

Distinctive features of recognition for documents printed in the Romanian transitional alphabets

Tudor Bumbu, Lyudmila Burtseva,
Svetlana Cojocar, u,
Alexandru Colesnicov, Ludmila Malahov

Abstract

In this paper, we summarize the research of digitization of documents printed by Romanian transitional alphabet. These printings are the most original Romanian historical documents, which makes our experience useful when researching OCR methods for similar alphabets.

The current work is focused to OCR that is the first stage of scanned documents digitization. The technique of OCR of documents, printed in the Romanian transitional alphabet, is presented. In particular, this technique is embedded in our digitization platform HeDy.

A series of examples is presented to demonstrate the application of the described technique.

Keywords: cultural heritage, OCR, Romanian transitional alphabets

MSC 2020: 68T50.

1 Introduction

The presented work concerns the modern research in European cultural heritage preservation: saving the originality of every country. The author of high detailed work *History of Romanian Spelling* [1] termed the transitional Romanian alphabet as the most original graphic systems in the modern history of European cultures.

Despite of mentioned high originality, the necessity of transitional alphabet sprang from the common European phenomenon. The majority of the first printed books were the cult books. The originality is based on a combination of two factors specific to the Romanian language. The cult books in Romanian regions are written by Cyrillic scripts. But Romanian language belongs to Latin Group and its natural script is Latin. The transition of all Romanian writing to Latin scripts was declared at 1830. But transition process was so complicated and intricate that the author of specific work [2], that concerned transitional alphabet, called it *saga*. This process was not even straightforward, being performed in *trial and error* way. After unsuccessful attempts, some printing houses temporarily returned to Cyrillic script.

Our initial research on the topic is described in [3]. The transition was performed by direct transliteration. This technique succeeded, for example, in transition of specific Cyrillic letters of diphthong: $\text{ѣ} \rightarrow \text{ia}$, $\text{ѥ} \rightarrow \text{ie}$. But direct transliteration was rejected as representation of *phonetic* sense. In the traditional Latin orthography the *phonetic* sense is usually represented by several letters, like, for example, in English: $\text{sh} \rightarrow /ʃ/$. But multi-letters printing was often not accepted by readers, so direct transliteration technique was rejected. The second technique was to establish the closest possible correspondence between the sound value of the old and new letters. This technique was accepted as the main one. The transition rules have been discussed and tested for a long time. For some sounds, the multi-letter representation was replaced by the design of own, specific Romanian, letters, and as a result, the Romanian diacritics was born.

Such a complex and lengthy transition process leads to difficulties with the formal definition of the transitional alphabet. But the general definition is necessary for reference and can be formulated as follows.

Romanian transitional alphabet is the alphabet, that was used in Romania in 1830-1870 and was designed for transition from Cyrillic script to Latin script, defined as 36 letters, 27 of which are modern ones, plus 9 old letters: Ъ ъ, ІѢ ѣ, А а, Ѧ ѧ, Ѣ ѣ, Ѥ ѥ, Ї ї, Ъ ѡ, ІІ іі, ІІ іі.

Digitization of texts printed in transitional alphabets is an important element of digitization of Romanian historical printed publications. The period of using transitional alphabets coincided with the era when printed publications became a necessary part of everyday life. Thus, this period gave a rich legacy of both periodicals and literature, which can be called the first purely

Romanian. So, digitized copies of transitional alphabets prints are interesting not only for linguists but also for historic and literary researchers.

Digitization of documents printed in the Romanian transitional alphabet has specific aspects, both common to all very original documents and specific to Romanian printings.

In this paper, we summarize our experience in digitizing historical documents printed in the Romanian transitional alphabet.

Digitization was implemented by our platform HeDy [4].

2 OCR specific aspects

As mentioned in the preceding section, the formal definition of the transitional alphabet serves as a reference point. This is due to the fact that Cyrillic and Latin letters were mixed in varying proportions, influenced by factors such as the time period, location, and the preferences of typographers, editors, or authors of texts. Book [2] counts up to 17 variants of the transitional alphabets, while some authors declared approx. 20 variants.

Such an irregular variety of transitional alphabets creates problems for all digitization elements, but especially for OCR.

The main problem of the transitional alphabet OCR is the setting of OCR engine. To get acceptable accuracy, OCR tools have to be prepared for a particular variant of transitional alphabet, which means to be: (1) configured, (2) trained and (3) supplied by the proper dictionary.

During the development of HeDy, two approaches to transitional alphabet OCR were tested.

The firstly tested approach is to reproduce the resulted text in its original variant of transitional alphabet. AFR, prepared as it was described above, produces 7% of erroneous words. This is a good result, but the preparation process takes a lot of time and resources.

To achieve more effectiveness, the second approach was tested. This approach uses the general feature of large OCR systems, in our case AFR, to output the result both using original glyphs and substituting them by any sequence of letters from the selected alphabet of recognition. This is called ligatures in AFR documentation. So, the second approach consists in using as ligature the Latinized version of the transitional alphabet that we

specifically developed. For example, both **т** (Cyrillic) and **t** (Latin) will be recognized as **t**.

The OCR using this intermediate alphabet can be the final step if the goal is to obtain a source for transliteration. To solve problems whose purpose is to reproduce the original, we have added a utility to our platform that converts the intermediate OCR output into the desired variant of the transitional alphabet.

The second approach proved fruitful. The OCR errors reduced to 4.8%.

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

The second approach as well solves technical problem of OCR engine setting. AFR, for example, does not support arbitrary Unicode glyphs in its dialogs and forms. Old Romanian letter **ⵀ** was introduced in Unicode only after 2009. Standard system fonts do not contain some Romanian Cyrillic letters. As a result, we see in AFR empty boxes instead of letters during training, alphabet formation, etc. The use of ligatures allows to employ fonts only at the stage of converting the output data, when we control the view.

3 OCR of Romanian transitional alphabet by examples

In this section the transitional alphabet OCR by HeDy is demonstrated by examples. The list of examples, which have some particular and interesting features, was extracted from the book [1]. The scanned sources according this list were obtained from free web bases, the links are referred at footnotes.

3.1 Initial usage

The arrival of a transitional alphabet, both in Muntenia and Moldova, dates back to 1829-1830. One of the most known examples of the initial transitional alphabet usage is Iasi newspaper *Albina româneasca*. Initially, the

appearance of Latin letters was very rare, as we can see from the example¹ in Fig. 1, which is a fragment of the first issue on June 1, 1829. The text is practically entirely printed in Cyrillic script, except for a single Latin letter *i*. The low quality of the original text has been improved by image pre-processing; however, two recognition errors remain.



Ешіи 1 юніе 1829.

АЛБИНА РОМЪНЪСКЪТЪ
ГАЗЕТЪ ПОЛИТИКО-ЛИТЕРАЛЪТЪ.

ДНАИНТЕ КЪВЖНТАРЕ.

Епоха д карѣ трѣим поартѣ семне жѣштите ши вредниче де мираре! Дорѣл лвъцѣтѣрилор нѣнѣмай къ лфрѣцѣще пе лѣкѣбитОрїи оуinei цери лтрѣ кжщигарѣ ачестеи моралниче авѣци, прин карѣ w наѣе се фаче пѣтертикъ ши феричитѣ, чи лѣкѣ ши вамени не асемънаци кѣ Релїгїа, кѣ лимба ши кѣ лециле сжнт

Въцїндѣсе ши сѣпѣндѣсе дрептелор леѣй. Оаре пѣтемѣ нои приви ла ачѣсте бѣне оурмате лнаинтѣ шкилор ностри, фърѣ а ни лѣчриста къ нѣмай наѣа ноастрѣ лѣчѣ маи маре парте есте лифитѣ де ачести лѣбнѣтѣцири ши лнапоетѣ де кжт тоате пѣмѣрїле Еуропей, ши де кжт мѣлте алтеле че лѣкѣдескѣ пре челе лалте пърѣї але пѣмжнтѣлѣї? Чине нѣ сжмте д цара ноастрѣ лиѣа ашезѣмжнтѣрилор

Figure 1. The first issue of *Albina româneasca*, June 1, 1829.

In later issues, especially in the literary supplement *Alăuta Româneasca*, Latin characters become more and more frequent (Fig. 2). One can see that the title is printed entirely in Latin letters and the following text is in Cyrillic with full replacement of the Cyrillic *и* with the Latin *i*. For the spelling of the letter *t*, two variants are used: *т* and *ш*. The word *кѣрѣ* is recognized erroneously as *коре*.

¹The foolowing examples are taken from <https://tiparituriromanesti.wordpress.com/>



ALĂUTA ROMÂNESCĂ.
 SUPPLEMENT LITTERAL
 ALBINEI ROMÂNESCI.
 IASSI. 1. IULIE 1838.

Ачест Суплемент а Газетеи, есь де доуѣ ори пе лунѣ ла

бантора Албінеі Ромънеші ꙗн Еши.

Дін партеа Редакціѣ.

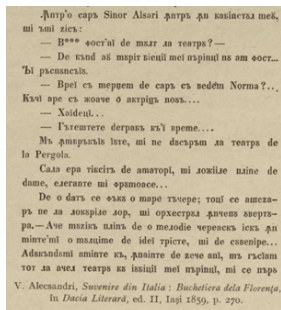
Редакція Албінеі Ромънеші, нѣ аѣ крѣчат, ꙗн коре де ноѣ ані, нічі о жѣртѣ спре а ꙗші мѣлцѣмі аБОНАЦІИ. Дорінд съ лі дее о ноѣ довадѣ де вѣна еї воинѣ, ел аѣ лѣат мѣсѣрїле кѣвїинчоасе ка Алѣшта Ромънеаскѣ, каре пѣнѣ акѣм се да нехотѣрїт, съ еасѣ ꙗн вігорїе, регѣлат де доѣѣ ори пе лѣнѣ, ла 1 шї ла 15 а лѣнеї, кѣспрінзїнд нѣмаї лѣкрѣрї літерале прекѣм ачест ꙗнѣгї нѣмѣр. Редакція се ва сїргѣї ка Алѣшта съ рѣсѣне прѣдѣктѣрїле дѣхѣлѣї челе маї несѣ шї челе маї интересанте пентрѣ четїторї. Еа ꙗшї ва ꙗмпліні скопосѣл кѣ ажѣторѣл мѣлтор тїнерї літерарї, карїї аѣ БИНЕВОїт а фагѣдѣї лѣкрѣрїле лор ачестеї пѣблїкації періодїче.

Figure 2. *Alăuta Româneasca*, July 1, 1838.

3.2 Complete usage

The next example presents the full version of the transitional alphabet how Heliade Radulescu, Balcescu, and Treboniu proposed it, and how Laurian, Asachi, Kogalniceanu, etc. used it. We selected as an example an excerpt from the famous author Vasile Alecsandri's novel *Suvenire din Italia: Buchetiera de la Florența* (*Souvenirs from Italy: A Florist from Florence*) (Fig. 3) because the works of the classical authors were re-published by modern Romanian alphabet. Modern re-publications are a useful source for verifying the transliteration tool. The transliteration tool, in turn, can be used for creation of datasets for both training and validation of the OCR engine. The quality of OCR is very good, with only one error in this fragment (the Latin letter **d** recognized as the Cyrillic letter **б**). Most of the letters are Cyrillic, with the letters **D**, **d**, **i**, **m**, **n** being written in Latin script. In the case of the last letter, the Cyrillic spelling **н** is also found. Note that all proper names: **Sinor Alsari**, **Norma**, and **Pergola** are written in Latin characters.

Another example of using the transitional alphabet in its mature state



̦лтр'о саръ Sinog AIsari ̦лтръ ̦лн кабинетѢл меѢ,
 ши ʔмї ʔїсѢ:
 — В*** фост'аї де мѢлт ла театрѢ? —
 — Де кѢнд аѢ мѢрїт бїеції меї пѢрїнци нѢ ам ФОСТ...
 ̦ї рѢспѢнсѢїѢ.
 — Вреї сѢ мерѢем де саръ сѢ ведем Norma?... КѢчї
 вре сѢ жоаче о актрїцѢ поѢъ...
 — Хаїдеці...
 — ГѢтешете деграбъ кѢ време...
 МѢ ̦лѢбрѢкѢїѢ їѢте, ши не дѢсѢрѢм ла театрѢ де In
 Pergola.
 Сала ера тїкїсїѢ де аматорї, ши ложїїле плїне de dame,
 елѢганте ши фѢрѢмоасе...
 Де о датѢ се ѡзкз о шарѢ тѢчере; тоцї се ашеѢсарѢ
 пе ла локѢрїле лор, ши орѢкестра дїчеѢз аверѢрѢ-
 рѢ. — Ave muzica talia de o melodie cereasca icu a
 mure'mi o maxime de idei triste, mi de esențe...
 Adaranduri amare ca, pasiune de acei ani, nu ruciam
 tot la ovela teatrului ca insuși mei părinți, mi se țira
 V. Alexandri, *Souvenirs din Italia: Buchetiera din Florența*,
 în *Dacia Literară*, ed. II, Iași 1859, p. 270.

Figure 3. Fragment of *Souvenirs from Italy: A Florist from Florence* by V. Alexandri, 1840

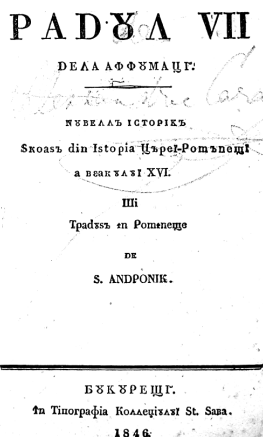
is interesting by the confident use of the transitional alphabet in secular literature. The example is from the novel *Radu VII from Afumați*² by Stefan Andronic (Fig. 4A, p. 347). The *Dictionary of Romanian literature* declares this novel as the first Romanian historical roman.

3.3 One direction conversion Cyrillic – Transitional – Latin

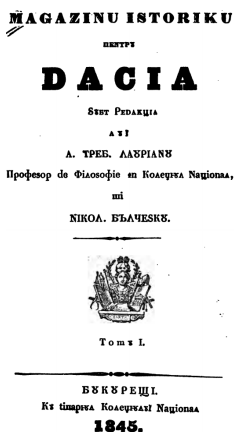
The book [1] presents the examples which are named *effective transition*. The definition of *effective* here means that the mentioned printed publications: started issuing using the transitional alphabet immediately after the announcement; used the transitional alphabet only for a few years; finally, were issued entirely in the Latin alphabet without any returns.

The presented example is the fragment from *Magazinu istoricu pentru*

²<https://revistatransilvania.ro/wp-content/uploads/2019/11/1846.-S.-Andronic-Radu-VII-de-la-Afumatii.pdf>



A.



B.

Figure 4. A. Cover of *Radu VII from Afumați* by S. Andronic, 1846
 B. Title page of *History Magazin for Dacia*, 1845

*Dacia*³ (Fig. 4B, p. 347), edited by Laurian and Balcescu, that used the transitional alphabet in 1845–1847 and then the pure Romanian Latin alphabet.

4 Accuracy evaluation

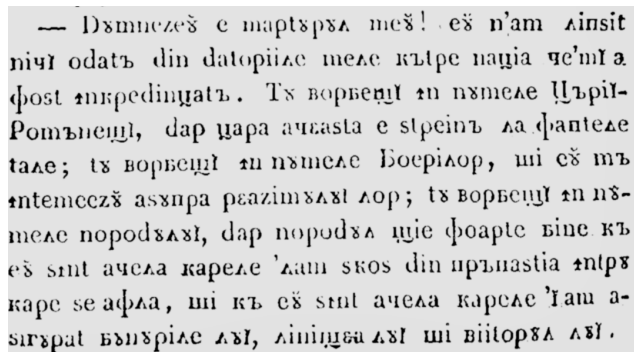
Here we propose analysis of accuracy in recognition of p. 20 of the novel *Radu VII from Afumați* (see Sec. 3.2 above).

The most frequent errors were observed in the letter **б** (changed to **в**) and the letter **н** (changed to **п**). With total 1172 characters on the page and 93 erroneous characters, we have the 92.1% accuracy.

This result could be improved by adding a user dictionary as part of the FineReader language model. With the dictionary, we got only 31 erroneous characters, and the accuracy became 97.4% that seems to be a good result.

However, after adding the dictionary, while most errors with the letter **н** were resolved (as similar words were added to the dictionary), there were still recognition errors with the letters **и**, **б**, **м**, and some others. Also, the

³https://archive.org/download/magazinuiistoric01unkngoog/magazinuiistoric01unkngoog_tif.zip



— Дѣмнезеѣ е мартѣрѣл меѣ ! еѣ н'ам лѣнсѣт
нѣчѣ одатѣ дѣн даторѣѣле меле кѣтрѣ наѣѣа че'мѣ а
ѣост ѣнкрѣдѣнѣатѣ . Тѣ ворбѣнѣѣ ѣн нѣмеле Цѣрѣѣ-
Ромѣнеѣѣѣ , дар ѣара ачеаста е стрѣнѣѣ ла ѣаптеле
тале ; тѣ ворбѣнѣѣ ѣн нѣмеле Боерѣлор , шѣ еѣ мѣ
ѣнтемеезѣ асѣпра реазѣмѣлѣѣ лор ; тѣ ворбѣнѣѣ ѣн нѣ-
меле норѣдѣлѣѣ , дар норѣдѣл ѣѣе ѣоарте бѣне кѣ
еѣ снѣт ачела кареле 'лам скос дѣн прѣпастѣа ѣнтрѣ
каре се аѣла , шѣ кѣ еѣ снѣт ачела кареле 'ѣам а-
сѣѣрат бѣнѣрѣле лѣѣ , лѣнѣѣеа лѣѣ шѣ вѣѣторѣл лѣѣ .

— Дѣмнезеѣ е мартѣрѣл меѣ ! еѣ нам лѣнсѣт нѣчѣ одатѣ дѣн даторѣѣле меле кѣтрѣ наѣѣа чемѣ а ѣост ѣнкрѣдѣнѣатѣ . Тѣ ворбѣнѣѣ ѣн нѣмеле Цѣрѣѣ-Ромѣнеѣѣѣ , дар ѣара ачеаста е стрѣнѣѣ ла ѣаптеле тале , тѣ ворбѣнѣѣ ѣн нѣмеле Боерѣлор , шѣ еѣ мѣ ѣнтемеезѣ асѣпра реазѣмѣлѣѣ лор ; тѣ ворбѣнѣѣ ѣн нѣмеле норѣдѣлѣѣ , дар норѣдѣл ѣѣе ѣоарте бѣне кѣ еѣ снѣт ачела кареле лам скос дѣн прѣпастѣа ѣнтрѣ каре се аѣла , шѣ кѣ еѣ ачела кареле 'ѣам асѣѣрат бѣнѣрѣле лѣѣ , лѣнѣѣеа лѣѣ шѣ вѣѣторѣл лѣѣ .

Figure 5. Digitization of *Radu VII from Afumati* fragment

word **ѣубенеѣ** was incorrectly recognized this second time, even though the model recognized it correctly without the dictionary.

It's worth noting that the model almost flawlessly recognizes special characters and punctuation, often erring in the letters **н**, **п**, **ш**, **м**, **б**, and **ц**. The model learned to recognize the letter **ѣ** well and also correctly recognizes the letter **л**.

The letter **ѣ** was not on this particular page, but it was often identified as **e** in other pages. In reality, later the **e** was used instead of this letter.

Later it was observed that the letter **ѣ** also appears in texts but was always replaced with **i**. The letter which looks like **ѣ** with a line over it was counted as a simple **ѣ**.

Ної кредем къ о аsемине Istopie este кѣ пѣтин-
цѣ. Pentрѣ асeаstа sokotim къ чeй чe се окѣпѣ къ
Istopia нѣстрѣ, нѣ трeвѣ а се цинe нѣмаї de чeea чe
аѣ лѣкрат шї аѣ зїs istorїчїї nostrїї чeї modernї;
dap, tot нѣтр'о вpeme, фoлoсїndѣse de аdевѣрѣрї des-
коперїte de дѣншїї сѣ мeргѣ маї департе, сѣ алер-
цe ла їsvѣрѣлe opїцїналe, сѣ каѣte шї сѣ адѣne tѣte
datѣpїлe пѣтинчїѣse, шї атѣнчї vor пѣтe чeсе o Бѣнѣ
Istopie.

Ної кредем къ о аsемине Istopie este кѣ пѣтинцѣ. Pentрѣ асeаstа sokotim къ чeй чe се окѣпѣ кѣ Istopia нѣстрѣ, нѣ трeбѣ а се цинe нѣмаї de чeea чe аѣ лѣкрат шї аѣ зїs istorїчїї nostrїї чeї modernї; dap, tot нѣтр'о вpeme, фoлoсїndѣse de аdевѣрѣрї дескоперїte de дѣншїї сѣ мeргѣ маї департе, сѣ алерцe ла їsvѣрѣлe opїцїналe, сѣ каѣte шї сѣ адѣne tѣte datѣpїлe пѣтинчїѣse, шї атѣнчї vor пѣтe чeсе o Бѣнѣ Istopie.

Figure 6. Digitization of *Magazinu istoricu pentru Dacia* fragment

5 Conclusion

Digitization of historical documents, printed by rare and unique alphabets, is an important part of the preservation of a specific national cultural heritage. The solutions of OCR problems, which arise during digitizing such documents, has scientific and practical value.

These solutions are taken into attention in our HeDy platform, which exists in free versions for both web and desktop. In addition to OCR tools presented in current work, HeDy platform provides the tool for transliteration, that simplifies reading the historical documents content.

Acknowledgments. This work was prepared as part of the research project 20.80009.5007.22 *Intelligent information systems for solving ill-structured problems, processing knowledge and big data.*

References

- [1] Pârvu Boerescu, *Din istoria scrierii românești*, București: Editura Academiei Române, 2014, 400 p. ISBN: 978-973-27-2459-0. (in Roma-

nian).

- [2] Ștefan Cazimir, *Alfabetul de tranziție; Jurnal de tranziție*, Oscar Print, 1996. 197 pp. ISBN: 9789739757348. (in Romanian).
- [3] S. Cojocaru, L. Malahov, A. Colesnicov, 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.
- [4] T. Bumbu, L. Burțeva, S. Cojocaru, A. Colesnicov, L. Malahov, “A platform for processing heterogeneous documents,” in *Proceedings of the 17th International Conference “Linguistic Resources and Tools for Natural Language Processing”*, Univ. “A.I.Cuza”, Iași, 2022, pp. 141–151.

Tudor Bumbu^{1,2}, Lyudmila Burtseva^{1,3},
Svetlana Cojocaru^{1,4},
Alexandru Colesnicov^{1,5}, Ludmila Malahov^{1,6},

Received November 29, 2023

¹“V. Andrunachievici” Institute of Mathematics and Computer Science, Chisinau, Republic of Moldova

²ORCID: <https://orcid.org/0000-0001-5311-4464>

E-mail: bumbutudor1@gmail.com

³ORCID: <https://orcid.org/0000-0002-9064-2538>

E-mail: luburtseva@gmail.com

⁴ORCID: <https://orcid.org/0009-0003-1025-5306>

E-mail: svetlana.cojocaru@math.md

⁵ORCID: <https://orcid.org/0000-0002-4383-3753>

E-mail: acolesnicov@gmx.com

⁶ORCID: <https://orcid.org/0000-0001-9846-0299>

E-mail: ludmila.malahov@math.md