, Tesseract; the Library of Congress project, which uses volunteers for virtual processing of historical manuscripts [7]. Volunteers create and review transcriptions virtually to improve search, access, and discovery of these pages from history.

Some attempts to interpret old Romanian manuscripts of the 19th century were made using the first version of the Transkribus platform [8]. However, the accuracy of the recognition was not good enough.

This article will discuss how the ABBY Fine Reader platform (versions 15,16) was used to recognize manuscripts with Romanian Cyrillic uncial writing. Two OCR (AI) models were created and trained for the recognition of Romanian uncial and semiuncial texts: *Floarea darurilor, Codicele Bratul, and Codex Sturdzanus.* The performance evaluation of a specific OCR model for Romanian uncial and semiuncial texts was developed, obtaining an accuracy of approximately 75-85%.

1. Manuscript „Floarea darurilor" recognition

Floarea darurilor is a popular book, with a moralizing character, which has enjoyed wide circulation in the Romanian cultural space. It was published in 1491 in Florence in Italian, under the title *Fiore di virtù*. Until 1701, the Romanian translation of the book was distributed in manuscripts [9].

The text is written in uncial script using superscripts without separating words and sentences, without dividing into paragraphs. The text is parallel in two languages, Church Slavonic and Romanian. The Church Slavonic text is written in black, and the Romanian text in red. Headings and initials are also written in red. An example of the original page is shown in Figure 1.

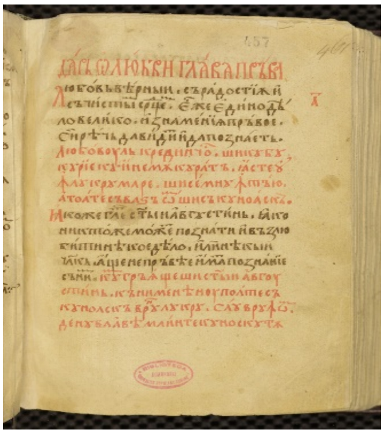


Fig.1.

Text recognition was performed by FineReader PDF v.16. The recognition model consisted of a user language and a dictionary. The alphabet of the language is presented in Figure 2. (To present it as an image. Is the first character “y” in the correct place?)

ɣабвгдежзийклмнопрстуфхцчш
щъыьюsiïwѣѧѫѯѱѳѵꙁꙃш̄ȥɤꙗꙛ

Fig.2

A small dictionary was created manually from the text in several iterations.

Difficulties are created by the superscripts (signs and letters), overlapping lines, and, to a lesser extent, overlapping the borders of letter boxes. Direct recognition without image editing did not give adequate results. Therefore, the image was preprocessed.
A feature of the existing texts was established: the text has superscripts, but no subscripts. This made it possible to formulate an algorithm for eliminating line overlap: descending from the reference line of the line, go to the background color and fix there the dividing line under the line. The resulting dividing line is stretched vertically. The side effect is that the initials are processed correctly, falling into their line. A string splitting program is currently being developed, and for test purposes, pages have been divided into rows manually (Fig. 3).

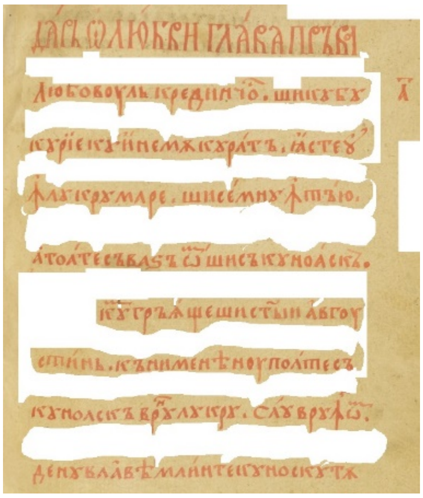


Fig.3

Text processing with extended lines was carried out in the FineReader training mode. Errors were corrected manually (Fig. 4). Also, word division was made by hand and superscript letters were ~~introduced~~ inserted into the text (Fig. 5). The latter requires a special, more complex solution. The training set consisted of about 2000 characters. The test set included about 300 characters. Recognition accuracy without dividing into words is 85%.

дарълюбвиглавапръва
любовоулькрединчос.шикɣбɣ
кɣріекɣинемѫкɣратъ.ꙗстеɣ
ꙛлɣкрɣмаре.шисемнɣꙛтъю.
атоатесъвasъwмшисъкɣноаскъ.
кɣмгръѧщешистыиавгоу

даръ любви глава пръва
любовоуль крединчос .
ши кɣ бɣкɣріе кɣ инемѫ
кɣратъ . ꙗсте ɣꙛ лɣкрɣ
маре . ши семнɣ ꙛтъю .
а тоате съ ваsъ wм ши

стинь.кънименѣноупоатесъ
кҮноаскъврҮнлҮкрҮ.саҮврҮѧwм
денҮваааавѣмаинтекҮноскҮтѫ

Fig. 4

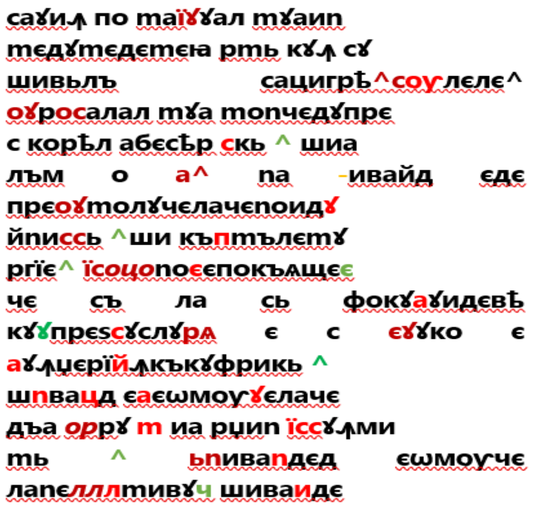
съ кҮноаскъ . кҮм
гръѧще ши стыи
авгоустинь . къ нименѣ
ноу поате съ кҮноаскъ
врҮн лҮкрҮ . саҮ врҮѧ
wм де нҮ ва авѣ маинте
кҮноскҮтѫ

Fig. 5

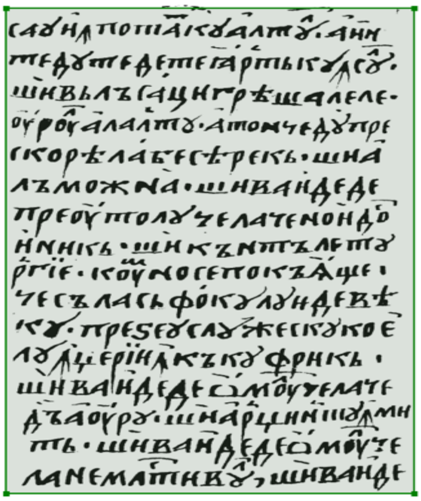
2. Manuscript „Codex Sturdzanus” recognition

Codex Sturdzanus is a collection of ancient texts, parabiblical manuscripts (apocryphal, hagiographic, and apocalyptic legends), from the sixteenth century copied between 1580 and 1619 by the priest Grigore from Măhaci.

Text recognition was performed by FineReader v.15. The character-level accuracy is just over 85% after training the 15-page model. We continue to work until the 20-page training to evaluate progress. The result of recognition is shown below in Fig. 6.



саѷиѧ по maïѷѷал mѷаип
mєдѷmєдєmєѩ pmь кѷѧ сѷ
шивьлъ сацигрѣ^соуלєлє^
оѷросалал mѷа monчєдѷпрє
с корѣл абєсѣр скь ^ шиа
лъм о а^ па -ивайд єдє
прєоѷmолѷчєлачєпоидѷ
йписсь ^ши къпmълєmѷ
pгïє^ ïсоцопоєєпокъѧщєє
чє съ ла сь фокѷаѷидєвѣ
кѷѷпрєsсѷслѷрѧ є с єѷѷко є
аѷѧцєрïйѧкъкѷфрикь ^
шпвацд єаєωмоуѷєлачє
дъа орpѷ m иа pцип ïссѷѧми
mь ^ ьпивапдєд єωмоучє
лапєлллmивѷч шиваидє



Considering the complexity of adding a dictionary of "words", we used regular expressions in FineReader. A model of word segmentation in a sentence has been developed based on Transformer neural networks with an accuracy of 99% for Romanian texts written in the sixteenth and seventeenth centuries.
The performance evaluation of a specific OCR model for the *Codex Sturdzanus* was developed, obtaining an accuracy of approximately 85.6%.

3. Manuscript „Codicele Bratul” recognition

Codicele Bratul is a miscellaneous manuscript from the 16th century, which includes several religious texts with the interleaving of the Slavonic and Roman languages. As in the case of *Codex Sturdzanus*, the Slavonic sentences are written in black, the Roman ones - in red.

Text recognition was performed by FineReader PDF 15. For this purpose, the dataset with pages from *Codicele Bratul* was prepared. The training set consisted of approximately 2520 glyphs. The test set included 504 characters. In the evaluation, the criterion considered was the accuracy of OCR at the character level. As a result, we found that 129 characters were misrecognized, hence FineReader 15 demonstrated an accuracy of 75%.

Conclusion

In conclusion, ABBYY FineReader 15,16 demonstrated an accuracy of 75-85% in recognizing old uncial Romanian text. This result shows that the tool performs well, even when it has been trained on a relatively small portion of text.

References

1. BUMBU, T.; BURTSEVA, L.; COJOCARU, S.; COLESNICOV, A.; MALAHOV, L. Digitization of Romanian Historical Printings. Vladimir Andrunachievici Institute of Mathematics and Computer Science, Moldova State University, Chișinău, 2023(Valinex), 176 p. ISBN 978-9975-68-497-2
2. Achim Rabus. Recognizing handwritten text in Slavic manuscripts: A neural-network approach using Transkribus. https://www.academia.edu/38835297/
3. Thibault Clèrice. Evaluating Deep Learning Methods for Word Segmentation of Scripta Continua Texts in Old French and Latin: https://jdmdh.episciences.org/6264/pdf
4. Deep Learning for Old Romanian (DeLORo): https://acadiasi.org/proiecte/deloro-deep-learning-for-old-romanian/
5. L. Tsochatzidis, S. Symeonidis, A. Papazoglou and I. Pratikakis (2021). HTR for Greek Historical Handwritten Documents. Journal of Imaging 7(12):260, 2021. https://doi.org/10.3390/jimaging7120260
6. Transcribus: https://www.transkribus.org/
7. The People Transcription Project: https://libguides.viterbo.edu/c.php?g=1206714&p=8835134
8. MALAHOV L., COJOCARU S., COLESNICOV A., BUMBU T. On Recognition of Manuscripts in the Romanian Cyrillic Scripts. In: Proceedings of MFOI-2017, Conference on Mathematical Foundations of Informatics, Chisinau, Rep. Moldova, and November 9-11, 2017
9. https://www.romaniatv.net/codex-neagoeanus-manuscrisul-romanesc-care-prevesteste-evenimentele-politice-si-starea-vremii-pina-in-anul-20_249799.html
10. https://ro.wikipedia.org/wiki/Codex_Sturdzanus