

**BRINGING THE OLD WRITINGS CLOSER TO US:  
DEEP LEARNING AND SYMBOLIC METHODS  
IN DECIPHERING OLD CYRILLIC ROMANIAN DOCUMENTS**

**DAN CRISTEA<sup>1,2</sup>, NICOLAE CLEJU<sup>3</sup>, PETRU REBEJA<sup>2</sup>, GABRIELA HAJA<sup>4</sup>,  
EDUARD COMAN<sup>5</sup>, ANCA VASILESCU<sup>5</sup>, CLAUDIU MARINESCU<sup>5</sup>  
and ANDREEA DASCĂLU<sup>5</sup>**

<sup>1</sup> Institute for Computer Science, Iași Branch of the Romanian Academy,  
2, Th. Codrescu Str., Iași, Romania

<sup>2</sup> Doctoral School of Computer Science, “Alexandru Ioan Cuza” University of Iași

<sup>3</sup> Department of Electronics and Telecommunications Engineering,  
“Gheorghe Asachi” Technical University of Iași

<sup>4</sup> “Alexandru Philippide” Institute of Philology and Literature, Iași Branch  
of the Romanian Academy

<sup>5</sup> Faculty of Mathematics and Computer Science,  
“Transilvania” University of Brașov

*dan.cristea@acadiasi.ro*

The paper addresses the problem of transliteration of scanned copies of old Romanian books written in the Cyrillic script into the Latin script. The motivation of this endeavor and attendees of such a technology are enumerated. Then, a number of peculiarities of these documents, which create difficulties for automatic processing, are exemplified. The proposed technology is presented in the form of a pipeline of modules, each applying AI or symbolic methods. Then, the component parts are discussed individually, and solutions are presented. The research is presented as work in progress, which leaves space for further enhancements. The data supporting training and evaluation of the modules is rooted in the former DeLORo project.

*Keywords:* written cultural heritage, old Romanian documents, transliteration of Cyrillic Romanian into Latin script, identification and recognition of graphical signs

## 1. INTRODUCTION

This article presents a new technology for automatically deciphering old Cyrillic Romanian documents, based on a combination of advanced image processing algorithms and machine learning techniques. The aim of the enterprise is to allow for a better understanding of the history and culture of Romania. The article discusses the methodology used, the results obtained until the present day, and the implications of the technology for the study of old Romanian documents written in the Cyrillic script.

The organization of the paper is as follows: the present section provides historical and linguistic data about the old Romanian writing in the Cyrillic script, as it was distributed in the Romanian provinces along the time. The motivation for this research and an enumeration of the possible users of the technology ends the section. Section 2 puts into evidence the major particulars that make automatic processing of old documents such a laborious and difficult task, necessitating departure from the classical optical character recognition (OCR) methods and adoption of artificial intelligence (AI) techniques. The section is very dense in commented excerpts of old documents. Section 3 presents the prerequisites of our contribution, first on a sketch of the overall architecture of a system working as a pipeline of modules, then presenting the hand-generated data that supports training of the neural models and finally briefly summarizing approaches similar to ours. Section 4 details one by one the component parts of the pipeline. A pre-processing phase dedicated to rectifying curved pages is followed by the recognition of the characters as graphical objects distributed on a page. An algorithm that segments the page into windows helps the neural model focus on only parts of the page, thus increasing detection accuracy. Then, the identified characters are classified into classes of Latin letters and finally they are linearized, and words of the old Romanian language are looked for in the decoded string of characters. The final Section 5 includes discussions and a presentation of envisioned future research.

#### 1.1. THE CYRILLIC ROMANIAN WRITING OVER TIME AND SPACE

The beginnings of the Cyrillic alphabet in the territories inhabited by Romanians are controversial. According to some scholars, they are said to date back to the X<sup>th</sup> century, while others believe that the introduction of the Cyrillic alphabet occurred three centuries later (Panaitescu 1965), (Gheție 1974). In any case, historical and political circumstances imposed the use of Slavonic as an official language in the royal courts of Moldavia and Wallachia, but also as a language of orthodox worship in all regions inhabited by Romanians, also including, in a generic sense, Transylvania (which encompasses Banat, Crișana and Maramureș regions). The cultural centers developed around the monasteries imposed the use of the Cyrillic alphabet (from various sources, *e.g.*, Bulgarian, Western Russian, Serbian) also in the writing of Romanian church texts, which circulated from one region to another, first in manuscript form then as printed volumes. Administrative needs required issuing of the first legal texts printed in the same alphabet. However, the resources available in the late Middle Ages and in the pre-modern period could not ensure large amounts of printed books, so that manuscripts continued to circulate in our country until the early decades of the XIX<sup>th</sup> century, unlike in the countries of Western Europe where printed books became widespread as early as the XV<sup>th</sup> century (Bianu 1907: V).

The documents preserved in libraries, personal, local and national archives, in Romania and abroad, are highly diverse in terms of content or stylistic framing of the texts, and especially as to the graphic form of the manuscripts. Subject to various influences, to the decisions of translators, copyists and authors to use one or other letter to render a sound or a group of sounds of Romanian words are so diverse that it is rather difficult to define general writing rules valid for a given region or a certain period. It was not until the end of the XVII<sup>th</sup> century that a trend towards standardization of the rules of writing can be observed, with the editing and publication of the complete text of the Bible in Bucharest in 1688. The first translation of the Septuagint into a vernacular language, by the Moldavian Nicolae Milescu Spătarul and revised by Greceanu brothers in Wallachia, was an important step in the unification and standardization of literary Romanian in its ecclesiastical version. In the areas of Catholic influence, texts written in the Latin alphabet prevailed (predominantly in Transylvania, but also in Moldavia until the 13<sup>th</sup> century). The spellings used are either Hungarian or Polish, and transcription and editing problems are present (Gheție, 1997), (Rosetti *et al.*, 1971/1961).

The work of cataloguing manuscripts written or copied in the territories inhabited by Romanians (more or less covered by the borders of Romania established after the Union of 1918), as well as old Romanian prints started late, only in the XIX<sup>th</sup> century, first atomized, at the level of some ecclesiastical or administrative institutions or of some localities, and then, towards the end of the century, remarkable personalities such as Gr. G. Tocilescu (Tocilescu *et al.*, 1886–1900), Eudoxiu Hurmuzachi (Hurmuzachi, 1887–1913), Nicolae Iorga (Iorga, 1901–1914), Ioan Bianu (Bianu *et al.*, 1903–1944), (Bianu *et al.*, 1907–1967) developed a systematic model for describing and indexing manuscripts and printed textual resources. Gabriel Ștrempel, trained under the guidance of Ioan Bianu, continued the work of cataloguing manuscripts carried out by Ioan Bianu and his team, in several volumes published until 1997 (Ștrempel, 1978–1992).

With its researchers and specialized institutes of the Section of Historical Sciences and Archaeology and the Section of Philology and Literature, the Romanian Academy has always played and continues to play a major role in the inventory and recovery of old Romanian documents, by launching projects dedicated to recording, describing, rendering or translating documents on the history of the Romanians (DRH A (1975–2006), DRH B (1966–2010), DRH C (1977–1985)), as well as through philological editing projects of old literary texts. The outcomes of these research projects provide the necessary resources for further studies in fields such as history, history of the Romanian literary language, history of literature and culture.

In the more than 150 years of life of the Romanian Academy, projects such as the above-mentioned ones have ensured the continuity of fundamental research in humanities, creating a respectable tradition of recovering the elements of heritage

that define us historically and culturally. The work founded at the end of the XIX<sup>th</sup> century has been continued to the present day by specialized researchers.

As evidenced by the bibliographical references of this paper, the main way in which the results of research are made available to the public is through the publication of classic printed volumes, with access limited by the small print runs and with a manner of editing texts for a specialized target audience. All collections of manuscripts and old printed texts (and we have confined ourselves only to the largest and best-known ones, without recording numerous other collections) here cited number tens of thousands of pages and record thousands of titles. An inventory of documents held in foreign libraries: apart from a small number, most of them receiving an extremely limited exposure (Cândea, 2011, 2012, 2014, 2016).

In the 2000s, a concerted effort has been made to digitize old textual resources and their editing in various academic centers in Romania. Among the first such projects are those initiated in Iași, materialized in the first volume of Romanian old texts, edited in Romania with specially constructed tools, exhaustively indexed and with an electronic version, published, initially, on DVD and online, in 2021 (Andriescu *et al.*, 2008). Volumes I–VI, IX of the *Monumenta linguae Dacoromanorum* series should also be mentioned here (1988–2003, 2005); they were funded and published by the “Alexandru Ioan Cuza” University of Iași, with the participation of researchers from the “Alexandru Philippide” Institute of Romanian Philology, and with the substantial support of Albert-Ludwigs University of Freiburg, Germany. Another important project of the “Alexandru Ioan Cuza” University of Iasi, aimed at digitizing old Romanian texts, is the *Electronic Corpus of The Old Romanian Texts (1521–1640) / Corpus electronic al textelor românești vechi (1521–1640)* (CETRV) coordinated by Professor Alexandru Gafton (Gafton, 2013–2016).

## 1.2. MOTIVATION FOR A TECHNOLOGY DEDICATED TO INTELLIGENT INTERPRETATION OF OLD DOCUMENTS

### **The difference between digitization and intelligent interpretation**

Digitization of the cultural heritage of a nation is meant at preserving on electronic media copies of its most representative artefacts. It embraces information created digitally or converted into digital form from the existing physical resources. If resources are born-digital, there is no other format except for the digital object but, in most cases, electronic copies should be created and preserved. This applies to textbooks, 3D reconstructions of architectural sites, records of musical interpretation, sculptures, paintings, etc., while the “digital materials include texts, databases, still and moving images, audio, graphics, software, and web pages, among a wide and growing range of formats. They are frequently ephemeral, and require purposeful production, maintenance, and management to be retained” (UNESCO, 2003). From this point of view, to digitize a written document means to scan and store the images

of pages as electronic documents, in order to preserve for future access and display the electronic version of the originals.

On the other hand, processing a document involves much more than that. It means to use the electronic version as an input artefact and to interpret the images of pages, decoding the content and intelligently looking inside it. This kind of knowledge, acquired and deciphered based on the digital form of an original object, could include structural constituents of the object (as given by its syntax), interpretation of the content (its semantics) and even practical usages (its pragmatics).

Therefore, the intimate knowledge over the original digital document, acquired through its syntactic, semantic, or pragmatic “readings”, allows for a wide range of post-processing operations to be applied to the original object, or to others that can be derived from the original one. To give one example, suppose one museum possesses a 3D reconstruction of an antiquity city, Jerusalem for instance. A mere electronic document would allow for a spatial view of buildings and streets of this city, but the distinction between them would be done only in the eye of the reader, the technology making no difference between churches and mosques, unless labels are hand-placed by curators on particular details. The artefact would not be able to distinguish streets from walls. As such, a virtual reality walk through the city would mean passing through the walls as they would be aerial. On the other hand, if the reconstruction is supported by a semantic interpretation as well (which allows for a distinction between how a street looks like and what a wall is like), and disposes of a pragmatic knowledge (which would contribute with the conceptual knowledge that physical entities cannot pass through walls and walking can be done only along streets), the user would be allowed to virtually “walk” through streets as in real life, and also to search for details which have not been annotated beforehand. One-time interrogations are thus possible out of unimaginable many others.

The next section will present several benefits of a technology, which, applied to old documents, can intelligently interpret parts of the old Romanian language writings.

### **The benefits of a knowledgeable technology: what should it bring and to whom?**

The development of an intelligent technology always starts from inventorying the goals: what do we intend to have and to whom should it be addressed?

In our case, the final goal of a technology would be to render, in an intimate understanding of the original Cyrillic Romanian content, the transliterated sequence of words, possibly offering notes to the reader that would signal and comment on special situations. So, the idea is not to display a sequence of decoded characters in the Latin script, but to produce instead a sequence of old Romanian language words. Unless specifically stated, no encroachment into the old form of spelling words is allowed; the technology is not supposed to “translate” the old language into the contemporary one.

Several types of users can benefit from such a technology, each having their possible demands. Let us note that the lists that follow do not exhaustively put in evidence any implementation – we are aware of this – but only represent behavior features which could become part of more or less complex systems.

The first category of users we have in mind is that of researchers, such as paleo-linguists – interested in studying old documents, or historians and archeologists – interested in collecting information about certain places and historical figures, or students – interested in enriching their knowledge from old books. For them, the technology should allow:

- transliteration in the Latin script, taking in consideration the context to interpret ambiguous Cyrillic letters;
- intelligent browsing of the document, displaying search features, as those supported by advanced corpora frontends. One such example is Korap, used in the queering of COROLA, the Corpus of Contemporary Romanian Language (Cristea *et al.*, 2019). This would allow, for instance, to search for all occurrences of flexed forms of a lemma in a collection of documents, the results expected being contextual windows of transliterated text;
- formation of parallel image-text collections, archived in the database, while keeping the original images of scanned Cyrillic pages paired with their Latin transcriptions; this could permit the display of fragments of images cut from their original page scans around the searched terms;
- generation of critical editions, as in (Macé *et al.*, 2019), (Seretan, 2020), which includes the ability to compare and align several editions of the same original document (possibly written at large time intervals, by different authors and in different places, sometimes even in different languages);
- *ad-hoc* generation of statistics, vocabularies, comparative studies, etc.

To editors, willing to edit, print and offer to the large public old Romanian books, therefore going further from simply reproducing images of original pages, the technology should also be possibly able:

- to offer help to interpret and index abbreviations, proper names, marginal notes, etc., linking them with the text or directing the reader towards secondary sources, for instance, by browsing the web, as in the Mapping Books project (Cristea and Pistol, 2014);
- in controlled situations, to produce, for instance, the modern equivalent of old, incomprehensible, word sequences from an XVIII<sup>th</sup> century book, in the form of notes that would make them readable to an uninformed contemporary reader.

## 2. WHY IS PROCESSING OF OLD DOCUMENTS DIFFICULT?

We will make in this section an inventory of the most frequent problems that can be faced in old documents. Part of these issues will be dealt with by our approach.

Old books are kept in special preservation conditions in libraries that offer well controlled and constant humidity, weak exposure to light, and lack of dust, as well as very limited access of researchers to touch them directly (usually intermediated by gloves and tweezers). Very old books sometimes present a poor physical condition, with damaged pages and even missing portions. There are frequent cases of dirty documents, spots, writing seen by transparency from the reverse page, or palimpsests – pages which have been reused by first deleting a previous writing. Figure 1 shows only some examples.



Fig. 1. Obturating spots and dirty or damaged pages in old documents

If digital cleaning of pages and reconstruction of very small portions from damaged pages is feasible where contexts can contribute with the missing information, guessing the missing words in largely damaged pages is questionable if scientific precision is needed. We mention here the AI generative technologies of our days, as GPT-4 (OpenAI, 2023), LaMDA (Thoppilan *et al.*, 2022), or LLaMA (Touvron *et al.*, 2023), capable of composing large spans of free text by learning from similar documents. However, no scientific certainty can be attributed to these artificial reconstructions.

Simplifying a lot, there are two types of writing in old Cyrillic Romanian documents:

- uncial and semiuncial manuscripts, in centuries XVI<sup>th</sup> and XVII<sup>th</sup> (Chivu *et al.*, 1978), and manuscripts with cursive letters and ligatures in centuries XVIII<sup>th</sup> and XIX<sup>th</sup>. *Scripta continua*, with no spaces between words, was used especially in the XVI<sup>th</sup> century;
- prints, between the XVI<sup>th</sup> and the XIX<sup>th</sup> (1862), with different matters from one printing house to another, from one century to another.

As one might expect, the greater the diversity of characters, the more difficult it is to automatically decipher these types of writing. Even for prints, which are supposed to be more manageable, the huge diversity of fonts often makes their interpretation difficult. To this, one should also add the variations in the set of letters, due to the continuous process of modernization of writing (and of the technical means of printing), and the gradual shift towards the Latin script, the most significant change in Romanian writing taking place between 1830 and 1862, when Latin characters were gradually introduced into writing, combined with the modern Cyrillic alphabet. This period is known as the transitional alphabet.

Of course, in the case of manuscripts, there are differences from one copyist to another, each with its own spelling, as in handwriting today, but also with personal rules for rendering the pronunciation of sounds and words in writing, because normative grammars appeared late in Romanian culture and, therefore, there was no uniform rule, unique for all centers of culture. There are also differences from one historical region to another, mainly generated by the differences in phonetics and lexical expressions, by the influence of spoken language upon writing and by the fact that, until the premodern period of Romanian culture (1780–1830), some marks of literary language regional variants can still be observed (Ivănescu, 2000/1980: 523, 567–594, 614–640).

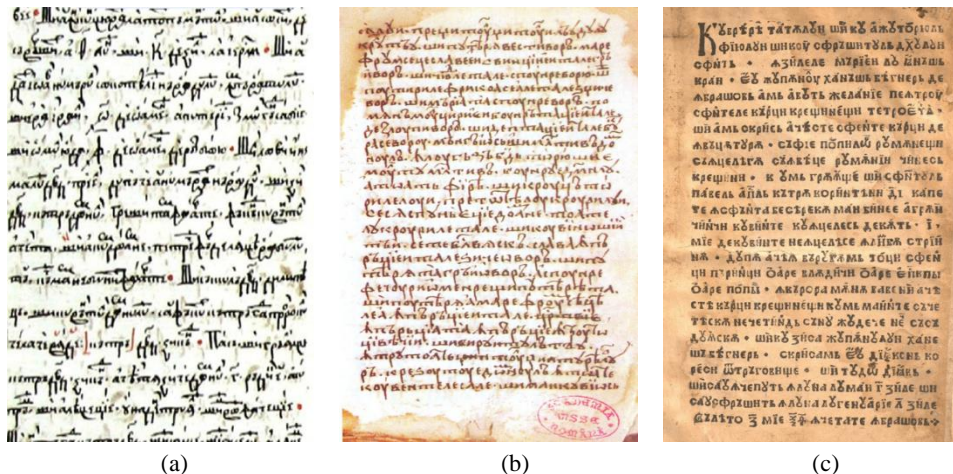


Fig. 2. Examples of different types of writing. (a) Manuscript (Ms. 45 – *Vechiul Testament*); (b) Uncial (Ms. 3077 – *Psaltirea Hurmuzaki*, XVIth century); (c) Printing (CRV 10 *Tetraevangheliar*, Braşov, 1561)

In the process of digitization, it is often the case that a scanned page remains with curvatures of lines due to the spines of thick books (for instance, the skew can be observed in Figure 3, more on the left page than on the right one). So, deskewing scans of curved pages is a challenge that has to be faced. The dedicated algorithm should be (or not) capable of realizing that a page presents skews, but the output should be similar in all cases.



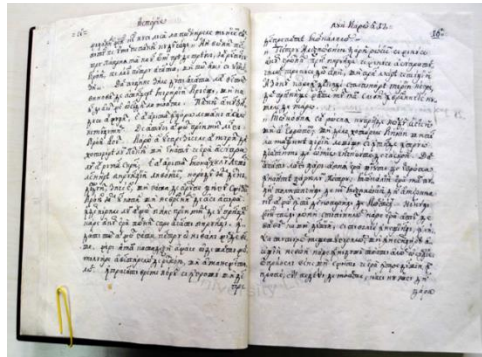


Fig. 3. Scan of a thick book: spines produce curvatures

In old manuscripts, the text is sometimes complemented by drawings, such as frontispieces or framing arabesques, large titles with drawn letters, ornate initial letters (see Fig. 4a). Marginal writing occurring occasionally contains indexes or notes that pair the content (as in Fig. 4a), where the numbers indicate verses. Interlinear writing was intensively used to manually correct mistakes made by copyists, usually to add missing characters (see Fig. 4b). Other difficult cases to interpret in relation with the content are signs with a value of modifiers (accents, tildes, accolades).

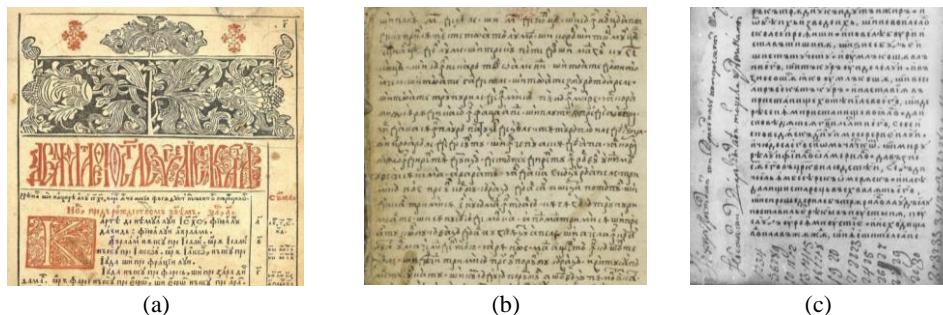







Fig. 4. Pages extracts: (a) from *Noul Testament de la Bălgrad*, 1648 – displaying large titles, frontispieces, drawn initial letters and marginal notes; (b) from *Hronograf*, ms.rom.3517, XVIIth century – displaying interlinear writing; (c) from *Psaltirea Voronețeană*, cca 1550 – displaying bilingual text (Romanian and Slavonic), interlinear text, *scripta continua* on the margin added by a different hand than the uncials of the main text, and aligned vertically

In such cases, the technology should be able to separate the text from the drawings, to interpret the ornate initial letters and decode the large titles, to find and relate the references from the margin inside the text and to correctly place in rows the isolated letters or the short sequences of letters written interlinearly.

The significance of many signs is contextual and should be decoded in the particular situations where they occur. This applies to letters (for instance, **ⲕ** = “ea” and **ⲕⲱ** = “ia”), as well as to special signs (for instance, the tilde sign

(~) is used both to mark abbreviations, as here:  “I[isus] h[ristos]”,  “H[ristos]”  “D[o]mnul”, and to change the significance of some graphemes: when placed above them, their meaning changes to numbers:  “200”,  “40”). In all cases, the automatic interpretation of the text should copy the expert’s knowledge in interpretative transcription, going down to recomposing the sequence of words from the sequence of characters, also in accordance with a dictionary, as would be, for instance, an Old Romanian lexicon (possibly organized diachronically and synchronically).

### 3. IN PREPARATION OF A TECHNOLOGY

#### 3.1. THE PHILOSOPHY OF THE WHOLE ENDEAVOUR

Our research is a follow-up of a two-year project (*DeLORo – Deep Learning for Old Romanian*, October 2020 – October 2022, see *Acknowledgements* for details), intended to create the computing infrastructure, to acquire digital and annotated data and to build up the processing components of a technology that would serve the goals announced in the introduction. The technology itself is designed as a pipeline, fed in input with images of scanned pages of old Romanian documents drafted in the Cyrillic script and producing in output the decoded text, as sequences of words of the old Romanian language, in the Latin alphabet (see Fig. 5). The architecture of this pipeline was designed in DeLORo, but the assembly of the modules as a system has never been done and, at the day of publishing this paper, it is still work in progress.

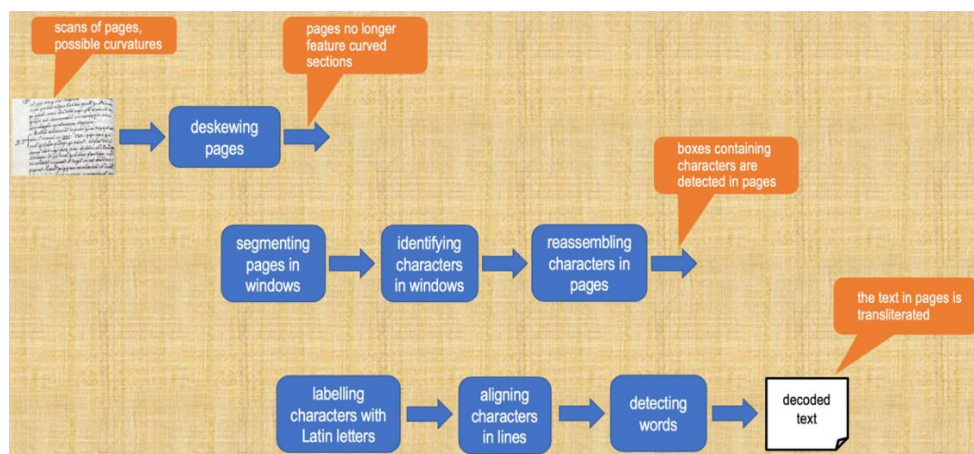


Fig. 5. The technology works as a pipeline of input-output modules, from (Coman *et al.*, 2023)

What Figure 5 hides is the support that the component modules have from external data, mostly acquired through manual annotation during the DeLORo project. The next sub-section details how this data has been acquired and stored, while Section 4 describes each module of this architecture. As mentioned in the title

of this paper, some of the component modules use classical methods, while the rest of them – deep learning methods.

### 3.2. ACQUISITION OF THE GROUND TRUTH

The external support of the modules that make up the pipeline in Figure 5 originates in a collection of 52 scanned Romanian books in the Cyrillic script, both in printed and uncial writings, together with their corresponding metadata (which marks: title, author, genre, place of publication, year, quality of the original document and the scan copy, etc.), originating from an interval that spans from the XVI<sup>th</sup> century until half of the XIX<sup>th</sup> century, and being of various quality conditions. This collection, totaling 16,864 images of original pages, was organized in a database called *Romanian Old Cyrillic Corpus – ROCC* (Cristea *et al.*, 2021). Many of these images include segmentation and content annotations to highlight a number of 202,042 graphical objects: rows, characters, marginal and interlinear writing, frontispieces, titles, ornaments, modifiers, initial ornate letters, etc., as well as the transliteration in context of a high number of rows and isolated characters (see Fig. 6). ROCC was acquired through extensive annotation by members of the DeLORo consortium, using a specially built frontend – the *Online Old Cyrillic Image Annotation Tool – OOCIAT* (DeLORo, 2022).

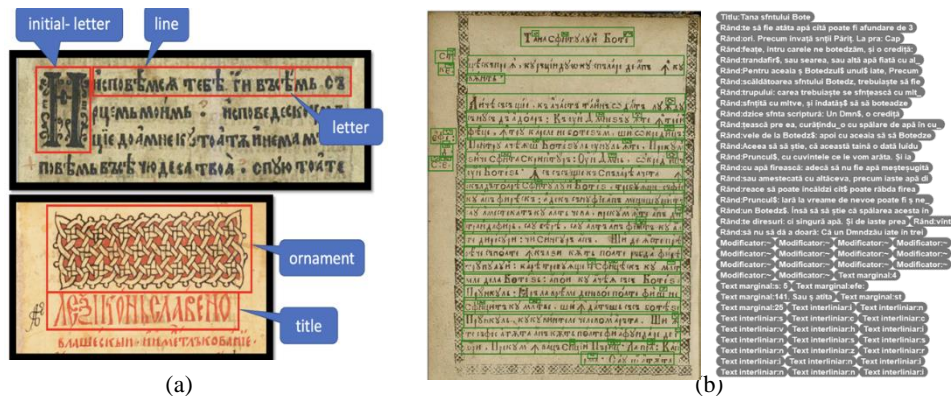


Fig. 6. The OOCIAT frontend: (a) Types of annotated objects; (b) One heavily annotated page with transliteration of lines

### 3.3. SIMILAR ENTERPRISES

Initially focusing on accurate identification of single characters in printed text, Optical Character Recognition (OCR) systems became widely available

when ABBYY<sup>1</sup> produced the first FineReader version in 1993. At the beginning, OCR systems presented rather poor effectiveness on complex scripts or scripts with ligatures (Blanke *et al.*, 2012). The OCR systems were limited in that characters had to be isolated and spatially separated, while processing of complex layouts and fonts was faulty (Cordell, 2017). Now-a-days, FineReader integrates intelligent document processing features and has become one of the major commercial products in content-centric processing. It leverages machine learning, natural language processing and computer vision techniques to decipher and correctly interpret the content of documents.

Tesseract<sup>2</sup> (Tesseract, 2021) and Kraken<sup>3</sup> (Kraken OCR, 2021) are other major commercial OCR packages, both embedding pre-processing steps.

The move from OCR to more advanced handwritten text recognition (HTR) technology coincided with the application of machine learning models, such as deep neural networks, to extract visual features and recognize characters and words in a segmented line of text *via* the calculation of overlapping probabilities. The lines of the text are accurately segmented and a wide range of glyphs can now be deciphered (Edwards, 2007). Like its OCR counterparts, HTR requires some manual intervention and training, yet lessens the need for full human transcription and bespoke recognition models developed at high cost.

Monk<sup>4</sup> (Monk, 2004), a tool developed at the University of Gröningen, is able to transcribe lines, broader zones for words and label them, allowing a scholar to speed up the process of indexing a documentation (Schomaker, 2020, pp. 226–227). Monk was more recently used to process Chinese and Arabic characters.

These tools are especially designed to process continuous writing, but perhaps the most known of all is Transkribus<sup>5</sup>, also an ABBYY AI-based system offering online and desktop transcribing services for historical documents. It decodes handwriting, even *scripta continua*, but the claim to work for any language induces the conclusion that it does not take into consideration lexical information, therefore it recognizes characters but does not assemble them in word forms of a specific vocabulary. Moreover, the documentation does not put into evidence the capacity to recuperate and place in sequence interlinear writing, and tests with a Cyrillic writing produced incomprehensible Latin strings.

---

<sup>1</sup> <https://www.abbyy.com/>

<sup>2</sup> <https://github.com/tesseract-ocr/tesseract>

<sup>3</sup> <https://medium.com/analytics-vidhya/unleashing-the-kraken-for-ocr-fba6bff73c8c>

<sup>4</sup> <https://www.ai.rug.nl/~lambert/Monk-collections-english.html>

<sup>5</sup> <https://readcoop.eu/transkribus/?sc=Transkribus>

A system that has very much in common with our project is the  $\mu$ Doc.tS platform (Pratikakis, 2021), (Tsochatzidis *et al.*, 2021). It decodes handwritten Old Greek documents, makes handwritten text recognition and keyword spotting. A consistent set of handwritten documents originating from the Stavronikita Monastery on Mount Athos has been recently made public, after decoding with the platform, for research purposes. The platform was trained on Greek, English, German and Finnish.

#### 4. PIECES OF THE TECHNOLOGY

##### 4.1. PRE-PROCESSING: REMOVING CURVATURES OF PAGES

Deskewing is the process of straightening images that are slightly tilted or skewed. In our case, written lines may look crooked towards an end, due to the scanning process. Straightening crooked rows is assumed to enhance the quality of the recognition process that follows.

The methods found in the literature that address the problem of deskewing range from simple heuristics to complex machine learning algorithms. Let us note that classical techniques widely used to detect the dominant lines in an image and compute their orientation, as those based on the Hough transform (Likforman-Sulem *et al.*, 1995), (Louloudis *et al.*, 2009) or the Radon transform (Deans, 1993), for instance, are not directly applicable for deskewing handwritten lines in scanned images, because the lines of text are not geometrical lines, and page scanning may curve the image, especially at the edges. However, they can be used as part of a more complex processing algorithm; for a survey of classical methods, see Louloudis *et al.*, 2009.

More recently, deep learning approaches have shown promising results in deskewing. Convolutional neural networks (CNNs) have been used to learn rotation-invariant features from images and estimate the rotation angle. One such method, called RotNet (Gidaris *et al.*, 2018), trains a CNN to predict the rotation angle of an image from four possible angles (0, 90, 180, and 270 degrees). Another method, called CornerNet (Law and Deng, 2020), uses a CNN to detect the corners of a document and estimate the rotation angle from their positions. End-to-end solutions have also been proposed in (Yan *et al.*, 2018) and (Li *et al.*, 2019).

For the sake of efficiency and simplicity, we opt for using a classical (non-deep-learning) method. Our approach in deskewing considers a chain of transformations described below.

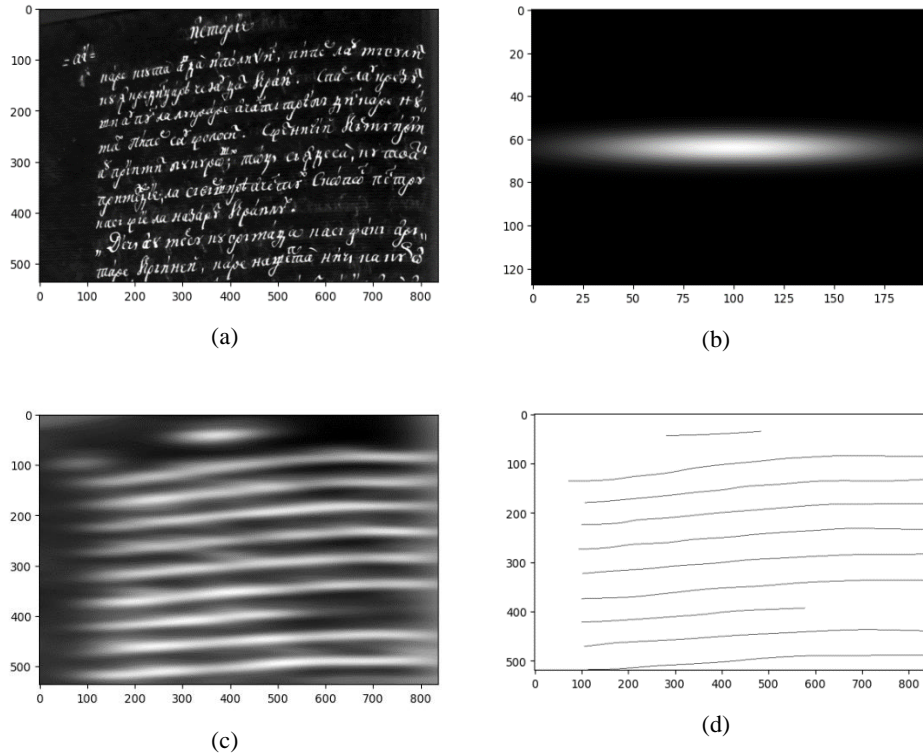


Fig. 7. Deskewing pre-processing steps: (a) Original image, inverted;  
 (b) Gaussian kernel used for low-pass filtering;  
 (c) Filtered image; (d) Detected rows

The deskewing approach consists of two major steps:

1. Detecting the lines of text in the input image;
2. Rectifying the image to make these lines horizontal.

For detecting the lines in the input image, we employ the following heuristics, illustrated in Figures 7 and 8. Starting from the inverted original image (see Fig. 7a), we apply a 2D Gaussian low-pass filtering, *i.e.*, we convolve with the kernel illustrated in Figure 7b. The filter is much more aggressive in the horizontal direction, the kernel being very elongated, in order to exploit the a-priori expectation that text lines are oriented more or less horizontally. Thus, the text lines are blurred horizontally, while the vertical separation between lines is maintained, and the lines are transformed into elongated blobs, as in Figure 7c.

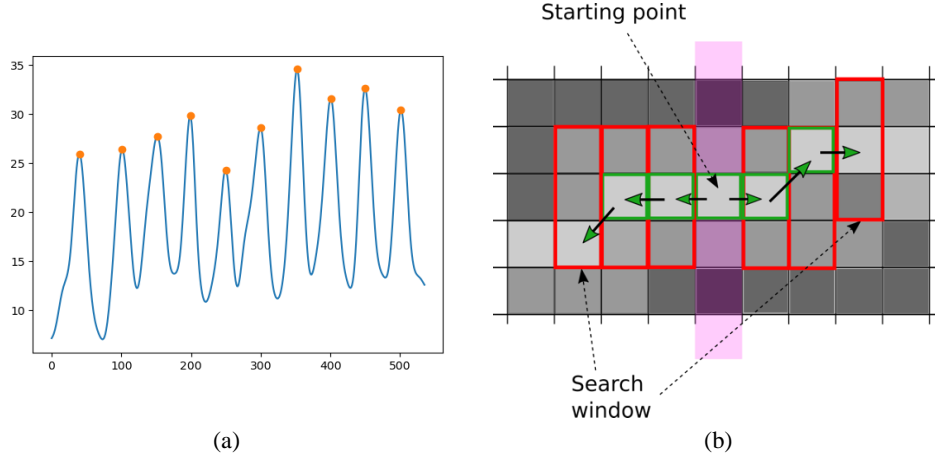


Fig. 8. Line detection: (a) histogram of a vertical band through the image, peaks correspond to lines; (b) ridge following strategy along the blurred lines

Next, we use a ridge following strategy to detect the line blobs, as explained in Figure 8. We pick some starting points by locating the peaks in a vertical cut across the center of the filtered image (see Fig. 8a), and we search for the largest pixel value (which is the level of white on a black-white scale) in an adjacent search window, which is then added to the path. We continue in this manner as long as the pixel values remain above a certain threshold, extending towards left and right (as in Figure 8b). We thus obtain a set of irregular paths, composed of adjacent pixels, which approximately follow the middle of the text lines. To improve robustness, we then use piecewise linear approximation to smooth these paths, effectively turning each path into a sequence of connected segments (as illustrated in Figure 7d). In the following, we refer to these detected paths as “ridges”.

The second step of deskewing is image rectification, where we seek to warp the image in such a way as to transform the detected ridges into horizontal lines. First, we compute the target horizontal lines, by aligning the points on each ridge horizontally. Note that we consider only the knots of the connecting segments that compose a ridge. We obtain therefore a set of matching points:

$$x_{in}, y_{in} \leftrightarrow f_1(x_{out}, y_{out})$$

where  $(x_{in}, y_{in})$  are points on the ridges detected in the input image, and  $(x_{out}, y_{out})$  are their target coordinates in the rectified image, horizontally aligned.

For the actual image dewarping, we implemented two approaches:

1. **Spline approximations.** We use bivariate cubic splines (Dierckx, 1995) to approximate the functions:

$$x_{in} = f_1(x_{out}, y_{out})$$

$$y_{in} = f_2(x_{out}, y_{out})$$

which map the output to the input coordinates of the points. By evaluating these splines in every point of the output image we can obtain the source pixel location for every position in the output image. The actual pixel values in the output image are computed by bilinear interpolation of the neighboring pixels of the source location.

An illustration of the two bivariate splines is provided in Figure 9, and the rectified image is shown in Figure 10.

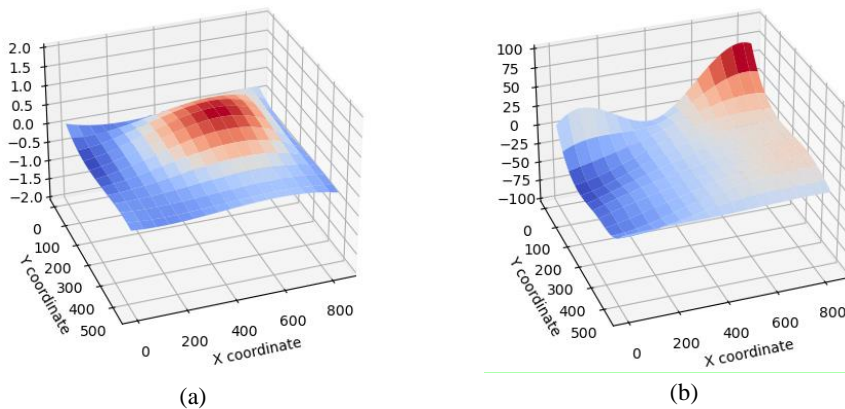


Fig. 9. Bivariate splines approximating the X-axis (a) and Y-axis (b) displacement of the source pixel location for every target location in the output image

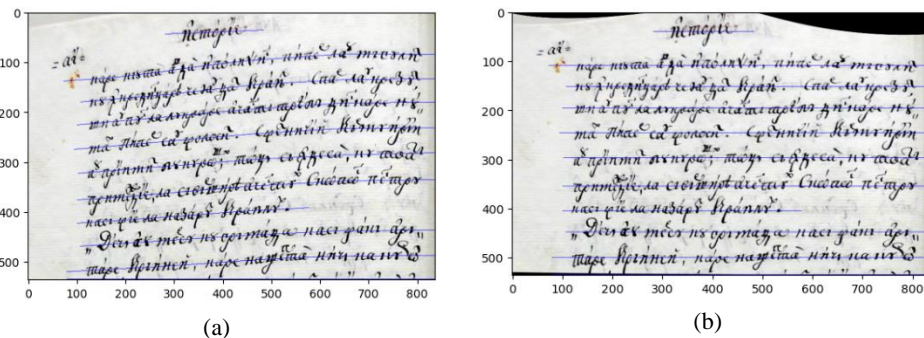


Fig. 10. Rectification using bivariate splines for pixel remapping:  
(a) Original image and detected ridges; (b) Rectified image

2. **Mesh decomposition and affine transformation.** We decompose the input image into a mesh of triangles, using Delaunay triangulation (Lee and Schachter, 1980) of each zone between two consecutive ridges. We obtain therefore pairs of matching triangles in the input and output images:



$$\{(x_1, y_2), (x_2, y_2), (x_3, y_3)\}_{in} \leftrightarrow \{(x_1, y_2), (x_2, y_2), (x_3, y_3)\}_{out}$$

Each of these triangle pairs defines an affine transformation, which describes the transformation of coordinates of the three vertices between the input and output images. Applying the transformation to an input triangle, we remap all the pixels from that triangle into their target location. Repeating this procedure for every triangle in the decomposition, we obtain the output image by stitching together the resulting output triangles (see Fig. 11).

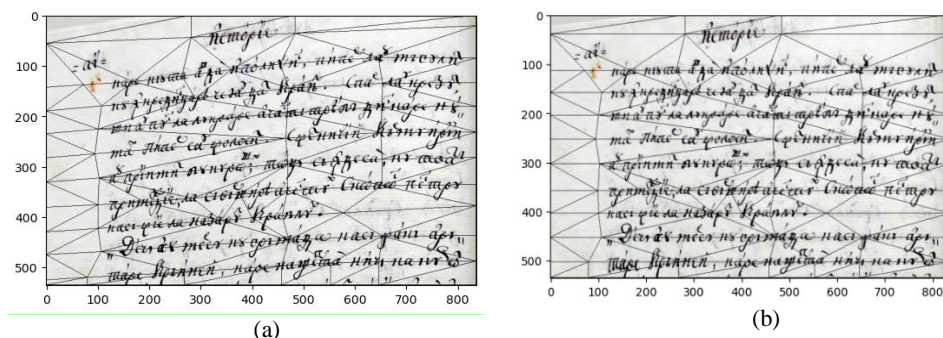


Fig. 11. Rectification using mesh decomposition and affine transformations:  
 (a) Original image and mesh decomposition; (b) Rectified image and mesh decomposition.

The implementation relies heavily on the Scipy library for mathematical approximations and OpenCV for actual image handling – available online as an open-source package<sup>6</sup>.

#### 4.2. IDENTIFYING CHARACTERS

As it could have been already evident from the examples presented, a classical OCR approach would prove ineffective for the tasks we envisage, especially because of the large diversity of types of objects to be identified in the scanned pages, and for the non-regular placement of these objects, each of them having a particular contribution to the overall understanding of the content. Of course, we are still dealing with texts, but the bidimensional placement of signs on a page takes us out from the idea of reading, seen as a linear, sequential process, to almost that of deciphering a picture in a bidirectional space. However, at the end of the automatic interpretation process, sequencing has to be recovered, because the output must still be made up of words placed in sequence.

Current approaches in this direction refer to an area of AI research in computer vision, which is usually called Object Identification: find a known object, more or less unobstructed, in a picture that displays an agglomeration of items of different forms. Our

<sup>6</sup> <https://github.com/nikleju/RowRectification>

goal is to identify and, where possible, label graphical objects in pages and to assemble an overall message out of their physical placement. In this section, we will focus on identifying graphical objects on a page that represent Cyrillic characters (either printed or hand-written, but well individuated, therefore showing no ligatures). These characters, as mentioned in Section 2, are placed mostly in rows, but often also in-between them.

We tried several techniques for identifying the characters in the scanned pages: cluster-based heuristics, Region-Based Convolutional Networks (Girshick *et al.*, 2016), and *You Only Look Once* (YOLO) (Redmon *et al.*, 2015). Out of these techniques, YOLO provided the best observable results.

A previous experiment has proved that the detection of written lines can be performed with a rather high confidence when using a self-paced strategy (Kumar *et al.*, 2010) implemented on a YOLO.v4 architecture (Găman *et al.*, 2022). Also, the attempt to use a Jetson architecture in order to recognize characters on the whole page, which in other contexts (Ciobanu *et al.*, 2022) proved efficient, had to be corrected by segmenting the page into  $224 \times 224$  pixels windows. Although the results improved a lot, even this attempt could not be positively finalized, but at least the experiment showed that, by reducing the area on which characters are searched, the trained neural network responds better. All these experiments made us propose and compare three possible strategies to choose the contexts in which characters should be detected and classified, detailed in (Cristea *et al.*, 2022): page segmentation (PS), row segmentation (RS), and window segmentation (WS). Let us note that a RS approach could be applied only when characters are well aligned in lines, because the line detection process could be seriously puzzled by the “noisy” characters left in-between the lines, as in situations like the ones evidenced in Figure 12.

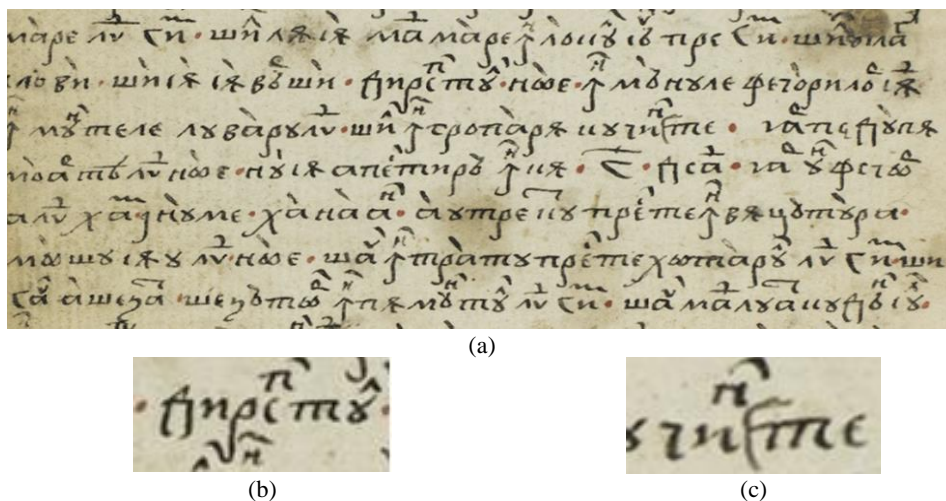
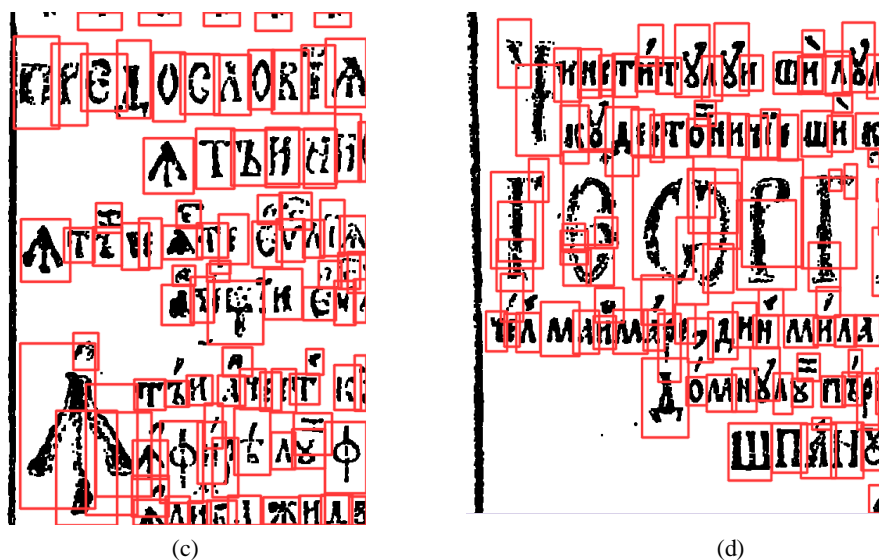


Fig. 12. Manuscript with a high density of interline signs: (a) The whole fragment; (b) A word with “noise” – letters of the lower line and sequences of letters of the upper line: *direptul* (“the righteous”) where *-p-* is written above *-e-* and *-l* above *-u-*; (c) A word with one letter added later by a different hand: *cinste* (“honour”), where *-n-* is written above *-i-*, and *-s-* is added

The results can be seen in Figure 13, on pages with different density of characters, displaying more or less interlinear writing, as well. As shown in these figures, the detection algorithm is still much sensitive to the size of letters (large letters, as those in Figure 13c and Figure 13d are multiply segmented) and to the binarization technique used (some letters are chopped in pieces, due to the disappearance of some black pixels, therefore introducing false characters). But, apart from these defects, the method gives good results, in the detection of interlinear letters included.



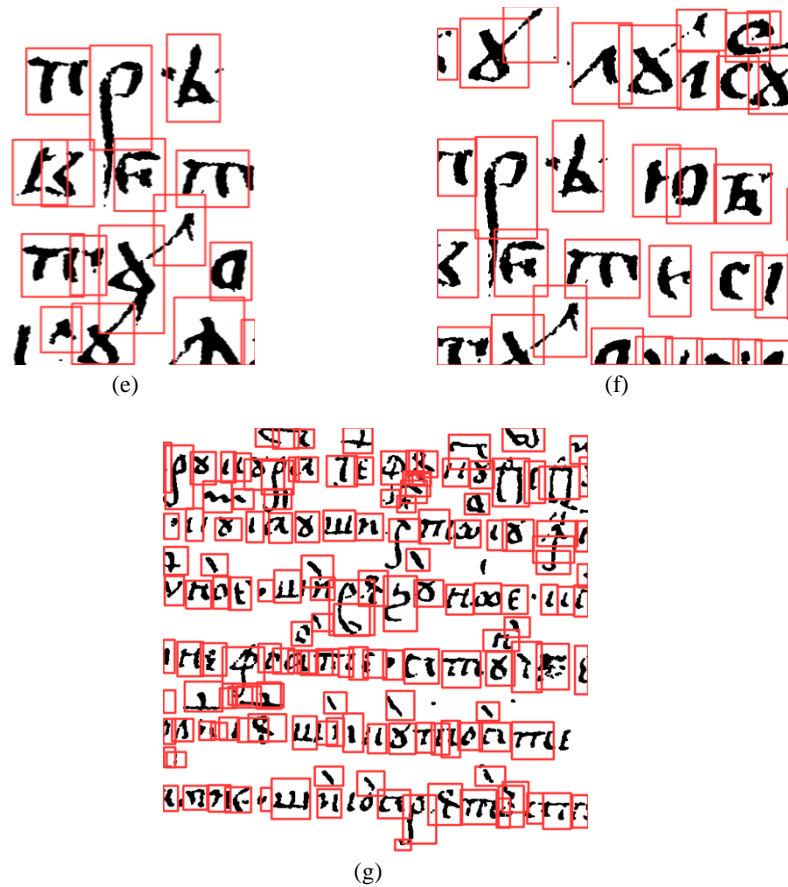


Fig. 13. Examples of pages displaying identified characters placed in bounded boxes: (a) High-density, uniform text; (b) Non-uniform, high-density text; (c) - (d) Non-uniform text with varying character size; (e) - (f) - (g) Uncial texts with interline characters

#### 4.3. CLASSIFYING CHARACTERS

After identifying the location of characters in pages, they can now be labelled. The Cyrillic alphabet used in Romanian writings includes 44 letters.

Given the variety of shapes a letter can have, as shown in Figure 13, the classifier (for which we have only preliminary results), besides discriminating between multiple letter classes, must learn to generalize for:

- **multiple shapes, and sizes of letters** – Figure 13.c provides an example of the different shapes and sizes for the letter  $\uparrow$  (transliterated as “în”, “im”, “i”);
- **skewed letters** – a skewed letter can appear due to the angle at which the page was scanned, and also to the transformations applied to the bounding box of the letter in order to impose uniform dimensions of the classifier input;

- **signal loss** – when applying preprocessing steps to window segmentation, some of letter’s pixels can be lost, as can be seen in Figures 13.c and 13.d, which does not only impact the classification of the letters but also their identification; an example of such loss can be seen in the top-left corner of Figure 13.d, where the letter Ч (transliterated as “ci”/”ce”) lost some of its pixels due to preprocessing, and the YOLO.v5 model identified it as two distinct letters.

In this section, we present a deep-learning solution for the classification of prints and handwritten (without ligatures) Cyrillic characters. The proposed solution leverages a Convolutional Neural Network (CNN) to capture relevant spatial information from input images, and includes a number of layers grouped in *feature extraction* (input layer, convolutional layer 1, max-pooling layer 1, convolutional layer 2, max-pooling layer 2, flatten layer) and *classification* (dense layer, output layer), as described below.

*Input layer:* the network accepts images of printed and handwritten Cyrillic characters, resized to  $28 \times 28$  pixels. The images are preprocessed, including normalization and conversion to grayscale, and then binarized using one of the techniques mentioned below (Grayscale, Fixed Threshold, Otsu's Method, or Adaptive Threshold).

*Convolutional layer 1:* consists of 32 filters, each with a  $3 \times 3$  kernel size, having applied a ReLU (Rectified Linear Unit) activation function to the output feature maps, introducing non-linearity and enhancing network's ability to capture complex patterns.

*Max-pooling layer 1:* performs  $2 \times 2$  max-pooling to reduce the spatial dimensions of the feature maps, thereby reducing the computational complexity and improving translation invariance.

*Convolutional layer 2:* uses 64 filters, each with a  $3 \times 3$  kernel size, and applies the ReLU activation function to the resulting feature maps.

*Max-pooling layer 2:* performs  $2 \times 2$  max-pooling, further reducing the spatial dimensions of the feature maps.

*Flatten layer:* reshapes the 2D feature maps into a 1D vector, which serves as input for the subsequent fully connected layers.

*Dense layer:* a fully connected layer with 256 neurons, employing the ReLU activation function to maintain non-linearity within the network.

*Output layer:* a fully connected layer with “num\_classes” neurons, where “num\_classes” corresponds to the number of distinct Cyrillic character classes. This layer utilizes the softmax activation function to produce class probabilities.

The network is trained using categorical cross-entropy loss and is optimized with the Adam optimization algorithm (Kingma and Ba, 2015).

The baseline architecture presented supports variations, which we have exploited in the search for the optimum model adapted to our needs. To improve classification performance, we explored various types of preprocessed input images,

and binarization techniques, and experimented by varying the depth of the model and filter sizes in the convolutional layers (the best results for each type of modification will be later showcased).

The following binarization methods have been used:

*Grayscale*: no binarization is applied to the input layer after preprocessing. Pixels have continuous values, with intensities ranging from 0 to 1 (after normalization).

*Fixed Threshold*: a global threshold value is applied to all pixels in the preprocessed grayscale image. Pixels with intensities above the threshold are converted to white (1), while those below are converted to black (0). This technique is simple and easy to implement but may not be optimal for images with varying lighting conditions.

*Otsu's Method*<sup>7</sup>: calculates an optimal threshold value based on the preprocessed grayscale image's histogram, minimizing the intra-class variance between the foreground and background pixels. It is an automatic thresholding technique that adapts to image's characteristics.

*Adaptive Threshold*<sup>8</sup>: calculates a local threshold value for each pixel in the preprocessed grayscale image based on the surrounding pixel intensities. As a result, it allows for different threshold values across the image and performs well in situations with varying lighting conditions or local image properties.

After applying the chosen binarization technique, the resulting binary or grayscale images are fed into the neural network for classification.

Varying the depths of the model and filters' sizes, we proceeded in two phases, as follows:

- 1) adding more convolutional layers – we created a modified version of the baseline model with an additional convolutional layer, which included 128 filters with a  $3 \times 3$  kernel size. This experiment aimed at investigating the impact of increased depth on network's performance. Deeper networks can potentially learn more complex features, leading to better classification performance;
- 2) applying different filter sizes – we experimented with varying filter sizes in the convolutional layers to analyze model's ability to capture different levels of detail in the input images. The modified model included  $5 \times 5$  and  $7 \times 7$  filter sizes. Larger filters can capture more contextual information and may be better at detecting larger patterns, whereas smaller filters may focus on finer details. It is essential to strike a balance between capturing both local and global patterns in the input images for effective classification.

Training and evaluation of the described sub-models have been performed on a dataset consisting of Cyrillic character boxes previously annotated with their

---

<sup>7</sup> <https://ieeexplore.ieee.org/document/4310076>

<sup>8</sup> <https://www.tandfonline.com/doi/abs/10.1080/2151237X.2007.10129236>

corresponding Latin letters. Before training the sub-models, the dataset was pre-processed to ensure compatibility with the neural network input requirements. This process may have included resizing the images to a fixed size, normalizing the pixel values, and applying any necessary data augmentation techniques.

To ensure a fair evaluation of the sub-models, the dataset was divided into three parts: training, validation, and testing. The training set is used to train the neural network, while the validation set is utilized during the training process to monitor model's performance and tune its hyperparameters. The testing set is reserved for the final evaluation of model's performance, ensuring an unbiased assessment.

The training process involved feeding the input images and their corresponding labels (the Latin letters) to the neural network. The network adjusts its weights during each epoch to minimize the loss function, which measures the difference between the predicted labels and the ground truth labels. The validation set is used to prevent overfitting and to determine when the training should be stopped, based on regular performance metrics such as accuracy, precision, recall, and F1-score.

Once the training process is complete, model's performance is evaluated on the testing set, helping determine how well the neural network generalizes to new, unseen data. The results of this evaluation, along with the results of other sub-models and their configurations, are summarized in Table 1, where we can identify the best-performing model or combination of techniques for our specific task.

Table 1

Comparison among sub-models

Binarization technique	Precision	Network architecture
Grayscale	0.9565	Initial architecture
	0.9461	Add conv layer 128 filters $3 \times 3$ kernel
	0.9555	Different filter sizes (best results)
Fixed Threshold	0.9279	Initial architecture
	0.9096	Add conv layer 128 filters $3 \times 3$ kernel
	0.9171	Different filter sizes (best results)
Otsu's Method	0.9430	Initial architecture
	0.9368	Add conv layer 128 filters $3 \times 3$ kernel
	0.9540	Different filter sizes (best results)
Adaptive Threshold	0.9526	Initial architecture
	0.9385	Add conv layer 128 filters $3 \times 3$ kernel
	0.9491	Different filter sizes (best results)

The above results provide some insights related to how the model reacts to different types of inputs and to slight modifications to its architecture. The model

tends to perform better on grayscale images or images binarized using Otsu's method or adaptive thresholding, leading to higher precision, compared to fixed threshold binarization.

Adding an additional convolutional layer with  $3 \times 3$  kernels seems to slightly decrease model's performance, as the precision is generally lower compared to the initial architecture, while experimenting with different filter sizes in the convolutional layers can lead to improvements in model's performance. The best results were achieved using grayscale images and Otsu's method for binarization.

Overall, the model performs well, particularly when using Grayscale or Otsu's method. Further experimentation with the architecture, preprocessing techniques, and hyperparameters could potentially lead to even better performance.

In future experiments, we plan to evaluate the performance of other well-known architectures, such as ResNet (He *et al.*, 2015) and VGG (Simonyan and Zisserman, 2014). These architectures, widely used in various computer vision applications, are known for their great performance in image classification tasks (Brownlee, 2020).

ResNet has residual connections that allow for the training of much deeper networks without facing the vanishing gradient problem. VGG, on the other hand, has a very simple architecture with small  $3 \times 3$  convolution filters, which can be beneficial in our case, as it may reduce overfitting.

Additionally, we are also exploring the possibility of implementing a Convolutional Recurrent Neural Network (CRNN) (Shi *et al.*, 2015) well-known for its ability to recognize patterns in sequential data, which is important for handwriting recognition tasks. The recurrent layers in the CRNN are able to capture the temporal relationships between the characters, allowing the model to recognize the patterns in the sequence the characters are written, thus helping in interpreting ambiguous characters.

CRNNs can not only learn from past patterns, thus exploiting regularities in writing styles, but can also adapt to different handwriting styles, catching variations in the way characters are written. This adaptability is especially important for handwriting recognition tasks, where the variability in the handwriting styles of different people is huge.

As said, this is work-in-progress. We will continue experimenting with various architectures, such as ResNet, VGG, and CRNNs with attention mechanisms, in search of the most reliable architecture, offering stable results to the problem of deciphering and transcribing handwritten Cyrillic characters to the Latin script.

Figure 14 shows a decoded window. Due to signal loss, some of the letters are misclassified, as can be seen on the first row, where the letter “c” is split in two shapes and its right side is interpreted as an “s”. The correct transcription of the first row is “curvie”, of the second is “nu le-au”, and of the third is “eu pa”. The last row contains only segments of letters, therefore no prediction should be taken in consideration. So, out of the 15 character boxes, 4 are wrongly transcribed.





Fig. 14. An example of a window after applying the letter classification model

#### 4.4. PUTTING CHARACTERS IN SEQUENCE AND FORMING WORDS

As shown in Figure 12, characters are not always linearly aligned in rows. This picture, but also the one illustrated in Figure 15, shows that, sometimes, rearrangement of characters in lines is not evident.

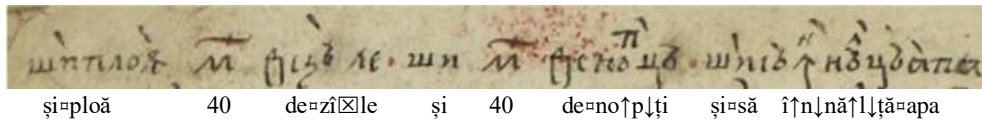


Fig. 15. An original line of text (from Hronograf, page 2, verso, DeLORo file 002v.jpg), with the original line (up) and the transliterated one (down); the notations used here are:  
 ↑ = the next character is placed above the line; ↓ = the reading continues in the line; = = a space is missing; ⊠ = a space should be removed

In this section we propose an approach for the linearization of characters in strings of words. Same as the preceding section, the work described here is still work-in-progress, as our models have not been fully tested yet. The following questions should be answered, in the following sequence:

- (q1) How many rows of text exist on the page?
- (q2) Which is the row each character belongs to?
- (q3) What is the position each character should be placed on its row?
- (q4) Where are word boundaries in each row?

To answer (q1), the vertical positions of each row should be estimated. As suggested by Figure 16, we do this by extracting the histogram of pixels on the vertical axis of the page, going downwards (our tests proved that counting not pixels but centers of gravity of the already detected character boxes statistically produces a less accurate result). Then, we approximate the positions of rows on the peaks of the histogram.

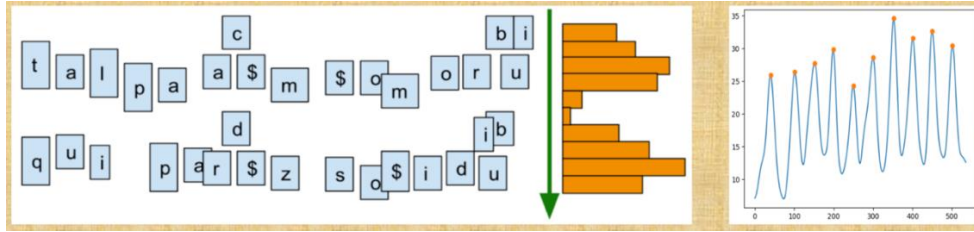


Fig. 16. Approximate the positions of rows by picks of a histogram

Once the axes of the rows are known, it remains to place on them, in sequence, the decoded characters and then to form words out of the sorted characters. It is clear that in answering (q2), the vertical coordinate of the center of gravity of character's box counts more than its horizontal one, in answering (q3), the horizontal position of character's box counts more than its vertical one, and, in answering (q4), the horizontal distance between neighboring characters is of importance. We say that one coordinate "counts more" than the other and not that it is the only one that counts to account for cases of characters placed in rows and in an evident sequence, but which display long tails under their own row, which could negatively balance the decision. Of course, in the great majority of cases, characters placed interlinearly should be attached to the line under them, not to the one above them, but, as explained before, sometimes the decision whether a character is interlinear or not is not straightforward. Finally, of decisive importance is how do the characters form words.

For answering (q2), suppose we pick up a character placed between two lines (as shown, given by the peaks of the histogram in Figure 16). The distance between the center of gravity of this character's bounding box and the two lines must favor its belonging to one row or the other, but the "attraction force" clearly has different strengths. Indeed, an object which is very close to one line has to belong to that one but, the more we depart from it, the stronger is the "attraction" towards the line below than towards the line above. One way to express this is by an equation like this:

$$F = \frac{w_1}{(y_c - y_l)^2 + \delta} * \left\{ (y_c - y_l) > 0 \Rightarrow w_2; (y_c - y_l) = 0 \Rightarrow 1; (y_c - y_l) < 0 \Rightarrow \frac{1}{w_2} \right\}$$

where  $w_1$  and  $w_2$  are positive sub-unitary weights summing up to 1,  $y_c$  and  $y_l$  are: the vertical coordinate of the center of gravity of the character box ( $c$ ) and the vertical coordinate of the line ( $l$ ), respectively. The first factor of the product expresses a force of attraction of the character by a line (as in the law of universal attraction, being inversely proportional with the squares of the distances). The very small entity  $\delta$  should prevent a null nominator when the character  $c$  is exactly aligned with the

line  $l$ . Then, this expression is corrected by a factor that is dependent on the position of the character with respect to the line: if  $c$  is placed under  $l$ , then  $y_c > y_l$  and the correction is sub-unitary, weakening the force; if  $c$  is placed above  $l$ , then  $y_c < y_l$  and the correction is supra-unitary, increasing the force; and if  $c$  is placed exactly on  $l$ , then  $y_c = y_l$  and there is no correction. The decision whether  $c$  should be attached to  $l$  or not can be negotiated by comparing  $F$  against a threshold  $T$ : if  $F \geq T$  then  $c \in l$ , otherwise  $c \notin l$ . Then, the weights  $w_1$  and  $w_2$  and the threshold  $T$  are determined during a calibration process that reproduces the actual belonging of characters to the line above or the one under, as given by an annotated set of ground truth examples.

For answering (q3), the formation of words should be the main criterion. Therefore, a look around in the context given by the neighboring characters in order to look for word matches with the entries of a lexicon, should be considered. The brute force alignment, which can be taken as a baseline, would sort the characters that we now know as belonging to a line in the order of their horizontal coordinates of the centers of the bounding boxes. In cases when short sequences of characters are placed interlinearly, this ordering could be however misleading. And let us also add to this already complicated equation some more issues: missing or false white spaces, abbreviations, characters not recognized or recognized with low confidence, and false characters (spots or remains of large characters which have misled the segmenter, as the ones noticed in Figure 13). If, in most of the cases, spaces indicate word boundaries, in some cases they are missing, while in others their existence does not necessarily imply a word boundary (see Fig. 15 for some examples). Moreover, abbreviated words are not found in a normal dictionary, which makes it necessary for a dictionary especially dedicated to abbreviations. Characters which cannot be classified or are classified with a very low confidence score produce empty boxes (marked with the '\$' sign in Figure 16). The character classifier could be designed to output more than one value in cases of low confidence, and to postpone the decision on the recognition of characters to the word recognition module.

The lexical confronting criteria are suggested by the reality that has to be admitted, namely that our lexicons of the old variants of language are incomplete (because the experimental vocabulary we are working with misses many word forms and because the old language had no writing rules, was not standardized, therefore no paradigmatic procedure that systematically generates all forms is available). When no match against a word of the old language has been recognized, it can be looked for even in the lexicon of the present-day language, choosing the closest word there (with respect to some lexical distances, Levenshtein (Adhitama *et al.*, 2014), for instance), on the ground that a native speaker of the contemporary language, even without expert linguistic knowledge, is still able to recognize words of the old language. This confrontation against a contemporary language dictionary should be

done, however, without trying to update the old writing. This is not allowed, because we do not modernize the language of the old documents<sup>9</sup>.

The discussion above indicates that the problem of linearization of characters and formation of words, therefore the answers to questions (q3) and (q4), can be drafted in a learning methodology, which would necessarily take into consideration, with different saliences: a) the vertical position of the character with respect to the neighboring rows; b) the vertical, horizontal, or Euclidean distances between neighboring character boxes; c) the constituency of a sequence of characters in a word that belongs to an old language dictionary; d) a lexical distance (for instance, the Levenshtein distance) of that combination of characters against a word that belongs to an old language dictionary; e) in extremis, even its lexical distance from a word of the contemporary language; f) the possibility that more than one content (label, *i.e.*, Latin letter) be determined for one character box, each with a different confidence score, but none above an accepted threshold.

Going a little deeper in the issue of word segmentation, one possible approach could be a variation of an (open source) splitting algorithm<sup>10</sup>, fed with a list of words as an external resource. This list is assumed to be sorted based on the frequency of the occurrence of the words in a corpus of deciphered old Romanian texts. With the aid of this list, a dictionary of words is generated, their corresponding cost being calculated as the logarithm of the word index in the dictionary (the rank of the frequency). This formula is based on Zipf's law (Thurner *et al.*, 2015), which describes the relationship between the frequency of a word in a given language and its position in the frequency table. Then, the input string, which lacks spaces, is divided into individual words by finding the best possible match at each position, iteratively. A list of costs is created for the input string, with each position's cost indicating the cost of the best match for that position. The best match at each position is defined as the word with the lowest cost from the dictionary, which ends on the character at the current position and may begin with any of the characters from the preceding positions.

The approach for determining the best match for each position iterates backwards from the current position, considering all potential word matches ranging in length from 1 to the maximum word length in the dictionary. It calculates the cost of each match by adding the cost of the prior position to the cost of the current matched word from the dictionary. Subsequently, it returns a pair structure (*co*, *le*), where *co* signifies the minimal cost discovered for the best match and *le* represents the length of the match. To retrieve the list of words, the algorithm conducts a

---

<sup>9</sup> However, as a side effect of the technology and for the sake of certain applications, the output string of words could be paired with a modern transcription and/or a string of lemmas of a contemporary dictionary. On the same level of considerations, once found in a dictionary of abbreviations, the non-abbreviated forms could also optionally complement the output.

<sup>10</sup> <https://stackoverflow.com/questions/8870261/how-to-split-text-without-spaces-into-list-of-words>

traversal over the input string, from right to left, at each iteration, removing from the string the word that is deemed to be the best match for the current position.

Although this provides good outcomes at first glance, the algorithm relies on a frequency dictionary that, in our case, has been generated from a limited amount of deciphered old Romanian texts. Consequently, there is a high possibility of encountering new words, which would not be recognized, that could affect the output in ways that are yet to be addressed. To achieve better results, the algorithm requires a frequency dictionary that encompasses a vast collection of old documents. Or, as suggested before in this section, one way to tackle this problem would be to approach the old Romanian language as not entirely distinct from the contemporary language. As such, when encountering a possible word that does not appear in the dictionary, it can be compared against a dictionary of the contemporary language, to see if it can directly be found there or, if not, whether it has a degree of similarity to a word that exists in the contemporary language above a certain threshold. Let us note that the lack of data to form a reliable dictionary is not an issue, anymore. However, this approach raises another concern, which is similar to the challenge of guessing words based on a partial match in the dictionary.

Finally, the matching algorithm suggested raises another problem in the case of words that include in their string of characters two or more smaller words, all existing in the dictionary. The algorithm may mistakenly identify the smaller words in the resulting split text, because the individual words may have a higher frequency and therefore a better cost compared to the concatenated longer word (for instance, in the current word list used by the algorithm, the word “cineva” ranks lower than the words “cine” and “va”).

## 5. DISCUSSIONS AND FUTURE WORK

In this paper we described the technological components of a platform designed to decipher prints and handwritten documents without ligatures (uncials) containing well individuated Cyrillic characters occurring in old Romanian language. Such a technology is expected to help the curation of old Romanian writings and to be used for scientific, didactic, and editorial purposes. Among many possible applications, we mention keyword search in a collection of handwritten documents, reverse search from text into image, different operations applied to objects inside a page, helping the editor for an intelligent display of the content.

The modules that make up the pipeline use a mixture of statistical and AI models. We begin by presenting a host of problems observable in old documents (dirty or obtruded pages, curved rows of writing, diversity of fonts and types of writing, diversity of graphical objects, interlinear and marginal writing, etc.). Then the proposal for an overall architecture is shortly described and the component

modules are outlined. Where possible, once the whole enterprise, as a coherent system, is still under development, we presented results and evaluation details.

Our approach is a follow up of a project that laid the foundations for this research by creating the technological infrastructure, designing the structure of a database and populating it with a considerable number of expert annotations, which have been used for training and evaluation purposes (ROCC).

A lot more has still to be done to reach our expectations. First, the modules that are only sketched (those described in Section 4.4, doing linearization of characters and formation of words) should be fully implemented and calibration experiments should be thoroughly organized on them. Then, all modules need careful testing, evaluation and refining iterations applied, until the best possible results are obtained.

Further envisioned developments include: extending the character recognition module to decipher characters of the "transition alphabet", used in Romanian provinces in the middle of the XIX<sup>th</sup> century, building POS-taggers and lemmatizers for old Romanian, with particularizations for diachronicity and synchronicity, for example answering questions like: "How can I find occurrences of the word *haină* in the 12 novels from the XIX<sup>th</sup> century included in the collection of *Astra Data Mining* of the journal *Transilvania Sibiu*?"<sup>11</sup>.

**Acknowledgments.** We mention the project "Artificial Intelligence Models (Deep Learning) Applied in the Analysis of Old Romanian Language (DeLORo - Deep Learning for Old Romanian)" PN-III-P2-2.1-PED-2019-3952, no. 400PED: *Deep Learning for Old Romanian* (2020–2022), official pages at <http://deloro.iit.academiaromana-is.ro/>, whose follow-up should be considered the developments described in this paper. We thank the members in the project from the Institute of Computer Science, the Institute of Philology - Iași branch of the Romanian Academy, and from the Faculty of Mathematics and Computer Science of the hrs. University of Bucharest<sup>12</sup>.

We thank the Romanian Academy Library in Bucharest for offering scans of old books and metadata, especially Mrs. Gabriela Dumitrescu and Marilena Bănică.

We thank assoc. prof. dr. Roxana Vieru and her students from the Master in Paleolinguistics, Faculty of Letters, "Alexandru Ioan Cuza" University of Iași, who contributed with hundreds of hours of manual annotation to acquire data used to train the technology.

**Authors' contributions:** Dan Cristea ([dan.cristea@acadiasi.ro](mailto:dan.cristea@acadiasi.ro)) designed the methodology and wrote a large part of the paper, Nicolae Cleju ([nclaju@etti.tuiasi.ro](mailto:nclaju@etti.tuiasi.ro)) designed, implemented and tested the technology of deskewing pages, and wrote Section 4.1, Petru Rebeja ([Petru.Rebeja@info.uaic.ro](mailto:Petru.Rebeja@info.uaic.ro)) designed, implemented and tested the technology of character localization and wrote Section 4.2, Gabriela Haja ([gabi.haja@acadiasi.ro](mailto:gabi.haja@acadiasi.ro)) wrote Section 1.1 and a large part of Section 2, Eduard Coman ([eduard.coman@student.unitbv.ro](mailto:eduard.coman@student.unitbv.ro)) designed, implemented and tested the technology of character recognition and wrote Section 4.3, Anca Vasilescu ([avasilescu@unitbv.ro](mailto:avasilescu@unitbv.ro)) supervised students' work and did the majority of formatting operations of the paper, Claudiu Marinescu ([claudiu.marinescu@student.unitbv.ro](mailto:claudiu.marinescu@student.unitbv.ro)) did many experiments of pages deskewing and partially implemented the linearization algorithm and Andreea Dascălu ([andreea-a.dascalu@student.unitbv.ro](mailto:andreea-a.dascalu@student.unitbv.ro)) proposed the algorithm of word formation and wrote part of Section 4.4.

<sup>11</sup> The query was suggested by Camelia Lăncrănjan-Mila, in the line with her PhD research.

<sup>12</sup> The names can be found in the official pages of the project, at <http://deloro.iit.academiaromana-is.ro/>.

## R E F E R E N C E S

- (Adhitama *et al.*, 2014) ADHITAMA, P., KIM, S.H., and NA, I.S. (2014). *Lexicon-Driven Word Recognition Based on Levenshtein Distance*. International Journal of Software Engineering and Its Applications, **8** (2), 11–20.
- (Andriescu *et al.*, 2008) ANDRIESCU, AL., HAJA, G., and MIRON, P. (coord.) (2008). *Monumenta linguae Dacoromanorum. Biblia 1688. Pars VI. Regum I, Regum II*, Iași, Editura Universității „Alexandru Ioan Cuza”, 2008, 560 p. + DVD. Autorii volumului: Tamara Adoamnei, Mădălina Andronic, Mioara Dragomir, Gabriela Haja, Elsa Lüder, Paul Miron, Alexandra Moraru, Mihai Moraru, Adrian Muraru, Veronica Olariu, Elena Tamba Dănilă. Consultant științific: Eugen Munteanu. Formatul electronic al volumului, pe suport DVD, a fost realizat de Vlad-Sebastian Patraș. ISBN 978-973-703-399-4; publicat online, pe platforma <https://solirom.ro>, august 2021, <https://biblia1688.solirom.ro/7/>.
- (Bianu *et al.*, 1944) BIANU I., HODOȘ N., and SIMIONESCU D. (1903–1944). *Bibliografia românească veche. 1508–1830*. Tom. I–V, Edițiunea Academiei Române, București, 2490 p.
- (Bianu *et al.*, 1967) BIANU, I., CARACAȘ, R., NICOLAIASA, G., and ȘTREMPEL, G. (1907–1967). *Catalogul manuscriptelor românești. I–IV. I: Numerele 1–300*. Biblioteca Academiei Române. Întocmit de Ioan Bianu. București, Edițiunea Academiei Române, Inst. de Arte Grafice Carol Göbl. S-sor, Ioan St. Rasidescu, 1907, VIII + 746 p.; II: *Numerele 301–728*. Biblioteca Academiei Române. Publicat de Academia Română. Întocmit de Ioan Bianu și R. Caracaș. București, Socec & Comp. și C. Sfetea, 1913, 666 p.; III: *Numerele 729–1 061*. Biblioteca Academiei Române. Publicat de Academia Română. Întocmit de Ioan Bianu și G. Nicolaiasa. Craiova, „Scrisul Românesc” S. A., 1931, 658 p.; IV: *Catalogul manuscriselor românești*. Întocmit de Gabriel Ștrempel, Fl. Moisil, L. Stoianovici. București, Editura Academiei Republicii Socialiste România, 1967, 703 p.
- (Blanke *et al.*, 2012) BLANKE T, BRYANT M, and HEDGES M (2012) *Open source optical character recognition for historical research*. J Doc 68(5):659–683. <https://doi.org/10.1108/00220411211256021>
- (Brownlee, 2020) BROWNLEE, J. (2020) *Deep Learning for Computer Vision. Image Classification, Object Detection, and Face Recognition in Python*. 2020, Jason Brownlee. <https://machinelearningmastery.com/deep-learning-for-computer-vision/>
- (Cândea, 2011, 2012, 2014, 2016) CÂNDEA, V. (2011, 2012, 2014, 2016). *Mărturii românești peste hotare*. Serie nouă. Vol. I–VI. I: Vol. I–IV, București, Editura Biblioteca Bucureștilor, 2011, 2012. Vol. V–VI. I, București, Editura Academiei Române, Editura Muzeului Literaturii, 2014, 2016.
- (Chivu *et al.*, 1978) CHIVU, Gh., GEORGESCU, M., IONIȚĂ, M., MAREȘ, A., and ROMAN-MORARU, A. (1978). *Documente și însemnări românești din secolul al XVI-lea*. Text stabilit și indice de Gheorghe Chivu, Magdalena Georgescu, Magdalena Ioniță, Alexandru Mareș și Alexandra Roman-Moraru. Introducere de Alexandru Mareș. Universitatea București. Institutul de Lingvistică. București, Editura Academiei Republicii Socialiste România, 1979, 498 p. [Documents between 1521 and 1600].
- (Ciobanu *et al.*, 2022) CIOBANU, A., LUCA, M., VULPOI, R.A., BĂRBOI, O., and DRUG, V.L. (2022). *Deep Learning in Colonoscopies: Improving Small Polyps Recognition Rate*, The 10th IEEE International Conference on E-Health and Bioengineering – EHB 2022, Romania, 17–18 nov. 2022. IEEE.
- (Coman *et al.*, 2023) COMAN, E., DASCĂLU, A., MARINESCU, C., REBEJA, P., VASILESCU, A., CLEJU, N., HAJA, G., and CRISTEA, D. *Bringing the Old Writings Closer to Us: Deep Learning in Deciphering Cyrillic Romanian*. Online communication at SMART-2023, Timișoara, 11 April, 2023.
- (Cordell, 2017) CORDELL R (2017) *Q i-jtb the Raven: taking dirty OCR seriously*. Book History John Hopkins University Press 20:188–225. <https://doi.org/10.1353/bh.2017.0006>

- (Cristea and Pistol, 2014) CRISTEA, D., and PISTOL, I.-C. *MappingBooks: Linguistic Support For Geographical Navigation Systems*. In Mihaela Colhon, Adrian Iftene, Verginica Barbu Mititelu, Dan Cristea, Dan Tufiş (eds.) Proceedings of the 10th International Conference “Linguistic Resources And Tools For Processing The Romanian Language, Craiova, 18–19 September 2014”, „Alexandru Ioan Cuza” University Publishing House, ISSN 1843-911X, 2014, pp. 189–198.
- (Cristea et al., 2019) CRISTEA, D., DIEWALD, N., HAJA, G., MĂRĂNDUC, C., MITITELU, V.B., and ONOFREI, M. *How to Find a Shining Needle in the Haystack. Querying CoRoLa: Solutions and Perspectives*. In Revue Roumaine de Linguistique (Romanian Review of Linguistics), vol. 64, nr.3, Publishing House of the Romanian Academy, ISSN: 0035-3957, 2019.
- (Cristea et al., 2021) CRISTEA, D., PĂDURARIU, C., REBEJA, P., SCUTELNICU, A., and ONOFREI, M. *Data Structure and Acquisition in DeLORo – A Technology for Deciphering Old Cyrillic Romanian Documents*. In Petru Rebeja, Mihaela Onofrei, Dan Cristea and Dan Tufiş (eds.) Proceedings of the 16th International Conference “Linguistic Resources and Tools for Natural Language Processing”, online, 13–14 December, “Alexandru Ioan Cuza” University Publishing House, ISSN 1843-911X, 2021, 2021, pp. 59–74.
- (Cristea et al., 2022) CRISTEA, D., REBEJA, P. and PĂDURARIU, C. *Applying YOLOv5 Learning in Detecting Old Cyrillic Romanian Characters*. In Svetlana Cojocaru, Victoria Bobicev, Tatiana Verlan, Dan Tufiş and Dan Cristea (eds.) Proceedings of the 17th International Conference “Linguistic Resources and Tools for Natural Language Processing”, online, Chişinău, 10–12 November, “Alexandru Ioan Cuza” University Publishing House, ISSN 1843-911X, 2022, pp. 115–122.
- (Deans, 1983, 1993) DEANS, S.R. *The Radon Transform and Some of Its Applications*, John Wiley & Sons Ltd., New York, © 1983, 1993.
- (DeLORo, 2022) ∅, *DeLORo – Deep Learning for Old Romanian*. Raport științific final (2020–2022). (in Romanian) Online at [http://deloro.iit.academiaromana-is.ro/rapoarte/Raportare\\_stiintifica\\_finala.pdf](http://deloro.iit.academiaromana-is.ro/rapoarte/Raportare_stiintifica_finala.pdf).
- (Dierckx, 1995) DIERCKX, P., 1995. *Curve and Surface Fitting with Splines*. Clarendon Press.
- (DRH, 1975-2006) ∅, DRH A (1975–2006), B (1966–2010), C (1977–1985). *Documenta Romaniae Historica*. A. Moldova. Vol. I–XXXVI (1247–1651). I: (1384–1448). Academia de Științe Sociale și Politice a Republicii Socialiste România. Institutul de Istorie și Arheologie „A. D. Xenopol” – Iași. Volum întocmit de C. Cihodaru, I. Caproșu și L. Șimanschi. București, Editura Academiei Republicii Socialiste România, 1975, LV + 607 p.; II: (1449–1486). Academia de Științe Sociale și Politice a Republicii Socialiste România. Institutul de Istorie și Arheologie „A. D. Xenopol” – Iași. Volum întocmit de Leon Șimanschi în colaborare cu Georgeta Ignat și Dumitru Agache. București, Editura Academiei Republicii Socialiste România, 1976, LVIII + 649; III: (1487–1504). Academia de Științe Sociale și Politice a Republicii Socialiste România. Institutul de Istorie și Arheologie „A. D. Xenopol” – Iași. Volum întocmit de C. Cihodaru, I. Caproșu și N. Ciocan. București, Editura Academiei Republicii Socialiste România, 1980, LVIII + 686 p.; VI: (1546–1570). Academia Română. Institutul de Istorie „A. D. Xenopol” – Iași. Volum întocmit de I. [= Ioan] Caproșu. București, Editura Academiei Române, 2008, LXXV + 1 043 p.; VII: (1571–1584). Academia Română. Institutul de Istorie „A. D. Xenopol” – Iași. Volum întocmit de Ioan Caproșu. București, Editura Academiei Române, 2012, LXXIX + 1 039 p.; VIII: (1585–1592). Academia Română. Institutul de Istorie „A. D. Xenopol” – Iași. Volum întocmit de Ioan Caproșu. București, Editura Academiei Române, 2014, LXXXIII + 1 057 p.; IX: (1593–1598). Academia Română. Institutul de Istorie „A. D. Xenopol” – Iași. Volum întocmit de Petronel Zahariuc, Marius Chelcu, Silviu Văcaru, Cătălina Chelcu, Sorin Grigoruță. București, Editura Academiei Române, 2014, LV + 664 p.; XVIII: (1623–1625). Academia Română. Institutul de Istorie „A. D. Xenopol” – Iași. Volum întocmit de I. [= Ioan] Caproșu și V. Constantinov. București, Editura Academiei Române, 2006, LXXXIII + 680 p.; XIX: (1626–1628). Academia Republicii Socialiste România. Institutul de Istorie „N. Iorga”. Volum întocmit de Haralambie Chirca. 1969, LXII + 800 p.; XXI: (1632–1633). Academia de Științe Sociale și Politice a




Republicii Socialiste România. Institutul de Istorie și Arheologie „A. D. Xenopol” – Iași. Volum întocmit de C. Cihodaru, I. Caproșu și L. Șimanschi. București, Editura Academiei Republicii Socialiste România, 1971, LXV + 736 p.; XXII: (1634). Academia de Științe Sociale și Politice a Republicii Socialiste România. Institutul de Istorie și Arheologie „A. D. Xenopol” – Iași. Volum întocmit de C. Cihodaru, I. [= Ioan] Caproșu și L. [= Leon] Șimanschi. București, Editura Academiei Republicii Socialiste România, 1974, XLVII + 488 p.; XXIII: (1635–1636). Academia Română. Institutul de Istorie „A. D. Xenopol” – Iași. Volum întocmit de Leon Șimanschi, Nistor Ciocan, Georgeta Ignat și Dumitru Agache. București, Editura Academiei Române, 1996, XCIX + 908 p.; XXIV: (1637–1638). Academia Română. Institutul de Istorie „A. D. Xenopol” – Iași. Volum întocmit de C. Cihodaru, I. [= Ioan] Caproșu. București, Editura Academiei Române, 1998, LXXXII + 741 p.; XXV: (1639–1640). Academia Română. Institutul de Istorie „A. D. Xenopol” – Iași. Volum întocmit de Nistor Ciocan, Dumitru Agache, Georgeta Ignat și Marius Chelcu. București, Editura Academiei Române, 2003, LXXXIV + 708 p.; XXVII: (1643–1644). Academia Română. Institutul de Istorie „A. D. Xenopol” – Iași. Volum întocmit de Petronel Zahariuc, Cătălina Chelcu, Marius Chelcu, Silviu Văcaru, Nistor Ciocan, Dumitru Ciurea. București, Editura Academiei Române, 2005, LXVI + 727 p.; XXVIII: (1645–1646). Academia Română. Institutul de Istorie „A. D. Xenopol” – Iași. Volum întocmit de Petronel Zahariuc, Marius Chelcu, Silviu Văcaru, Cătălina Chelcu. București, Editura Academiei Române, 2006, LXVI + 728 p.; B. Țara Românească: Vol. I–XXVIII (1384–1646). I: (1247–1500). Academia Republicii Socialiste România. Institutul de Istorie „Nicolae Iorga”. Volum întocmit de P. P. Panaitescu și Damaschin Mioc. București, Editura Academiei Republicii Socialiste România, 1966, LXIV + 637 p.; II: (1501–1525). Academia de Științe Sociale și Politice a Republicii Socialiste România. Institutul de Istorie „Nicolae Iorga”. Volum îngrijit de Ștefan Ștefănescu și Olimpia Diaconescu. București, Editura Academiei Republicii Socialiste România, 1972, LXII + 602 p.; III: (1526–1535). Academia de Științe Sociale și Politice a Republicii Socialiste România. Institutul de Istorie „Nicolae Iorga”. Volum întocmit în cadrul seminarului de paleografie slavă, condus de Damaschin Mioc. București, Editura Academiei Republicii Socialiste România, 1975, XXXII + 452 p.; IV: (1536–1550). Academia de Științe Sociale și Politice a Republicii Socialiste România. Institutul de Istorie „Nicolae Iorga”. Volum întocmit în cadrul seminarului de paleografie slavă, condus de Damaschin Mioc. București, Editura Academiei Republicii Socialiste România, 1981, XXXIII + 430 p.; V: (1551–1565). Academia de Științe Sociale și Politice a Republicii Socialiste România. Institutul de Istorie „Nicolae Iorga”. Volum întocmit de Damaschin Mioc și Marieta Adam Chiper. București, Editura Academiei Republicii Socialiste România, 1983, XXXII + 455 p.; VI: (1566–1570). Academia de Științe Sociale și Politice a Republicii Socialiste România. Institutul de Istorie „Nicolae Iorga”. Volum îngrijit de Ștefan Ștefănescu și Olimpia Diaconescu. București, Editura Academiei Republicii Socialiste România, 1985, XXVIII + 371 p. + XX fotocopii; VII: (1571–1575). Academia de Științe Sociale și Politice a Republicii Socialiste România. Institutul de Istorie „Nicolae Iorga”. Volum îngrijit de Ștefan Ștefănescu și Olimpia Diaconescu. București, Editura Academiei Republicii Socialiste România, 1988, XXVIII + 440 p.; VIII: (1576–1580). Institutul de Istorie „Nicolae Iorga”. Volum întocmit de Damaschin Mioc și Ioana Constantinescu. București, Editura Academiei Republicii Socialiste România, 1996, XXXIX + 632 p.; XI: (1593–1600). *Domnia lui Mihai Viteazul*. Academia de Științe Sociale și Politice a Republicii Socialiste România. Institutul de Istorie „Nicolae Iorga”. Volum întocmit de Damaschin Mioc, Ștefan Ștefănescu, Marieta Adam, Constantin Bălan, Maria Bălan, Sașa Caracaș, Ruxandra Cămărășescu, Olimpia Diaconescu, Coralia Fotino. București, Editura Academiei Republicii Socialiste România, 1975, LI + 747 p.; XXI: (1626–1627). *Domnia lui Mihai Viteazul*. Academia Republicii Socialiste România. Institutul de Istorie al Academiei Republicii Socialiste România. Volum întocmit de Damaschin Mioc. București, Editura Academiei Republicii Socialiste România, 1965, XLII + 596 p.; XXII: (1628–1629). Academia Republicii Socialiste România. Institutul de Istorie „Nicolae Iorga”. Volum întocmit de Damaschin Mioc. București, Editura Academiei


- Republicii Socialiste România, 1969, XLV + 864 p.; XXIII: (1630–1632). Academia Republicii Socialiste România. Institutul de Istorie „Nicolae Iorga”. Volum întocmit de Damaschin Mioc. București, Editura Academiei Republicii Socialiste România, 1969, LI + 831 p.; XXIV: (1633–1634). Academia de Științe Sociale și Politice a Republicii Socialiste România. Institutul de Istorie „Nicolae Iorga”. Volum întocmit de Damaschin Mioc, Sașa Caracaș și Constantin Bălan. București, Editura Academiei Republicii Socialiste România, 1974, LVII + 715 p.; XXV: (1635–1636). Academia de Științe Sociale și Politice a Republicii Socialiste România. Institutul de Istorie „Nicolae Iorga”. Volum întocmit de Damaschin Mioc, Maria Bălan, Ruxandra Cămărășescu, Coralia Fotino. București, Editura Academiei Republicii Socialiste România, 1985, XXXVII + 598 p.; XXX: (1645). Academia Română. Institutul de Istorie „Nicolae Iorga”. Volum întocmit de Violeta Barbu, Marieta Chipier, Gheorghe Lazăr. 1998, LXVI + 520 p.; XXXI: (1646). Academia Română. Institutul de Istorie „Nicolae Iorga”. Volum întocmit de Violeta Barbu, Constanța Ghițulescu, Andreea Iancu, Gheorghe Lazăr, Oana Rizescu. București, Editura Academiei Române, 2003, LXXX + 519 p.; XXXIII: (1648). Academia Română. Institutul de Istorie „Nicolae Iorga”. Volum întocmit de Gheorghe Lazăr, Constanța Vintilă-Ghițulescu, Andreea Iancu. București, Editura Academiei Române, 2006, LXVII + 518 p.; XXXIV: (1649). Academia Română. Institutul de Istorie „Nicolae Iorga”. Volum întocmit de Violeta Barbu, Gheorghe Lazăr, Oana Rizescu. București, Editura Academiei Române, 2002, LXX + 373 p.; XXXV: (1650). Academia Română. Institutul de Istorie „Nicolae Iorga”. Volum întocmit de Violeta Barbu, Constanța Ghițulescu, Andreea Iancu, Gheorghe Lazăr, Oana Rizescu. București, Editura Academiei Române, 2003, LXXV + 487 p.; XXXVI: (1651). Academia Română. Institutul de Istorie „Nicolae Iorga”. Volum întocmit de Oana Rizescu și Marcel-Dumitru Ciucă. București, Editura Academiei Române, 2006, LXI + 405 p.; XXXVII: (1652). Academia Română. Institutul de Istorie „Nicolae Iorga”. Volum întocmit de Violeta Barbu, Constantin Bălan, Florina Manuela Constantin. București, Editura Academiei Române, 2006, LXXIX, 459 p.; XXXIX: (1654). Academia Română. Institutul de Istorie „Nicolae Iorga”. Volum întocmit de Violeta Barbu, Gheorghe Lazăr, Florina Manuela Constantin, Constanța Ghițulescu, Oana Mădălina Popescu. București, Editura Academiei Române, 2010, CX + 815 p.; C. Transilvania: vol. X–XII (1351–1365). X: (1351–1355). Academia de Științe Sociale și Politice a Republicii Socialiste România. Institutul de Istorie și Arheologie – Cluj-Napoca. Sub redacția acad. Ștefan Pascu. București, Editura Academiei Republicii Socialiste România, 1977, XLIII + 463 p.; XI: (1356–1360). Academia de Științe Sociale și Politice a Republicii Socialiste România. Institutul de Istorie și Arheologie – Cluj-Napoca. Sub redacția acad. Ștefan Pascu. București, Editura Academiei Republicii Socialiste România, 1981, LXIV + 660 p.; XII: (1361–1365). Academia de Științe Sociale și Politice a Republicii Socialiste România. Institutul de Istorie și Arheologie – Cluj-Napoca. Sub redacția acad. Ștefan Pascu. București, Editura Academiei Republicii Socialiste România, 1985, LIX + 540 p.
- (Edwards, 2007) EDWARDS, J.A. *Easily adaptable handwriting recognition in historical manuscripts*. PhD Thesis, University of California Berkeley, 2007.
- (Gafton, 2013–2016), *The Corpus electronic al textelor românești vechi (1521 – 1640)* (CETRV) / *Electronic Corpus of The Old Romanian Texts (1521–1640)* (CETRV) Project Site. [http://media.lit.uaic.ro/?page\\_id=3914](http://media.lit.uaic.ro/?page_id=3914)
- (Găman et al., 2022) GĂMAN, M., GHADAMIYAN, L., IONESCU, R. T., and POPESCU, M. (2022). *Self-paced learning to improve text row detection in historical documents with missing labels*. In Proceedings of TiE (ECCV Workshop).
- (Gheție, 1984) GHEȚIE, I. (1984), *Începuturile scrisului în limba română*, București, Editura Academiei R.S.R.
- (Gheție coord., 1997) GHEȚIE, I. (coord.) (1997). *Istoria limbii române literare. Epoca veche. (1532–1780)*. Întocmită de Gheorghe Chivu, Mariana Costinescu, Constantin Frîncu, Ion Gheție,

- Alexandra Roman Moraru și Mirela Teodorescu. Coordonator: Ion Gheție. București, Editura Academiei Române, 1997, 496 p.
- (Gidaris *et al.*, 2018) GIDARIS, S., SINGH, P., and KOMODAKIS, N. (2018). *Unsupervised Representation Learning by Predicting Image Rotations*, 6th International Conference on Learning Representations, {ICLR} 2018, Vancouver, BC, Canada, April 30 – May 3, 2018.
- (Girshick *et al.*, 2016) GIRSHICK, R.B., DONAHUE, J., DARRELL, T., and MALIK, J., *Region-Based Convolutional Networks for Accurate Object Detection and Segmentation*. IEEE Transactions on Pattern Analysis and Machine Intelligence 38 (2016): pp 142–158.
- (He *et al.*, 2015) HE, K., ZHANG, X., REN, S., and SUN, J. (2015). *Deep Residual Learning for Image Recognition*. arXiv. <https://doi.org/https://arxiv.org/abs/1512.03385v1>
- (Hurmuzachi, 1887–1913) HURMUZACHI, E. (1887–1913). *Documente privitoare la istoria românilor*. I: Partea I: (1199–1345). Culese de Eudoxiu de [= Docsachi (Eudoxiu)] Hurmuzaki. București, Editura Socec, 1887, XXX + 704 p.; Partea II: (1346–1450). București, Editura Socec, 1890, XLVIII + 889 p; II: Partea I: (1451–1575). Culese de Eudoxiu de [= Docsachi (Eudoxiu)] Hurmuzaki. București, Editura Socec, 1891, XLIV + 824 p. [Documente dintre 1500 și 1575]; Partea II: (1451–1510). Culese de Eudoxiu de [= Docsachi (Eudoxiu)] Hurmuzaki. București, Editura Socec, 1891, XLVIII + 889 p. [Documente dintre 1500 și 1575]; Partea III: (1510–1530). Culese, adnotate și publicate de Nic. [= Nicolae] Densusianu. Cu un apendice *Documente slavone însoțite de traduceri latine (1510–1527)*. București, Editura Socec, 1892, XL + 748 p; Partea IV: (1531–1552). Culese, adnotate și publicate de Nic. [= Nicolae] Densusianu. 1894, XXXIII + 756 p; Partea V: (1552–1575). Culese de Eudoxiu de [= Docsachi (Eudoxiu)] Hurmuzaki. București, Editura Socec, 1897, XXX + 700 p; III. Partea I: (1576–1599). Culese de Eudoxiu de [= Docsachi (Eudoxiu)] Hurmuzaki. București, Editura Socec, 1880, XXX + 600 p.; Partea II: (1576–1600). Cu portretul lui Petru-Vodă Șchiopul. Publicate sub auspiciile Ministerului Cultelor și Instrucțiunii publice și ale Academiei Române. Culese de Eudoxiu de [= Docsachi (Eudoxiu)] Hurmuzaki. București, Editura Socec, 1888, XXXII + 575 p.; IV. Partea I: (1600–1649). Culese de Eudoxiu de [= Docsachi (Eudoxiu)] Hurmuzaki. București, Editura Socec, Publicate sub auspiciile Ministerului Cultelor și Instrucțiunii publice și ale Academiei Române. 1882, XXXVI + 708 p.; Partea II: (1600–1650). Cu portretul lui Vasile-Vodă-Lupul. Publicate sub auspiciile Academiei Române și ale Ministerului Cultelor și al Instrucțiunii Publice. Culese de Eudoxiu de [= Docsachi (Eudoxiu)] Hurmuzaki. București, Editura Socec, 1884, XL + 686 p; V. Partea I: (1650–1699). Cu portretul lui Gheorghe Ștefan Voevod. București, Editura Socec, 1885, XXXII + 547 p.; Partea II: (1650–1699). Cu portretul lui Constantin-Vodă Brâncoveanu. București, Editura Socec, 1886, XXVI + 406 p.; VI: (1700–1750). Publicate sub auspiciile Ministerului Cultelor și Instrucțiunii publice și ale Academiei Române. București, Ed[itura] Socec, Sander și Teclu, 1878, XXIV + 697 p.; VII: (1750–1818). Publicate sub auspiciile Ministerului Cultelor și al Instrucțiunii Publice. București, Ed[itura] Socec, Sander și Teclu, 1876, XXII + 585 p.; VIII: (1376–1650). București, Institutul de Arte grafice Carol Göbl., 1894, XLVIII + 540 p. [Documente dintre 1500 și 1650]; IX. Partea I: (1650–1747). 1900, LI + 691 p.; XI: Acte din secolul al XVI-lea (1517–1612) relative mai ales la domnia și viața lui Petru Vodă Șchiopul. Adunate, adnotate și publicate de Neculai [= Nicolae] Iorga. Sub auspiciile Ministerului Cultelor și Instrucțiunii Publice și ale Academiei Române. București, Editura Socec, 1900, CLIV + 909 + XLIV p. [Documente dintre 1588 și 1594]; XII: Acte relative la războaiele și cuceririle lui Mihai Vodă Viteazul: (1594–1602). Adunate, adnotate și publicate de N. [= Nicolae] Iorga. Sub auspiciile Ministerului Cultelor și Instrucțiunii Publice și ale Academiei Române. București, Editura Socec, 1903, LXXXIX + 1 281 + XXXIV p. [Documente dintre 1594 și 1602]; XV: Acte și scrisori din arhivele orașelor ardeleni (Bistrița, Brașov, Sibiu): Partea I: (1358–1600). București, Editura Socec, LXXXVIII + 775 p. [Documente dintre 1594 și 1600]; Partea II: (1601–1825). București, Editura Socec, 1913, CIII + 1168 p. [Documente dintre 1601 și 1750].
- (Iorga, 1901–1914) IORGA, N. (1901–1914). *Studii și documente cu privire la istoria românilor*. Vol. I–XXXI. Publicate de N. [= Nicolae] Iorga. I: *Socotelile Bistriței (Ardeal)*. București, Editura

Ministeriului de Instrucție. Stabilimentul Grafic I. V. Socec, XLIX + 53 p. [Documente dintre 1524 și 1692]; II: *Acte relative la istoria cultului catolic în Principate*. Adunate și tipărite cu o prefață despre propaganda catolică până la 1500 de ... București, Editura Ministeriului de Instrucție. Stabilimentul Grafic I. V. Socec, p. 54–535. [Documente dintre 1600 și 1847]; III: *Fragmente de cronici și știri despre cronicari*. Adunate și tipărite cu o prefață despre istoria munteană în legătură cu istoriografia sârbească de ... București, Editura Ministeriului de Instrucție. Stabilimentul Grafic I. V. Socec, 1901, 2 f. + LXXXI + 104 p. [Documente dintre 1667 și 1821]; IV: *Legăturile Principatelor Române cu Ardealul de la 1601 la 1699. Povestire și izvoare de ...* București, Editura Ministeriului de Instrucție. Stabilimentul Grafic I. V. Socec, 1902, 2 f. + CCCXIX + 345 p. [Documente dintre 1600 și 1767]; V: *Cărți domnești, zapise și răvașe ...* Partea I. București, Editura Ministeriului de Instrucție. Stabilimentul Grafic I. V. Socec, 1903, 1 f. + VII + 720 p. [Documente dintre 1600 și 1850]; VI: *Cărți domnești, zapise și răvașe ...* Partea II. București, Editura Ministeriului de Instrucție. Stabilimentul Grafic I. V. Socec, 1904, 1 f. + XI + 661 p. [Documente dintre 1602 și 1853]; VII: *Cărți domnești, zapise și răvașe ...* Partea III: *Istoria literaturii religioase a românilor până la 1688*[1]. București, Editura Ministeriului de Instrucție. Stabilimentul Grafic I. V. Socec, 1904, 8 p. + CCXLIII + 383 p. [Documente dintre 1614 și 1851]; VIII: *Scrisori de boieri și negustori olteni și munteni către Casa de negoț sibiiană Hagi Pop*, București, Atelierile Grafice Socec & Comp., Societate anonimă, 1906, 2 f. + LXXI + 203 p. [Documente dintre 1712 și 1836]; IX: *Povestiri, scrisori și cronici*. București, Editura Ministeriului de Instrucție. Stabilimentul Grafic I. V. Socec, 1905, 2 f. + 225 p. [Documente dintre 1638 și 1707]; X: *Brașovul și românii*. Scrisori și lămuriri. București, Stabilimentul Grafic I. V. Socec, 1905, 2 f. + 455 p. [Documente dintre 1644 și 1825]; XI: *Cercetări și regeste documentare*. București, Editura Ministeriului de Instrucție (Atelierile Grafice Socec & Comp., Societate anonimă), 1906, 2 f. + 307 p. [Documente dintre 1602 și 1846]; XII: *Scrisori și inscripții ardelenne și maramureșene*. București, Atelierile Grafice Socec & Comp., Societate anonimă, 1906, 1 f. + LXVII + 303 p.; [Documente dintre 1628 și 1847]; XIII: *Scrisori și inscripții ardelenne și maramureșene*. II. *Inscripții și însemnări*. București, Atelierile Grafice Socec & Comp., Societate anonimă, 1906, 3 f. + 336 p. [Documente dintre 1650 și 1848]; XIV: *Hârtii din arhiva Mănăstirii Hurezului precum și din a Protopopiei Argeșului, din a boierilor brâncoveni și altor neamuri găsite în casele proprietății din Brâncoveni*. Publicate cu o introducere, note și indice. București, Atelierile Grafice Socec & Comp., Societate anonimă, 1907, 2 f. + XLV + 386 p. [Documente dintre 1625 și 1853]; XV: (I<sub>1</sub>) *Inscripții din bisericile României*. Adunate, adnotate și publicate de ... Fascicula I. N-rele 1–764. București, Institutul de Arte Grafice și Editură „Minerva”, 1905, VIII + 312 p. [Documente dintre 1602 și 1857]; (I<sub>2</sub>) *Inscripții din bisericile României*. Adunate, adnotate și publicate de ... Fascicula II. N-rele 766–944. București, Atelierile Grafice Socec & Comp., Societate anonimă, 1907, p. 317-374; [Documente dintre 1650 și 1873]; (II) *Inscripții din bisericile României*. Adunate, adnotate și publicate de ... Volumul II cuprinzând 1079 nre. București, Editura Ministeriului de Instrucție Publică Tip[ografia] „Neamul Românesc”, 1908, 1 planșă + 2 f. + 381 p. [Documente dintre 1614 și 1870]; XVI: *Chemarea lui Constantin-Vodă Mavrocordat către olteni (1737). Trei scrisori privitoare la lenăchiță Văcărescu. Un manuscris de leacuri. Două acte privitoare la Unire (1857-1858). O moșie a lui Mihai Viteazul: Bârca din Dolj. Un memoriu de avocat al lui Mihail Kogălniceanu (1847). Mărunțișuri istorice. Documente amestecate. Acte botoșănene și dorohoiene privitoare mai mult la familia Cananău. Alte mărunțișuri istorice. Documente comunicate de d. Simionescu-Râmniceanu. Din actele expuse la expoziția istorică din Iași. Documente ale familiei Palade. Un catastif de dajde ale orașului Galați (c. 1683). Pomelnicul românesc al mănăstirii Bisericiani. Inscripții și însemnări. Documente amestecate. Varia. Miscellanea, (1600–1874)*. București, Atelierile Grafice Socec & Comp., Societate anonimă, 1909, IX + 452 p.; XVII [pe copertă XVIII]: *Constatări istorice cu privire la viața agrară a românilor ... (1600–1789)*. București, Atelierile Grafice Socec & Comp., Societate anonimă, 1908, 1 f. + 91 p.; XVIII: *Scrisori și alte acte privitoare la Unirea Principatelor... (1856–1865)*. București, Atelierile Grafice Socec & Comp., Societate anonimă, 1910, 1 f. + 104 p.; XIX: *Documente felurite. Câteva inscripții și însemnări de biserici. Condica de menzilar*

- a lui Scarlat-Vodă Callimachi (1635–1849)*. București, Atelierile Grafice Socec & Comp., Societate anonimă, 1910, 2 f. + 130 p.; XXI: *Documente interne. Miscellanea (1624–1846)*. București, Editura Ministeriului de Instrucție Publică Tip[ografia] „Neamul Românesc”, 1911, 2 f. + 613 p. + erată; XXII: *Documente interne (1633–1846)*. București, Editura Ministeriului de Instrucție Publică Tip[ografia] „Neamul Românesc”, 1913, 474 p.; XXIV: *Basarabia noastră. Scrisă la 100 de ani de la răpirea ei de către ruși*. Vălenii-de-Munte, Editura și Tipografia Societății „Neamul Românesc”, 1912, 2 f. + 178 p.; XXV: *Corespondența lui Dimitrie Aman negustor din Craiova (1804–1844)*. București, Tip[ografia] „Neamul Românesc”, 1913, XV + 253 p. + erată; XXIX: *Viața și domnia lui Constantin-Vodă Brâncoveanu ...* București, Tip[ografia] „Neamul Românesc”, 1914, 215 p.
- (Ivănescu, 2000/1980) IVĂNESCU, G. (2000/1980). *Istoria limbii române*, Ediția a II-a, Îngrijirea ediției indice de autori și indice de cuvinte: Mihaela Paraschiv, Iași, Editura Junimea. The first edition was published by the same publishing house in 1980.
- (Jaccard, 1912) JACCARD, P. *The Distribution of the Flora in the Alpine Zone. I*. New Phytologist 11, no. 2 (1912): 37–50. <https://doi.org/10.1111/j.1469-8137.1912.tb05611.x>.
- (Kingma and Ba, 2015) KINGMA, D. P., and BA, J. (2015) *Adam: A Method for Stochastic Optimization*. 3rd International Conference for Learning Representations, San Diego, 2015, <https://doi.org/10.48550/arXiv.1412.6980>.
- (Kraken OCR, 2021) ☞, *Kraken OCR, Unleashing the Kraken for OCR*. <https://medium.com/analytics-vidhya/unleashing-the-kraken-for-ocr-fba6bff73c8c>. Accessed 5 Mar 2021.
- (Kumar et al., 2010) KUMAR, M.P., PACKER, B., and KOLLER, D. (2010). *Self-Paced Learning for Latent Variable Models*. In: Proceedings of NIPS, vol. 23, pp. 1189–1197.
- (Law and Deng, 2020) LAW, H., and DENG, J. (2020). *CornerNet: Detecting Objects as Paired Keypoints*, Int J Comput Vis 128, 642–656.
- (Lee and Schachter, 1980) LEE, D.T., and SCHACHTER, B.J. (1980). *Two algorithms for constructing a Delaunay triangulation*, International Journal of Computer & Information Sciences 9: 219–242.
- (Li et al., 2019) LI, X., ZHANG, B., LIAO, J., and SANDER, P.V. (2019). *Document rectification and illumination correction using a patch-based CNN*, ACM Transactions on Graphics 38: 168:1–168:11.
- (Likforman-Sulem et al., 1995) LIKFORMAN-SULEM, L., HANIMYAN, A., and FAURE, C. (1995). *A Hough based algorithm for extracting text lines in handwritten documents*, Proceedings of 3rd International Conference on Document Analysis and Recognition, vol.2., 774–777.
- (Louloudis et al., 2009) LOULUDIS, G., GATOS, B., PRATIKAKIS, I., and HALATSIS, C. (2009). *Text line and word segmentation of handwritten documents*, Pattern Recognition 42: 3169–3183.
- (Macé et al., 2019) MACÉ, C., ROUQUETTE, M., SERETAN, V., AMSLER, F., ANDRIST, P., and ANTONELLI, C. (2019). *Critical digital editions of Christian apocryphal literature in Latin and Greek: Transcription and Collation of the Acts of Barnabas*. *Storie e Linguaggi*, 5(1):125–145.
- (Monk, 2004) ☞, *Monk wiki*. <https://www.ai.rug.nl/~lambert/Monk-collections-english.html>. Accessed 20 Nov 2020.
- (OpenAI, 2023) ☞, *OpenAI. “GPT-4 Technical Report.” ArXiv abs/2303.08774*, 2023.
- (Panaiteescu, 1965) PANAITESCU, P. P. (1965), *Începuturile și biruința scrisului în limba română*, București, Editura Academiei R.P.R.
- (Pratikakis, 2021) PRATIKAKIS, I. *Accessing Greek historical handwritten documents using the μDoc.tS platform (DeLORo -> DeLOGr)*. Abstract of the invited communication in the DeLORo Workshop section of: Petru Rebeja, Mihaela Onofrei, Dan Cristea and Dan Tufiș (eds.) Proceedings of the 16th International Conference “Linguistic Resources and Tools for Natural Language Processing”, online, 13–14 December, “Alexandru Ioan Cuza” of Iași Publishing House, ISSN 1843-911X, 2021, pp 3.
- (Redmon et al., 2015) REDMON, J., DIVVALA, S.K., GIRSHICK, R.B., and FARHADI, A. *You Only Look Once: Unified, Real-Time Object Detection*. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015): 779–788.

- (Rosseti *et al.*, 1971/1961) ROSETTI, Al., CAZACU, B., and ONU, L. (1971/1961). *Istoria limbii române literare*. Vol. I: *De la origini până la începutul secolului al XIX-lea*. Ediția a doua, revăzută și adăugită. București, Editura Minerva, 1971, 672 p. Ediția I: 1961.
- (Sanasam *et al.*, 2020) SANASAM, I., CHOUDHARY, P., and SINGH, K.M. (2020). *Line and word segmentation of handwritten text document by mid-point detection and gap trailing*, Multimedia Tools and Applications 79: 30135-30150.
- (Schomaker, 2020) SCHOMAKER L (2020) *Lifelong learning for text retrieval and recognition in historical handwritten document collections*. In: Fischer A, Liwicki M, Ingold R (eds) *Handwritten historical document analysis, recognition and retrieval – state of the art and future trends*. World Scientific, London, pp 221–248.
- (Seretan, 2020) SERETAN, V. *UNINE, SIB, DARIAH dans Sharing the Experience: Workflows for the Digital Humanities*. Proceedings of the DARIAH-CH Workshop 2019, Neuchâtel, 2020.
- (Shi *et al.*, 2015) SHI, B., BAI, X., and YAO, C. (2015). *An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition*. arXiv. <https://doi.org/https://arxiv.org/abs/1507.05717v1>
- (Simonyan and Zisserman, 2014) SIMONYAN, K., & ZISSERMAN, A. (2014). *Very Deep Convolutional Networks for Large-Scale Image Recognition*. arXiv. <https://doi.org/https://arxiv.org/abs/1409.1556v6>
- (Strempele, 1978–1992) ȘTREMPELE, G. (1978–1992). *Catalogul manuscriselor românești* [Vol. I–IV]. I: *B.A.R. 1–1600*. București, Editura Științifică și Enciclopedică, 1978, 431 p.; II: *B.A.R. 1 601–3 100*. București, Editura Științifică și Enciclopedică, 1983, 504 p.; III: *B.A.R. 3101–4413*. București, Editura Științifică și Enciclopedică, 1987, 495 p.; IV: *B.A.R. 4414–5920*. București, Editura Științifică, 1992, 543 p.
- (Tesseract, 2021).  *Tesseract Open Source OCR Engine*, <https://github.com/tesseract-ocr/tesseract> (main repository). Accessed 6 Dec 2021.
- (Thoppilan *et al.*, 2022) Thoppilan, R., De Freitas, D., Jamie Hall, Noam M. Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, Yaguang Li, Hongrae Lee, Huaixiu Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Yanqi Zhou, Chung-Ching Chang, I. A. Krivokon, Willard James Rusch, Marc Pickett, Kathleen S. Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Hartz Sriraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravindran Rajakumar, Alena Butryna, Matthew Lamm, V. O. Kuzmina, Joseph Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguerre-Arcas, Claire Cui, Marian Croak, Ed Huihsin Chi and Quoc Le. “LaMDA: Language Models for Dialog Applications.” *ArXiv abs/2201.08239* (2022)
- (Turner *et al.*, 2015) THURNER, S., HANEL, R., LIU, B. and COROMINAS-MURTRA, B. (2015). *Understanding Zipf’s law of word frequencies through sample-space collapse in sentence formation*. *Journal of the Royal Society Interface* 12(108), published 06 July 2015.
- (Tocilescu *et al.*, 1886–1900) TOCILESCU, Gr.G., ODOBESCU, A.I., STURDZA, D. A., and COLESCU-VARTIC, C. (1886–1900). *Documente privitoare la istoria românilor*. Urmare la colecțiunea lui Eudoxiu de [= Docsachi (Eudoxiu)] Hurmuzaki. Vol. I–VI. Suplement I–II. București, Ed[itura] Socec. Suplement I. I: (1518–1780). Cu portretul lui Ioan Nicolae Alexandru Mavrocordat Voevod. Documente culese din diferite publicațiuni și din Biblioteca Națională din Paris de Gr. G. Tocilescu. Documente culese din Arhivele Ministeriului Afacerilor Străine din Paris de A. I. Odobescu. Publicate sub auspiciile Ministeriului Cultelor și Instrucțiunii Publice și ale Academiei Române. 1886, LXXI + 1 003 p.; II: (1781–1814). Documente culese din Arhivele Ministeriului Afacerilor Străine din Paris de A. I. Odobescu. 1885, XLVIII + 757 p.; III: (1709–1812), 1889, XIII + 596 p.; IV: (1802–1849), 1891, XXX + 596 p.; V: (1822–1838). Documente adunate și coordonate de D. A. Sturdza și C. Colescu-Vartic. 1894, XXIV + 664 p.;

- VI: (1827-1849). Documente adunate și coordonate de D. A. Sturdza și C. Colescu-Vartic. 1895, XXXI + 633 p.; Suplement II, vol. I-III. Publicate sub auspiciile Ministeriului Cultelor și Instrucțiunii Publice și ale Academiei Române. I: (1510-1600). 1893, XXII + 652 p.; II: (1601-1640). 1895, XXXII + 624 p.; III: (1641-1703). 1900, 312 p.
- (Touvron *et al.*, 2023) TOUVRON, H., LAVRIL, T., IZACARD, G., MARTINET, H., LACHAUX, M.-A., LACROIX, T., ROZIĆRE, B., GOYAL, N., HAMBRO, E., AZHAR, F., RODRIGUEZ, A., JOULIN, A., GRAVE E., and LAMPLE, G. *LLaMA: Open and Efficient Foundation Language Model*, *ArXiv* abs/2302.13971 (2023)
- (Tsochatzidis *et al.*, 2021) TSOCHATZIDIS L, SYMEONIDIS S, PAPAZOGLU A, and PRATIKAKIS I. (2021). *HTR for Greek Historical Handwritten Documents*. *Journal of Imaging* 7(12):260. <https://doi.org/10.3390/jimaging7120260>
- (UNESCO, 2003) , UNESCO, *Charter on the Preservation of Digital Heritage*, Paris, 15 October, 2003, online at: <https://en.unesco.org/about-us/legal-affairs/charter-preservation-digital-heritage>
- (Yan *et al.*, 2018) YAN, C., HU, J., and ZHANG, C. (2018). *Deep transformer: A framework for 2D text image rectification from planar transformations*, *Neurocomputing* 289: 32-43.