Challenges and Solutions in Transliterating 19th Century Romanian Texts from the Transitional to the Latin Script

Marc Frincu
Department of Computer Science School of Science and Technology Nottingham Trent University, UK
marc.frincu@ntu.ac.uk

Simina Frincu and **Marius E. Penteliuc**
Department of Computer Science
Faculty of Mathematics and Computer Science West University of Timisoara, Romania
ioana.frincu@e-uvt.ro
marius.penteliuc@e-uvt.ro

Abstract

During the 19th century, the Romanian script has undergone a massive yet uneven transition from the Cyrillic to the current Latin alphabet. The amount of existing literature written in that script as well as the problems it poses for OCR and transliteration engines make the problem highly challenging from a Big Data perspective. In this paper, we discuss the issues and propose and test a machine-learning solution trained on small datasets using either transfer learning from Latin/Cyrillic or from scratch.

1 Introduction

Until the early 19th century Romanian texts were written in the Romanian Cyrillic Script (RCS) containing around 43 characters, a version of the script different from the standard Church Slavonic or Russian scripts. By the end of the 18th century, the first attempt to simplify the script to 38 letters comes from (Văcărescu, 1787). In 1823, to meet didactic purposes, I. H. Rădulescu highlights the same necessity for a reduced 30-letter script. Nonetheless, the reforms (to optimize or simplify the alphabet) proposed over time by different cultural figures ((Iorgovici, 1799), (Budai-Deleanu, 1812), (Diaconovici Loga, 1818), (Rădulescu, 1828) or (Pleșoianu, 1828)) remained until the official adoption in 1860 at the stage of individual and unofficial initiatives. The drive behind the change pertained also to the desire to reassert the Latin values of Roman origin of Romanian people, in the context of the sociopolitical events unfolding across Europe.

The alphabet transition did not occur abruptly (several versions coexisted between authors, publishing houses, editors, and regions) or simultaneously across the historical Romanian regions of Wallachia, Moldavia, and Transylvania (Cazimir, 2006). Yet, all these versions were based on the Simplified Modern RCS and a variable, increasingly higher in time, proportion of Latin letters.

The alphabet transition is extremely interesting for researchers studying the diachronic evolution of the language and encompasses thousands of typed manuscripts (some not digitized) written in various transitional script versions. Understanding these manuscripts starts with scanning, converting the scanned images into digital documents, reading the documents, and analyzing their content based on the researcher's objectives. The OCR process is the main driver behind digitization and it is here that existing software fails to recognize the Romanian Transitional Script (RTS), partly due to the quality of the original paper. Thus, Machine Learning (ML) models are better suited to handle different types of scanned documents and script versions (e.g. font type, publishing house, region). Tools like Transkribus (Miloni, 2020) and the open-source Tesseract (Smith, 2007) have been designed for such cases. However, the accuracy of the models depends on the volume and variety of training data. This turns the process into a Big Data problem where most data preparation is manually handled before training and testing the models.

The RTS digitization process consists of: (1) conversion to RTS characters (preserving the original text); and (2) interpretative phonetic transcription into Latin (transforming the original text into a version readable by modern researchers).

Following the CRISP-DM methodology (Wirth and Hipp, 2000) focusing on data understanding and preparation, we compare 2 approaches that lead to promising Tesseract models trained on few data and digitized to Latin/RTS.

2 Related Work

2.1 RTS Studies

Several studies (Cazimir, 2006), (Boerescu, 2014) refer to a formal "modernization" of the RCS after 1830. More precisely, the typographical Cyrillic capital letters were "carved" using the Latin-type model, namely redesigned to resemble the Latin letters. Thus, the graphical overlay can be explained by the fact that some Latin capitals were identical in sound and meaning to Cyrillic ones (A, E, I, K, M, O, T). In contrast, others coincided graphically yet differed semantically (Cyrillic B for V, C for S, H for N, P for R, X for H). The purpose of this initiative, sometimes leading to surprising approaches (cf. Fig. 1), was to prepare the readers for the alphabet transition about to take place.

Two methods can be used to render a text written in the Cyrillic alphabet into Latin: transliteration or interpretative phonetic transcription. The first implies a one-to-one mapping (IRS, 1997), a character-by-character conversion, more precisely each Cyrillic letter to be replaced with one and the same Latin letter, irrespective of the context within the converted system. The latter demands an accurate determination of the phonetic values represented by the Cyrillic letters (Ursu, 1960). Both methods present disadvantages and are not entirely satisfying. The shortcomings of the transliteration method (the Latin script counts fewer letters than the Cyrillic one, therefore the same Latin letter with various diacritics attached to it can stand for two or even three Cyrillic letters) and the difficulties of the phonetic transcription lead to a hybrid approach and a composite solution.

2.2 Automated Transliteration and ML

Most works on automating the RTS transliteration were done by researchers in Rep. Moldova as the script was used both there and in Romania.

Boian et al. (2014) mention at least 7 versions for RTS, provide a first look into the challenges of transliterating RTS, and mention that except for one (for which they used a replacement), all RTS characters are available in Unicode (UTF-16). The reported percentages using the proprietary paid ABBYY FineReader with and without training range between 63 and 95.4%.

Cojocaru et al. (2016) identify challenges when transliterating older scripts using OCR tools not supporting them. They mention the RTS versions and 3 existing fonts that cover the RTS characters, focusing on every script version starting from the RCS to the Moldavian Cyrillic Script in use in Rep. Moldova in the 20th century. Their approach targets ABBYY FineReader and experiments use both one-to-one mapping and rule-based context transliteration but they do not provide the number of tested documents and errors only showing the upper limit of 96% in terms of accuracy without providing an error distribution plot or mean value.

Demidova and Burteva (2017) also focus on historical documents written in RTS. In addition to the previous paper, they briefly describe their transliteration module written in the Java language but do not present comprehensive results for their experiments. It is unclear if the module only transliterates already digitized documents or goes through the entire OCR process too. The reported accuracy is 99% without mentioning the dataset size.

Gîfu and Plamada-Onofrei (2017) focus on creating a corpus of transliterated text to facilitate the automatic recognition and interpretative transcription from RTS to the modern Latin script.

While focusing on the older RCS and not on RTS the work of Burlacu and Rabus (2021) is interesting as it uses Transkribus, another online tool with limited free access that we considered. Their study involves handwritten manuscripts and the provider CER (Character Error Rate) is around 10%. We note here that Transkribus requires thousands of words for training its models (the authors used up to 30,900 words for one of their models) which calls for a significant upfront effort.

Compared to existing work using paid software and briefly discussing results, we focus on the open-source Tesseract Engine proposing a 2-phase automatic transliteration process: (1) to Latin/RTS characters followed by an interpretative phonetic transcription; (2) a corpus-based correction to improve the accuracy of the final text in Latin script.

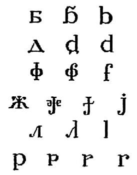


Figure 1: Example of transitional characters invented and used in some of his texts by I. H. Rădulescu to visually ease the alphabet transition and familiarize readers with the Latin script (Cazimir, 2006).

3 Current Challenges

3.1 Processing

When dealing with large collections of historical books several preprocessing and processing challenges occur. Foremost, these documents must be digitized so that OCR and transliteration tools can generate documents readable by present-day researchers (and the general public for that matter). This phase is largely manual and implies a significant amount of time and effort. Next, the ML model must be trained and validated on a relevant data sample covering the problems identified in Sec. 3.2. This process requires a manual transliteration of the training and validation data sets that will act as ground truth in the training and validation steps of the model. Finally, the best models need to be tested on a test data set which must also be manually transliterated to have a ground truth for automatically computing the errors. Our experiments have shown that the manual process takes around 30 minutes for 1 page with the time spent improving as users get accustomed to the RTS.

While a lot of manual transliteration is required, the computational and storage space also becomes an issue. Depending on the image format a scanned color page takes between 100 KB (jpeg) and ≈2 MB (tif) with the transliterated text file taking ≈2 KB. This means that a single book of 100 pages will occupy 10-200 MB. When it comes to thousands of books from the alphabet transition period storing all the data is a concern too. The Tesseract OCR process is fast taking between 0.18-0.59 secs per page while the training of a k-fold model ranges from 13.5-17.2 to 613-2,200 secs per fold times the number of folds and iterations (cf. Sec. 5).

3.2 OCR and Transliteration

All the titles printed between 1828-30 and 1860 used for the validation, training, and test phases have been selected by applying the "transitional alphabet" filter in the electronic catalogs of the libraries hosting rare/old book collections. The different degrees and types of paper alterations impact the ML-based OCR process and demand for additional processing of the images subject to further training. Hence, we have aimed at selecting scanned pages bearing a wide variety of physicochemical and a few physicomechanical types of age-related damage. These include (e.g., Fig. 3):

1) Thick binding, ripped stitching, or broken spine which led to poor quality scans, i.e. text deformations (crooked/bent text).

2) Creases, folds, wrinkles, and undulation due to humidity changes.

3) Moisture halos, ink discoloration, foxing, burns, tearing, grease stains, glue residue.

4) Presence of post-printing elements, e.g. signatures, institutional stamps, inventory numbers, notes in pencil/soluble ink/pen, etc.

We have also considered printing aspects likely to make the OCR process more difficult, some of which needed to be tackled individually:

1) Typesetting using various inks (usually black or red), typefaces, and fonts (e.g. drop caps, enlarged and illustrated initial letters meant to mark the beginning of a book/chapter/section).

2) Text visible from the verso of the sheet due to thin physical support.

3) Two-column versus single-column printing approach, framed and/or manually underlined text.

4) Glossing with marginal/interlinear notations, either numbered or marked by typographical symbols and sometimes separated from the main text by a separator line.

4 Proposed Solution

The existing literature on RTS transliteration / phonetic transcription is lacking a clear description of the datasets used for training and testing and relies in some cases on paid software (cf. Sec. 2). We present our approach for testing and assessing two scenarios, using either a Latin or RTS baseline for training through transfer learning or from scratch the models in the open-source Tesseract 5.2.

4.1 Improving Transliteration Accuracy

Transliterating from RTS to Latin poses several challenges including character ambiguity (cf. Sec. 5) and phonetic transcription (rule-based approach depending on the subsequent characters). As Tesseract can only perform OCR the phonetic transcription must take place afterward and therefore its efficiency depends on the accuracy of the OCR process. This second step requires replacing the transliterated character with another single or group of characters based on context. E.g., ч is interpreted as: *c* if followed by e or i; *ce* if followed by a; *ci* otherwise (Cojocaru et al., 2016).

To assess Tesseract's ability to accurately perform OCR we propose two approaches. Each uses a different baseline, Latin or RTS. The reason is that many documents have mixed Latin and RTS texts causing the phonetic transcription to fail as the text sections are neither automatically nor manually tagged with the script they use. For instance, the title can be in Latin, while the text itself is in RTS (cf. Fig. 4). In such a case, Latin *c* for instance is unnecessarily (and wrongly) phonetically analyzed

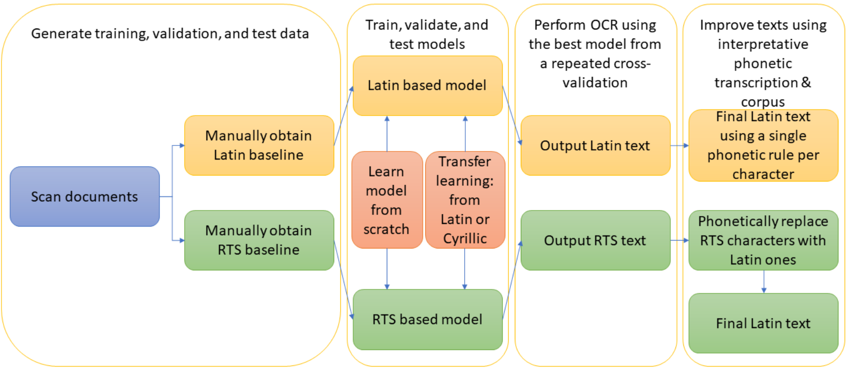
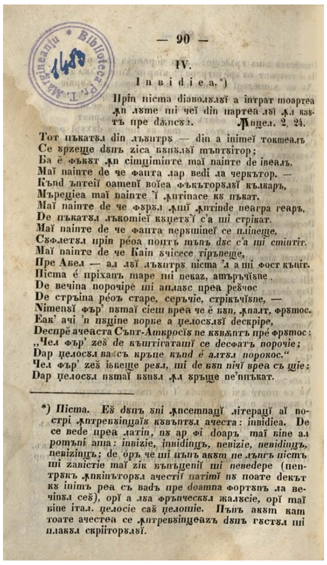

Generate training, validation, and test data
Scan documents
Manually obtain Latin baseline
Manually obtain RTS baseline
Train, validate, and test models
Latin based model
Learn model from scratch
Transfer learning: from Latin or Cyrillic
RTS based model
Perform OCR using the best model from a repeated cross-validation
Output Latin text
Output RTS text
Improve texts using interpretative phonetic transcription & corpus
Final Latin text using a single phonetic rule per character
Phonetically replace RTS characters with Latin ones
Final Latin text

Figure 2: Overview of the two proposed approaches.



— 90 —

IV.

Inbidiea.*)

Прin пiсma diаволȣлȣї a intрат moартea
ꙟп лȣme шi чеї din партea лȣї ꙟл казъ пре dꙟпсȣл. Ꙟцел. 2, 24.

Тот пъкатȣл din лъȣнтрȣ — din a inimeї тоkmeaлъ
Се ȣрzeще dȣпъ zica вȣнȣлȣї mънтȣiтор;
Ба ё фъкȣт ꙟп сiмџiminте maї nainте de iвeалъ,
Maї nainте de че фапта лар вedi ла черкътор. —
Кънd ъnтеiї oamenї воiea фъкъторȣлȣї кълкаръ,
Мърeџiea maї nainте 'ї ꙟнтinace кȣ пъкат.
Maї nainте de че фȣрȣл ꙟшї ꙟнтinde neaгра геаръ,
De пъкатȣл лъкomieї кȣџетȣ'ї с'а шi стрiкат.
Maї nainте de че фапта нерȣшinеї се плiнеще,
Сȣфлетȣл прin реоа пoнтъ mънъ dȣс с'а шi сminтiт.
Maї nainте de че Каin ȣчiсесе тiрънеще,
Пре Авел — ал лȣї лъȣнтрȣ пiсma 'л a шi фост кънiт.
Пiсma é прiхапъ mаре шi некаz, amъръчiȣне,
De вечinа порочiре шi аплаȣс преа реȣчос
De стрънна реоъ старе, серъчiе, стрiкъчiȣне, —
Nimenȣї фър' пȣmaї сieш врea че ё вȣн, ꙟнaлт, фрȣmос.
Еак' ачi 'n пȣџiне ворве а џелосȣлȣї deскрiре,
Deспрё ачеаста Сънт-Амвросiȣ пе кȣвꙟнтъ пре фрȣmос:
„Чел фър' zeȣ de къштiгаташї се deсфатъ порочiе;
Dар џелосȣл ва съ крънe кънd é алтȣл порокос."
Чел фър' zeȣ iьвеще реȣл, шi de вȣн пiчї врea съ щiе;
Dар џелосȣл пȣmaї вȣнȣл ꙟл ȣръще пе'нпъкат.

*) Пiсma. Еȣ dȣпъ ȣнi ꙟнсemnaџї лiтераџї аї ноcтрi ꙟнтревȣinџаїȣ кȣвънтȣл ачеста: inbidiea. De се веde преa латin, пȣ ар фi doаръ maї вiне ал ромънi апа: inbizie, inbidinџъ, nebizie, nebidinџъ, nebizinџъ; de орь че шi пънъ акȣm пе лънгъ пiстъ шi zавiстie maї ziк вънъџенiї шi neвеdере (пентрȣкъ ꙟнкiпъторȣл ачестiї nатimї нȣ поате dекът кȣ inimъ реa съ вadъ пре doamna фортȣнъ ла вечinȣл сеȣ), орї а лȣа фрънческȣл жалȣсie, орї maї вiне iтал. џелосie саȣ џелошie. Пънъ акȣm кam тоате ачестеa се ꙟнтревȣinџеazъ dȣпъ гȣстȣл шi плакȣл скрiiторȣлȣї.

Figure 3: Glossed text from 1847 written in RTS with notation separated by a line and marked by an apostrophe. Also present moisture halo, deformed text, institutional stamp, and text visible from verso.

in the title. Due to constraints, for the Latin baseline (impossible to interpret Latin characters), the phonetic transcription is based on a single selected rule, e.g., ч → *ci*. The RTS approach focuses instead on Cyrillic but it too can misinterpret Cyrillic characters for Latin ones. The key difference is the output from Tesseract and the fact that the RTS approach performs the phonetic interpretation and transliteration in one step during the Latin conversion ignoring any Latin characters. A major issue during transliteration is the character similarity between scripts, e.g. Latin C and Cyrillic C – Latin S (cf. Sec. 2) which can be solved by providing the model with enough and varied training data.

Both texts are improved by using a corpus from the training and validation documents. At the moment, candidate words are selected based on the Levenshtein distance (cf. Sec. 5) but other methods (e.g., based on n-grams) are possible.

5 Experiments

CER is a metric for assessing OCR quality. There is no consensus on what a good CER value is. Burlacu and Rabus (2021) mention a rate less than 5% (or <10% for a text to be manually corrected in a time less than that needed for manual transliteration), while (Halley, 2009) mentions 2% as a good result and 10% as average.

$$CER = (S + D + I)/N \tag{1}$$

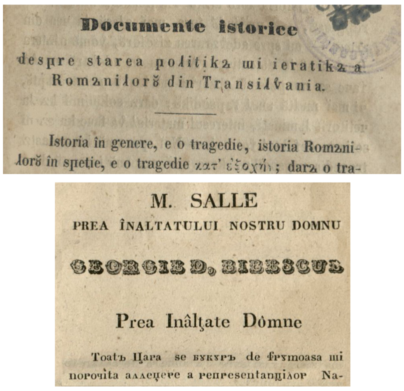
where $S + D + I$ represents the Levenshtein distance and corresponds to the number of substitutions (S), deletions (D), and insertions (I) required to make two texts equal; and N is the length of the baseline (ground truth) text. Tesseract computes by default BCER (Bag of Characters Error Rate):

$$BCER = \sum_{i=1}^{no_{words}} \left(\frac{S_i + D_i + I_i}{N_i} \right) / no_{words} \tag{2}$$

It can be shown that $BCER \geq CER$. CER is a function of the overall quality and BCER penalizes text where the error is less uniformly distributed.

5.1 Setup

To test our approaches we collected a corpus of over 3,000 pages from distinct documents (1837-1861) from the Timisoara Central University Library. In this paper, we used a small subset of 30 pages of 24,148 characters (out of which 64.4% are Cyrillic). The Cyrillic characters' percentage per page was $61.8 \pm 9.6\%$. Each page was scanned and manually converted/transliterated (into RTS/Latin) to obtain two baselines. Unfortunately, the existing corpus from Gîfu and Plamada-Onofrei (2017) does not



Documente istorice

desпre starea пoлitikă шi ieratikă a
Romăniлorŭ din Transiлvania.

Istoria în genere, e o tragedie, istoria Romăni-
лorŭ în sпeţie, e o tragedie κατ' ἐξοχήν; dară o tra-

M. SALLE
PREA ÎNALTATULUI NOSTRU DOMNU
GEORGIE D. BIBESCUL

Prea Inâlţate Dômne

Toatъ Цara se бѹкѹръ de фрѹmoasa шi
пoрoчita aлeџeрe a reпрesentanџiлoр Na-

Figure 4: RTS text in which the title is written only with Latin characters. 1850 (top) and 1844 (bottom).

include the scanned pages making it unusable for our experiments. While small, our dataset allowed us to assess Tesseract's potential to create good models from a few data. Tesseract uses LSTM deep network architecture. We trained our models either from scratch or starting from existing models through transfer learning (Latin or Cyrillic) and stopped the training after 10,000 iterations. One test page containing 745 characters (out of which 71.14% Cyrillic) was used.

Several model validation scenarios were used:

(S1): Initial 5-fold cross-validation of a randomly picked 15-page dataset for creating a model and using a single test page.

(S2-k): A repeated k-fold cross-validation for creating the model where $k \in \{3, 10, 29\}$. One page was omitted as it was unreadable by Tesseract.

We name the models for each baseline S1-L and S2-k-L, respectively S1-RTS and S2-k-RTS. Our aim is to assess if there are differences in CER when performing the ML-based conversion into RTS (followed by a Latin transliteration) or directly transliterating into Latin (Romanian). We also evaluated if using a corpus comprising the trained data can improve CER. We considered two cases, one containing a corpus from various regions and publishing houses, and one from Rădulescu's publishing house. Color pages and their b/w counterparts were tested separately. As results were better for color pages we present exclusively these.

CER was computed using the Levenshtein distance (Eq. 1) after removing all spaces from baseline and transliterated texts. The BCER value was computed automatically by Tesseract.

5.2 Results

The test service and data are available online[1]. Table 1 shows the results of our experiments. For repeated k-fold cross-validation we show the best results (k=3). As the number of folds increased both CER and BCER dropped indicating the sensitivity of our models to the small dataset. For Latin, the best model started from an existing Latin model enriched with our dataset and provided a CER=1.8 for S2-Lat. For RTS the best model was also one trained by enriching a Latin model and achieved a CER=17.7 for S2-3-RTS. The models starting from Cyrillic performed slightly worse for RTS. The reason for the high CER can be traced to the similarity of vocals in Cyrillic and Latin, e.g., а – a; е – e; і – i; о – o. As CER was computed based on the Unicode value it produced high values as most Cyrillic vocals were identified as Latin characters. Ignoring them reduces the number of wrongly classified characters by 52–59% depending on the base model. The RTS model trained from the Cyrillic base model performed slightly worse than the Latin-derived RTS model, partly due to wrongly classifying more Latin (e.g., *t*) characters. Improving these misclassifications would make the Cyrillic-derived model better. This would be ideal due to the non-existing phonetic transcription available for the Latin baseline. Overall, the Latin base model misidentified 52 characters compared to 54 by the Cyrillic-based one.

When using the training corpus to reduce CER for the test page we noticed that this happened only for a single model in the 5-fold and led to a 0.1% improvement. When using a model trained only for Rădulescu (2^{nd} fold of a 3-fold) no CER improvement was noticed except when assuming that the corpus already contained all the words in the test page (0–2.3%). The reason is that the Levenshtein distance is unsuited for the task as it compares the words in terms of changes in characters not semantically. Even assuming a corpus containing the correct test page does not lead to a $CER = 0$ across the board as the OCR process can introduce additional erroneous words (cf. Sec. 3).

From a formal, script-related perspective, a typology of the recognition failure cases consists of: 1) Errors due to the graphic similarity between letters, accented letters mistaken for other letters, or for numbers resembling them visually, e.g., i – î,

[1] https://transitional-romanian-transliteration.azurewebsites.net/

Target	Latin				RTS					
From	scratch		Latin		scratch		Latin		Cyrillic	
Scenario	S1	S2-3	S1	S2-3	S1	S2-3	S1	S2-3	S1	S2-3
CER %	–	10.6	2.5 ± 0.4	**1.8**	56.0 ± 7.3	19.4	27 ± 4.0	**17.7**	33 ± 2.0	19.6
BCER %	–	15.5	4.5 ± 1.3	8.2 ± 0.7	21.8 ± 6.4	20.7	13.9 ± 2.4	13.8	13.5 ± 1.8	15.5

Table 1: Test results for our two approaches including the model we started from, scenario, and error metrics.

n – п (*p*), *m* – ш (*ș*), í – *l*, ó – *6*, k – к (c/ch/k).
2) Errors caused by a lack of previous training. E.g. Greek symbols, and Latin script fragments.
3) Errors encountered in transliterating certain double consonants. It was noted that while double *s* and double *n* were 100% recognized, double *l* was always rendered faultily.

6 Conclusions

In this paper, we addressed the problem of transliterating 19th century Romanian texts. We proposed a solution based on Tesseract and demonstrated it on two targets: Latin and RTS. Initial results for Latin on a small dataset are very good but phonetically interpreting the text is challenging due to the mix of Latin and RTS phrases in some documents. Results for RTS indicate the need for a richer training dataset due to the similarity between Latin and Cyrillic characters. Future work will consider these aspects. We will also assess other methods for corpus-based text improvement such as n-grams and TF-IDF.

Acknowledgements

This work was supported by a grant of the Romanian Ministry of Research, Innovation and Digitization, CCCDI - UEFISCDI, project number PN-III-P2-2.1-PED-2021-0693, within PNCDI III.

References

P. Boerescu. 2014. *About the History of Romanian Writing (Ro.)*. Editura Academiei Române.

E. Boian, C. Ciubotaru, S. Cojocaru, A. Colesnicov, and L. Malahov. 2014. Cultural and historical heritage digitization, recognition and conservation. *Akademos*, 1(32):61–68.

I. Budai-Deleanu. 1812. *The fundamentals of Romanian grammar (Ro.)*. În Buda Sau tipărit la Crăiasca Universităţii Tipografie.

C. Burlacu and A. Rabus. 2021. The digitization of documents written in the romanian cyrillic script by using transkribus: new perspectives (ro.). *Diacronia*, (14):(1–10).

S. Cazimir. 2006. *Transitional Alphabet 2nd ed. (Ro.)*. Humanitas.

S. Cojocaru, L. Burteva, C. Ciubotaru, A. Colesnicov, V. Demidova, M. Ludmila, M. Petic, T. Bumbu, and S. Ungur. 2016. On technology for digitization of romanian historical heritage printed in the cyrillic script. In *Procs. of the Conference on Mathematical Foundations of Informatics*, pages 160–176.

V. Demidova and L. Burteva. 2017. The digitization and presentation of the romanian transitional cyrillic script (ro.). *Akademos*, 1:24–29.

C. Diaconovici Loga. 1818. *Orthography or the correct spelling to guide Romanian language writers (Ro.)*. În Crăiasca Typografie a Universitatei Ungariei.

D. Gîfu and M. Plamada-Onofrei. 2017. Developing a technology allowing (semi-) automatic interpretative transcription. In *TDDL/MDQual/Futurity@TPDL*.

R. Halley. 2009. How good can it get? analysing and improving ocr accuracy in large scale historic newspaper digitisation programs. Last accessed 01 December 2022.

P. Iorgovici. 1799. *Observations on the Romanian language (Ro.)*.

IRS. 1997. *Information and documentation. Transliteration of Cyrillic characters into Latin characters. Slavic and non-Slavic languages (Ro.)*. IRS.

N. Miloni. 2020. *Automatic transcription of historical documents: Transkribus as a tool for libraries, archives and scholars*. Ph.D. thesis, Uppsala University.

Gr. Pleșoianu. 1828. *Primer to facilitate learning for children (Ro.)*.

I.H. Rădulescu. 1828. *Romanian Grammar (Ro.)*.

R. Smith. 2007. An overview of the tesseract ocr engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pages 629–633.

N.A. Ursu. 1960. The problem of interpreting romanian cyrillic texts written around 1800 (ro.). *Limba Română*, 3:33–46.

I. Văcărescu. 1787. *Observations or considerations on the rules and regulations of Romanian grammar (Ro.)*.

R. Wirth and J. Hipp. 2000. Crisp-dm: towards a standard process model for data mining. In *Practical application of knowledge discovery and data mining*, pages 29–40.