MOLDOVA STATE UNIVERSITY
DOCTORAL SCHOOL OF PHYSICAL, MATHEMATICAL,
INFORMATION, AND ENGINEERING SCIENCES

Presented as manuscript
U.D.C.: 004:[94(478):008]

BUMBU TUDOR

TECHNOLOGIES AND INFORMATION RESOURCES FOR THE DIGITIZATION AND PROCESSING OF TEXTS FROM THE CULTURAL HERITAGE

Summary of the Ph.D. thesis in Computer Science

121.03 − COMPUTER PROGRAMMING

Author: Bumbu ____________ Bumbu Tudor

PhD Supervisor: S. Cojocaru ____________ Cojocaru Svetlana, Ph.D. hab. in computer science, professor, Corresponding Member of the Academy of Sciences of Moldova

Guidance commission: ____________ Gaindric Constantin, Ph.D. hab. in computer science, professor, Corresponding Member of the Academy of Sciences of Moldova

____________ Titchiev Inga, Ph.D. in Computer science, university professor

____________ Burtseva Lyudmila, Ph.D. in Computer science, associate professor

CHIȘINĂU, 2023

The thesis was elaborated at the Doctoral School of Physical, Mathematical, Information, and Engineering Sciences, Moldova State University.

Ph.D. Commission:

The Chairman of the Commission: LOZOVANU Dmitrii, Ph.D. hab. in physical and mathematical sciences, university professor, Corresponding Member of the Academy of Sciences of Moldova, Vladimir Andrunachievici Institute of Mathematics and Computer Science, USM.

PhD Supervisor: COJOCARU Svetlana, Ph.D. hab. in computer science, professor, Corresponding Member of the Academy of Sciences of Moldova, Academy of Sciences of Moldova.

Official reviewers:

GAINDRIC Constantin, Ph.D. hab. in computer science, professor, Corresponding Member of the Academy of Sciences of Moldova, Vladimir Andrunachievici Institute of Mathematics and Computer Science, USM.

IFTENE Adrian, Ph.D., university professor, Alexandru Ioan Cuza University, Iaşi, România.

PETIC Mircea, Ph.D. in computer science, associate professor, Alecu Russo State University, Bălți, Moldova.

Scientific secretary: NOVAC Ludmila, Ph.D. in physical and mathematical sciences, associate professor, Moldova State University.

The thesis defense will take place on September 19, 2023, at 14-00, bureau 340, Vladimir Andrunachievici Institute of Mathematics and Computer Science, USM, Academiei str. 5, Chisinau, Moldova.

The doctoral thesis and the summary can be consulted at the Library of Moldova State University and on the website of the National Agency for Quality Assurance in Education and Research (www.cnaa.md).

Summary sent on ____________________

Secretary of Doctoral Commission: ____________ Novac Ludmila

Author: Bumbu ____________ Bumbu Tudor

Contents

KEYWORDS

Keywords: Digitization, Romanian Cyrillic documents, historical-cultural heritage, OCR technology, transliteration, text alignment, classification of old fonts, neural networks, platform for digitization.

1. RESEARCH GOALS AND OBJECTIVES

Actuality and importance of the research topic. Digitization occupies a leading position in 21st-century technologies. As early as 2011, the European Commission presented a recommendation document on the digitization and online accessibility of cultural material and digital preservation [1], in which was stated that the development of the digitization process of material located in libraries, archives, and museums should be further encouraged to ensure that Europe maintains its position as a leading player internationally in the field of culture and creative content and that it utilizes its wealth of cultural material in the best possible way, urging member states to intensify their investments in this field.

The recommendation has been included as a policy action in several countries (not just those in the EU), with an entire industry developing that offers scanning, recognition, and other related services. The issue of digitization and preservation of cultural heritage represents a priority area in the digital agenda for Europe [2].

Large-scale digitization, which initially was limited to scanning and storing images, began with the Gutenberg project [3], initiated in the 1970s, and later, the Million Books Collection[1] and the Google Books[2] digitization projects. Although these projects solve the problem of preserving the printed heritage, the scanning of printed materials can only be considered a starting point in terms of preserving the knowledge they contain and facilitating access to them.

Even though many documents can be found and read online, they cannot be processed automatically, as in most cases they are only presented in image format, not in machine-readable text format. Therefore, the challenge of automating the process of transforming documents into computer-readable, hence editable text, falls to machine learning and computer vision applications, namely those of Optical Character Recognition (OCR). We will demonstrate in this work that this task cannot always be considered trivial, as the range of variation of the source material (the quality and volume of the document, the period of its editing, the storage conditions, etc.) is extremely wide. However, making digitized cultural heritage documents available in an editable format has been and continues to be considered a necessity. This is particularly emphasized in the report [4], which warns that Europe is in danger of entering a new dark age if sufficient means are not created to preserve and facilitate access to cultural heritage material. As a result, several large-scale projects have been funded that deal with OCR of historical prints in the context of mass digitization, the most important being

[1] Digitization project "Million Book Collection," http://ulib.isri.cmu.edu/ (Accessed 23.03.2023).

[2] Digitization project "Google Books," https://books.google.com/ (Accessed 23.03.2023).

the IMPACT project [2, 5], which aims to improve access to text, and the OCR project for early modern prints - eMOP[3].

Addressing this issue for the Romanian heritage faces specific challenges and aspects: many periods in language evolution, a relatively small and very scattered number of deposited resources, and a great diversity of alphabets used for document printing. The obstacles encountered in the digitization and preservation of this treasure are related to the correct recognition of Cyrillic letters, but also to the non-existence of a lexicon suitable for the period of printing of the old resources [6]. In particular, the problem of creating linguistic resources, digitizing, and processing cultural heritage texts from various historical periods is actual in several European countries [7-10].

By the Government Decision of the Republic of Moldova no. 857 of October 31, 2013, the National strategy for the development of the information society "Digital Moldova 2020", as well as the action plan for its implementation, was approved. Even though its provisions have not been fully implemented, this regulation has stimulated the digitization activities of document collections from the country's libraries and archives. However, solving the issue of digitizing the printed cultural heritage of the Republic of Moldova remains relevant, which would provide computer tools capable of processing documents from different historical periods, with different alphabets, with diverse vocabularies, preserved in various conditions – tools, which could benefit both researchers and the general public, offering them the possibility to operate with large indexed data collections.

The research goals and objectives. The goal of this research is to establish and develop computer tools for processing the Romanian language heritage printed in the 17th-20th centuries. The proposed goal determined the need to formulate the following objectives:

- analysis and determination of the main methods of preprocessing old documents;
- creation of a collection of scanned resources for training OCR models and elaborating dictionaries for OCR technology;
- development of an OCR technology for Romanian documents printed in the 17th-20th centuries;
- development of algorithms for transliteration from Cyrillic to Latin spelling;
- research and development of methods for aligning old texts with a contemporary vocabulary, developing support for alignment;
- development of a platform for the digitization and processing of Romanian Cyrillic documents.

[3] https://emop.tamu.edu/

The achievement of the proposed goals and objectives has contributed to obtaining important applicative results, incorporated in the digitization platform, the use of which facilitates access to the Romanian cultural heritage printed in Cyrillic script.

2. SCIENTIFIC RESEARCH METHODOLOGY

During the research carried out within the thesis, methods from the field of natural language processing and machine learning were used. The research process was comprehensive, with a rigorous approach to each stage: defining the problem, the documentation, formulating working hypotheses, analyzing and testing the results, and disseminating them.

The documentation phase and the formulation of working hypotheses are based on the goals and objectives of the research. One of the main objectives was the development of OCR technology for Romanian documents printed in the 17th-20th centuries. Achieving this objective required an in-depth investigation of existing OCR techniques and adapting them to the specifics of old Romanian texts. These texts include a variety of fonts (especially those from the 17th century), the printing format and the quality of the scanned document, and the linguistic specifics, with an orthography and syntax different from the contemporary ones.

During the testing phase, experiments were carried out using various training sets and word dictionaries in learning OCR models with the help of two versions of ABBYY FineReader software to identify the most suitable approach for OCR. In addition to OCR testing, the rules from the transliteration algorithm from Cyrillic to Latin spelling based on the generated Cyrillic lexicons were also tested. The analysis of the results consisted of evaluating the performance of the trained OCR models and comparing the results with test sets. The performance of OCR models was tested also using word dictionaries.

The research results have been disseminated through a series of publications in specialized journals and presentations at national and international conferences. This allowed exchanging ideas with other researchers and paved the way for further improvements to the methods and techniques used in the research.

The scientific novelty and originality of the thesis consist in researching and developing the technology for solving the problem of recognition and transliteration of Romanian Cyrillic documents printed in the 17th-20th centuries, which allows efficient and rapid processing of the mentioned documents. The degree of novelty and originality is represented by:

- developing OCR technology for Romanian documents printed in the 17th-20th centuries;

- developing algorithms for transliterating from the Romanian Cyrillic alphabet to the modern Romanian (Latin) alphabet;
- elaborating a method of classifying the fonts used in printing old texts;
- creating a method of aligning old texts to the modern vocabulary using sequence similarity techniques;
- developing a web digitization platform for processing Cyrillic documents.

The significant scientific problem solved in the field of research is the development of optical character recognition technology and transliteration from Cyrillic to Latin script of Romanian Cyrillic documents printed in the 17th-20th centuries, given the existence of a wide variety of alphabets and fonts.

The theoretical significance is determined by the development of a technology that enables the conversion of Romanian documents from the Cyrillic to the Latin alphabet, with the application of methods based on neural networks.

The applicative value of the paper consists in the development of a platform for digitization, which significantly contributes to the automation of reprinting old documents. It is a useful tool for libraries and archives in creating digital content, for researchers in the field of history, philology, etc., and for a wide range of users, offering them assistance at the stages of preprocessing, recognition, and postprocessing of documents.

Approval of scientific results. The scientific results obtained by the author in this thesis were presented at national and international scientific conferences and were published in peer-reviewed journals. The main results included in the thesis were presented at the following scientific conferences:

- *Development of a platform for heterogeneous document recognition using convergent technology*. Workshop on Intelligent Information Systems WIIS 2022, October 06-08, 2022, Chisinau, Republic of Moldova;
- *Platform for Digitization of Heterogeneous Documents*. The 29th Conference on Applied and Industrial Mathematics CAIM 2022, August 25-27, 2022, Chisinau, Republic of Moldova;
- *Punctilog Compared to Dependency Grammar and Constituency Grammar*. Symposium on Logic and Artificial Intelligence SLAI2022, January 12-16, 2022, Louisiana, USA;
- *User Interface to Access Old Romanian Documents.* The 4th Conference of Mathematical Society of Moldova CMSM4'2017, June 25-July 2, 2017;

- *Generation of the Romanian Cyrillic lexicon for the period 1967 – 1989.* The Fifth Conference of Mathematical Society of the Republic of Moldova, September 28-October 1, 2019, Chisinau, Republic of Moldova;
- *On Classification of 17th Century Fonts using Neural Networks*. Workshop on Intelligent Information Systems (WIIS2021), October 14-15, 2021, Chisinau, Republic of Moldova;
- *Discover the Moldovan Cultural Heritage through e-Moldova Portal by Using Crowdsourcing Concept*. Workshop on Intelligent Information Systems WIIS2021, October 14-15, 2021, Chisinau, Republic of Moldova;
- *Evaluarea Corpusului Diacronic Paralel cu Texte Românești din Noul Testament din 1648 & 1990*. The 9th edition of the Scientific Conference of Ph.D. Students "Contemporary Trends in Science Development: Visions of Young Researchers", Vol. 1, June 10, 2020, Chisinau;
- *Building a Diachronic Parallel Corpus for the Alignment of the Old Romanian Texts.* Conference on Mathematical Foundations of Informatics MFOI-2019, July 3-6, 2019, Iasi, Romania.

Publications on the topic of thesis research. The results obtained in the thesis are published in 17 scientific papers (see [50-66]): 5 articles in scientific journals (see [51, 52, 56, 59, 62]), 2 of which are single-author articles (see [51, 56]); 12 papers at international conferences (see [50, 53-55, 57, 58, 60, 61, 63, 64, 65, 66]).

Thesis structure and volume. The thesis is written in Romanian, under the title of the manuscript. The paper has the following structure: the introduction, three chapters, general conclusions and recommendations, annotations in Romanian, Russian, and English, bibliography comprising 140 titles. The total volume of the thesis is 136 pages, of which 115 pages are the main text.

3. SYNTHESIS OF CHAPTERS

In the **Introduction**, the relevance and importance of the research theme are highlighted, presenting concise and up-to-date information about the recent state of digitization of the historical-cultural heritage. The goal and objectives of the thesis are defined, and the scientific novelty of the obtained results, as well as the theoretical and practical value of the thesis, are presented. This is accompanied by the demonstration and validation of the results.

The first chapter, "Tools and methods for processing historical documents," contains an analysis of scientific studies on the methods, tools, and resources for digitizing documents from historical-cultural heritage. The concepts of digital heritage [11], digitization of old documents, image preprocessing from old printed documents, optical character recognition (OCR), ground truth, OCR accuracy evaluation metrics, and postprocessing are defined.

Subsequently, the methods and tools for optical character recognition (OCR, hereinafter referred to as recognition) in historical documents are described. Many such documents have been scanned and stored in databases and portals, and their recognition is essential to make them accessible. It is noted that the recognition of modern printed documents is very efficient, with an accuracy of over 99%, due to the similarity between the learned and recognized characters, clear separation of characters from the background, and modern spelling of words. However, historical documents pose a serious challenge for OCR.

There are specified two main methods of training OCR models: training on synthetic data (images generated from electronic text and fonts available on computer) and training on real data (pairs of glyph shapes or character images and their transcription - the Unicode character) [12]. While training on synthetic data is more efficient, the recognition quality is lower for historical documents compared to training on real data [13]. Two major problems are identified in applying the technology of recognizing printed historical documents: the need to train a specific model for each book and the difficulty of transferring the model's accuracy from one book to another. To solve these problems, recognition algorithms based on recurrent neural networks are being experimented with [14].

Individual models provide excellent accuracy on the books they were trained on, but cannot be successfully generalized to other documents. One solution is the training of mixed models, which use for training a variety of documents printed at different printing houses in order to achieve better generalization. Training mixed models can overcome the typographic barrier, so the OCR results can be used to train more accurate OCR models. Many works focused on the recognition of historical documents are based on Tesseract [15, 16]. It is shown that the FineReader OCR engine can provide better accuracy (at the character level) than Tesseract [17]. For these reasons, it was decided to use the FineReader engine for the recognition of Romanian Cyrillic documents.

The use of Ocropy [18], an OCR method for processing historical documents, which has considerable accuracy, is illustrated. The Calamari software [19] is presented, an OCR toolkit that outperforms Ocropy by using a CNN[4]-LSTM[5] neural network architecture based on TensorFlow[6].

[4] A convolutional neural network (CNN or ConvNet) is a class of artificial neural networks, most used for image analysis and recognition.

[5] Long short-term memory (LSTM) is a type of artificial recurrent neural network used in the field of artificial intelligence and deep learning. Such a neural network (recurrent) can process not only individual data points (like images), but also entire sequences of data (like audio or video).

[6] https://www.tensorflow.org/learn

Calamari improves computational performance, especially on a GPU, and offers additional features such as early stopping of the training process, cross-validation, and pre-training. Compared to Ocropy, Calamari proves to be faster and more efficient [20]. In tests, Calamari was trained on a corpus of historical newspapers from Finland, providing a character accuracy between 87% and 92%. The datasets on which the neural networks in Ocropy and Calamari are trained consist of text lines, not individual glyphs, as are the data sets for FineReader and Tesseract. The performance is described through cross-validation with the character error rate (CER) and word error rate (WER) to measure the efficiency of the method. The results of the experiments indicated that the mixed models achieved an average error of 2.6% CER and 10% WER. Applying the voting mechanism leads to improved results, and post-recognition error correction further improves accuracy. The new versions of FineReader and Tesseract also use deep learning, with multilayer networks of the CNN and LSTM types.

Next, a series of digitization and processing platforms or frameworks for historical documents are analyzed. An important example is the *Historical Document Processing and Analysis Framework* (HDPA[7]), described in [21], a complex web framework for the management and analysis of historical documents, with an emphasis on OCR (Optical Character Recognition). HDPA is free for research and efficient in preparing data sets for OCR. HDPA has eight modules, which facilitate image preprocessing and segmentation, the creation of the dataset for training the OCR model, and the recognition itself. The framework is developed in Django, allowing for the development of individual Python modules. HDPA does not include a post-processing module, but it supports a simple-to-implement integration of new modules, allowing system customization for the specific needs of the user. The OCR module in HDPA is based on machine learning technologies, using CNN networks for feature extraction and a bidirectional LSTM recurrent neural network for sequential recognition of text lines. One of these platforms is also Aletheia [22], focusing on analyzing the appearance of the document page and page segmentation, identifying and classifying areas of interest in a scanned image of a text document. The process includes detecting and labeling various blocks, such as text blocks, illustrations, mathematical symbols, and tables. Aletheia can automatically detect objects on four levels: areas of interest, text lines, words, and glyphs. Creating ground truth data sets, stored in PAGE XML format, is another feature of Aletheia [23].

Transkribus[8], another platform, was developed at the University of Innsbruck and includes tools for recognizing, transcribing, and searching historical documents. However, it does not offer support for generating synthetic data, a function available in HDPA. Also mentioned here is the OCR-

[7] The HDPA framework is available at http://ocr-corpus.kiv.zcu.cz/ (accessed on June 20, 2022).
[8] The "Transkribus" project, https://readcoop.eu/transkribus (accessed on May 26, 2023).

D project developed in Germany, which includes 8 specialized modules for different OCR stages. As part of this project, the OCR4all platform was created, an open-source tool for semi-automated processing of historical documents [24]. There are also other more specialized tools, such as those for generating artificial OCR datasets for Russian, Arabic, and Romanian languages [25, 26, 52]. However, they are limited to certain tasks and do not consider the entire digitization process. Therefore, the value of digitization platforms, which offer a broader spectrum of functionalities, is greater, which convinced us to work on such a platform (described in Chapter 3).

The OCR post-processing is also examined - a critical stage in verifying and improving the text recognized by an OCR engine, adding value to the system by increasing its robustness and utility for the digitization of historical documents. Approaches vary, with some viewing post-processing as a spell-checking task, while others, like the sequence-to-sequence method, might use a dictionary or lexicon to detect and correct OCR errors [27-30]. Most post-processing methods include at least two steps: the generation of candidates for error replacement and the decision to accept corrections. Other approaches add additional steps, such as expanding the word dictionary and ranking the candidates based on analytic rules, as described in [31], where the post-processing of documents in German is discussed. These approaches yield good results when the Levenshtein distance between the candidate token and the correct result was no greater than 2. Manual OCR post-processing can give high-quality results, but it requires time, effort, and specialized knowledge, especially for historical documents with old alphabets. Semi-automatic approaches, like PoCoTo [32], a tool for semi-automatic correction of OCR text, make this task easier and more efficient. An enhanced and fully automated version, A-PoCoTo [33], was developed in 2019. Other approaches include grouping OCR errors in vector space, as presented in [34, 35], where a Word2Vec model is used to obtain groups of errors and synonyms of words.

Next, the DeLORo project [36, 37] is examined, dedicated to the processing of historical texts printed in Romanian Cyrillic characters and their transliteration into Latin characters. As part of this project, an online tool for annotating images from Romanian Cyrillic documents, called OOCIAT, was developed, and a significant corpus, named ROCC, was created, which includes 367 pages of scanned historical documents annotated with transcribed text. The ROCC corpus contains documents from the XVI-XIX centuries, organized according to difficulty, type of writing, and level of annotation. It is recognized that large datasets are needed to train neural networks. For this, both annotations made through the OOCIAT interface are used, including the UAIC-RoDia Treebank corpus [38]. Training experiments of the OCR model used a combination of a statistical model for feature extraction and a neural network with a CNN architecture, to discover objects and assign them labels. However, despite the progress made, the results for the manuscript collections were less satisfactory, which is an issue that the authors have not yet addressed. Methods for word separation [39] are also proposed, using a sequence-to-sequence approach; and they also, there is a plan to apply string kernels and spectral clustering to group old word forms belonging to the same lemma and the same part of speech. Dedicated researchers are actively working on integrating the developed tools into DeLORo, with the goal of providing unrestricted access for all.

The last section of this chapter describes the need for tools for digitizing and processing Romanian documents printed with Cyrillic alphabets, considering their large number and diversity. In the evolution of the Romanian language, we distinguish two eras: old and modern, the first lasting until 1650 [40]. On Romanian territory, the first book was printed in 1508, and the first in Romanian - in 1535 [41]. Significant collections of Romanian documents with Cyrillic alphabets are found in libraries both in the Republic of Moldova and Romania, as well as in libraries in other countries, such as those in St. Petersburg (Russia). For example, the Library of the Romanian Academy has over 1960 prints from 1508-1830, 79% of which are in Cyrillic script [42]. All these documents require digitization and processing to valorize them in the Romanian heritage.

In Chapter 2, "Technologies for processing Romanian documents from the 17th-20th centuries", our approaches in designing the technology for processing historical texts (printed in Romanian with Cyrillic characters, starting from the 17th century) are substantiated, describing the developed methods, and justifying the use of certain modules from the existing ones. These include image preprocessing modules; OCR models; neural network models for font classification; technology for transliteration from the Romanian Cyrillic alphabet to the modern one; support for aligning old texts to modern ones.

Initially, the basic actions carried out in the recognition of 17th-century texts are described. The process is organized on the principle of using convergent technologies, that is, the interconnection within a platform of applications from certain fields, alongside the development of its own components. The use of the ABBYY FineReader Professional (hereinafter FR) program for the optical recognition of Romanian Cyrillic characters is analyzed. The following actions involve testing and adapting version FR 12, as well as newer versions, such as FineReader 14 and FineReader 15. Since these versions are not initially oriented towards the processing of old Romanian texts, it is necessary to extend their capabilities in order to adapt them to the solution of the problems mentioned. Operating with documents from a certain historical period has imposed the creation of new components (such as alphabets and dictionaries) and training the software on additional data sets, in order to ensure the highest possible quality of results. These actions lead to the elaboration of one or more models oriented towards the recognition of texts from a certain historical period. The process of optical character recognition for a new language in FR consists of the following steps: image preprocessing, which involves editing, cleaning, and adjusting the image resolution to optimize results; creation of the language and alphabet, which will contain all characters of the new language; preparation and creation of a word dictionary for the new language; training of models, where supervised learning of each character takes place to allow their proper segmentation from the image.

The section dedicated to processing images from old documents describes the specifics of applying two tools for image processing - the one incorporated in FR and ScanTailor[9], elucidating the specifics of their application in the processing of old texts. Here it is concluded that FR provides some options necessary for preprocessing old texts, but it does not offer the entire useful spectrum, requiring the involvement of additional tools. One of the essential modules for preprocessing old documents that FR lacks refers to the thickening of characters. This problem arises from the fact that some image binarization methods can thin the lines in glyphs, and to thicken them again in the image preprocessing stage, a special module from ScanTailor is used, a tool further described in this section of the thesis. Image conversion to black and white (binarization) is more efficient in ScanTailor than in FR. The implementation of this functionality in ScanTailor is based on illumination normalization [43], Savitzky-Golay smoothing[10], actual binarization based on Otsu's Method, and, finally, the removal of broken edges. It is mentioned that it is advisable to save documents in “black and white” format, as this demonstrates better OCR accuracy. However, the application of this option requires special attention, as decolorization can lead to the loss of some text elements. This can be somewhat compensated for by thickening the characters. Another peculiarity is configuring the resolution so that the OCR engine can correctly detect the text lines in the image, especially if diacritics, accents, or other elements persist above the text line.

Later, in the section about creating the user language and adding the word dictionary, it is mentioned that some characters, such as Ꙟ and Ꙋ from the Romanian Cyrillic alphabet do not exist in the alphabet addition system in FR and cannot be displayed by the fonts in its system. Therefore, it was necessary to identify and adapt them from *BabelMap*[11]. Next, three methods for expanding the word dictionary in the OCR process from FR are described. The first method involves transliterating existing vocabularies from the modern (Latin) alphabet to the Cyrillic one and adding them to the dictionary in FR; the second method involves creating the dictionary from the text of an already recognized document; the third method is based on including words from recognized portions of the document from the FR graphical interface.

[9] https://scantailor.org/

[10] https://en.wikipedia.org/wiki/Savitzky%E2%80%93Golay_filter

[11] https://www.babelstone.co.uk/Unicode/babelmap.html

Next, the particularities of OCR models applied to texts printed in the 17th century are described. The case study of the optical recognition of 17th-century books, described in the thesis, was based on the book “New Testament” printed in 1648, from which data sets were created taken from the first 257 pages. The data sets consist of the character cut from the page and the corresponding UNICODE character. A common aspect of books printed in the 17th century is the writing of certain letters above others. Also, abbreviations are used that use tildes or other diacritical signs. Considering this aspect, it is proposed to significantly increase the resolution (over 1200 DPI) so as to include all elements of a character in the training process.

Next, the process and results of OCR evaluation for 17th-century Romanian Cyrillic documents are discussed. For this purpose, a data set is created from 15 pages of books from this period. A single page from the set of pages contains on average 1400 characters and 260 words. The evaluation criteria considered were OCR accuracy both at the character level and at the word level. The overall accuracy is calculated using the method described in [44]. In the experiments demonstrated, OCR accuracy is calculated both with and without the word dictionary. Four experiments are presented with an increase in the dataset size for each experiment, where the last experiment shows the OCR model trained with 7 pages from the training set and the number of glyphs in the set is over 3600. In this experiment, the best accuracy at the character level was found, which is 96%, using a word dictionary. It is concluded that increasing the training set and word dictionaries leads to an increase in overall accuracy.

Next, the issue of classifying 17th-century fonts is addressed. The typographies of the 17th century have different fonts, among which two completely different fonts stand out, both in terms of writing/printing style and character usage [50, 51]. This problem cannot be solved in FR by training mixed models, therefore individual models are trained for each font. Some solutions for font classification are proposed next. One solution is a program for selecting the appropriate OCR model depending on the typography [50], and another solution consists in classifying fonts using neural networks [53], a solution that is detailed in the thesis. The dataset is created from 10 scanned books, selected from the *Digital Library of Romania*[12]. When creating the dataset, clustering methods such as PCA and K-Means are used. The obtained dataset consists of over 21,200 training examples and over 9 thousand testing examples. Next, a multilayer neural network (MNN) based on Keras and TensorFlow is trained to classify characters into two different fonts. The construction of the neural network begins with a transformation of the input data from an x-by-y matrix into a vector of length x*y. A hidden layer with 128 neurons with a *ReLU*[13] activation function is added, which is fully

[12] http://digitool.bibnat.ro/ (Carte românească veche şi bibliofilă/Sec. XVII)

[13] https://keras.io/api/layers/activations/#relu-function

connected to the last layer. As it is a binary classification, the output layer contains a single neuron and a *sigmoid*[14] activation function. After training, an accuracy of 96.7% is observed, an accuracy that can be improved by a more complex architecture, such as CNN-LSTM, where the order of characters is also considered.

Next, the process of transliteration is described, defining it as a conversion of a text from one alphabet to another, which involves changing letters in predictable ways [45]. Regarding the Romanian language, the changes that have taken place in the writing systems throughout history are highlighted [46]. Several methods of transliteration are then introduced, including techniques used in transliterating English proper names into Chinese, Japanese, Korean, or Arabic [47]. “Direct Orthographic Mapping” techniques that use n-gram-based models for transliteration are also mentioned [48]. Standardizing transliteration procedures is essential to ensure a rigorous, univocal, and completely reversible conversion [49]. The process of transliteration in Moldova began in 1989, with the adoption of the Law on the functioning of languages spoken on the territory of the Moldovan SSR.

Subsequently, some specific difficulties in transliterating texts printed with the Romanian Cyrillic alphabet are described, with particular emphasis on the problem of correctly representing Cyrillic text on a computer, especially those from 17th-century documents. Most letters (37 out of 43) are transliterated using simple, context-independent rules, and the rest are transliterated using context-dependent rules. The most problematic letter is “ѧ” which can be transliterated as “a”, “e”, “ea”, “ia”. Although some contextual dependency rules have been established at the character level, there are still exceptional cases where its transliteration falls outside the usual patterns. For example, “чѣѧ” and “кърꙋѧ” will both be transliterated using the same rule (“ѧ” => “ia”), although the word “чѣѧ” could become “ceia” and “ceea”. Ideally, we would also need dependency rules at the level of neighboring words. However, the transliteration accuracy exceeds 98%. In addition to the rules, exception dictionaries are also used to transliterate words that cannot be correctly represented based solely on rules. At the end of the compartment, two applications for transliterating from Cyrillic to Latin for the 17th-20th centuries are discussed. A desktop application is developed in Java, and another one with a web interface, but with limited functionality.

At the end of this chapter, a series of papers regarding the alignment of old texts to modern ones are discussed. A diachronic parallel corpus [54] and alignment tools are introduced. Aligning an old text to its modern representation involves translating it into a contemporary language, replacing outdated lexical variants with modern expressions. Parallel texts, like these, are valuable resources

[14] https://keras.io/api/layers/activations/#sigmoid-function

for machine translation and diachronic analysis of natural languages. The first proposed step in this direction is the creation of a diachronic parallel corpus of about 8400 sentences, based on the New Testament printed in 1648 in Bălgrad aligned to its modern electronic version from 1990.

Subsequently, word alignment tools [55, 56] are discussed, which are software programs that assist in aligning words from a source text with those from a target text. The tools examined are the *Berkeley Word Aligner*[15], a Java program that uses hidden Markov models (HMM) for word alignment in a sentence-level parallel corpus, and *GIZA* ++[16], a tool that uses HMM models for text alignment.

Considering that our object of study is diachronic parallel texts, it was decided to create a special alignment tool, which includes specific features such as calculating the BLEU score between texts, sentences, expressions, and words, interactive visualization of n-grams, and others. The developed tool is a web application based on Django, with three main modules: the parallel text editing and parallel corpus formation module, the text processing module, and the machine learning module.

There are plans to expand this tool with a text annotation component using the Punctilog methodology [57, 58] and features that use the BLEU score to evaluate and improve the similarity between diachronic parallel texts.

Chapter 3, "Platform for Digitizing Romanian Cyrillic Documents," is the final chapter of the thesis – dedicated to the design and description of a platform that includes the digitization toolkit for documents in the Romanian language printed in Cyrillic script [71, 73, 80-82]. The platform is presented as the main practical result of the thesis; it allows access to tools and resources for digitizing these documents through an interactive interface based on web technologies. The architecture of the platform is described, which includes four functional groups (G1-G4), namely: image processing, optical document recognition, text transliteration, and saving and publishing digitized documents.

Next, each functional group is described. A functional group consists of integrated modules that contain software developed by the author and third parties. The first functional group described refers to image preprocessing, namely the steps taken to prepare an image so that the OCR engine can analyze it. The OCR engine can sometimes have difficulty correctly interpreting images that are blurry, distorted, or have low contrast. Preprocessing helps improve OCR accuracy by preparing the image to be more suitable for recognition. Some important modules included in this functional group

[15] https://github.com/mhajiloo/berkeleyaligner
[16] https://github.com/moses-smt/giza-pp

are: adjusting the contrast or brightness of the image to improve text reading; binarization, which involves converting the image into a black and white version to improve contrast; noise removal by removing additional black pixels from the image that can lead to the recognition of unnecessary characters, such as some punctuation marks; distortion correction which involves rotating the image to correctly align it. By preprocessing the image before sending it to the OCR engine, the accuracy and reliability of the OCR process can generally be improved. This functional group integrates preprocessing modules from software such as ScanTailor, FineReader 15, and the Python package - OpenCV. The purpose of the G1 functional group is to prepare the document for OCR.

Next, the G2 functional group is described, which is about the optical recognition of documents. This group includes modules for selecting the OCR model depending on the historical period; using word dictionaries for recognition; editing the recognized text, using OCR exceptions dictionaries. The OCR engine is based on FineReader 15. G2 is initialized with 8 OCR models trained with data sets collected from documents printed in the 17th, 18th, 19th, and 20th centuries. Users can also add new OCR models. These models have the FineReader XML format (.fbt files) containing the OCR model configurations, the training data set, the necessary alphabet, as well as word dictionaries.

The actions in the G2 group start with the selection of the document period. So, usually, the user knows the historical period of printing of the document that is to be subjected to digitization, moreover, he can even indicate the year when it was printed. However, it is not excluded that the user has several images from a random document about which he knows nothing more than the fact that this document is in Cyrillic script. For such cases, an automatic period detection module would be useful. An approach that can be useful in solving this problem is the experience of detecting fonts in 17th-century Cyrillic printed documents, where certain neural network models have been trained to automatically recognize the document's font. A module that deals with the detection of 17th-century fonts is included in G2.

The recognition process, which can be divided into several parts that can be run in parallel to increase processing speed, is discussed next. This process is managed through *ABBYY Hot Folder*[17] (hereafter HF). This can be accomplished by using multiple instances for each OCR model, thus splitting the preprocessed images into several folders, which can improve efficiency when multiple users are working on the platform simultaneously. Recognizing a single page of text takes on average 30 seconds, although sometimes recognizing such a page could take up to two minutes. This is because the instances created in Hot Folder are verified every minute (this is the minimum time option

[17] https://help.abbyy.com/en-us/finereader/15/user_guide/hotfolder/

in HF) if newly processed images have appeared in the folders with processed images. A PDF document with 50 pages of text will be recognized in about 90 seconds; a PDF with 100 pages of text - in 150 seconds; PDF with 360 pages of text - over 385 seconds (more than 6 minutes). Text documents in PDF format attest to a duration of approximately 1.2 seconds per page. No stable duration was observed in PDFs with images. The OCR accuracy criterion at the character and word level is analyzed in Chapter 2 of the thesis. For example, the OCR model for the 20th century gives us a character-level accuracy of over 98%; 18th-century models offer over 92% at the word level; and the model for the 17th century offers an accuracy of over 95% at the character and word dictionary level, considering the appropriate image preprocessing, the document's scan quality, its wear, etc. Next, the use of word dictionaries used within the OCR engine and some OCR exception dictionaries consisting of tuples formed from an expression containing "recognition ambiguities" and the correct version of this expression is demonstrated. We used the phrase "recognition ambiguities" for the simple reason that some letters have a very close graphic similarity, and sometimes the OCR engine recognizes the wrong version with a very high probability. In such a case, the internal dictionary cannot propose the correct candidate even if the correct version would have been in the dictionary. For example, the letter **и** is confused with the letter **н** in the expression "сърачїн", so the OCR exception dictionary could contain the exception: (сърачїн, сърачїи). To handle such situations, an OCR postprocessing component is included in G2 using the exception dictionary. Exception dictionaries are also used in transliteration, and we have a similar module in G3. The G2 functional group also includes a text editing module. This text editor has a web virtual keyboard that adjusts its character composition according to the document period.

The next section presents the transliteration modules included in the G3 functional group. Along with the transliteration operation itself, the G3 offers spelling updates, management of exception dictionaries for transliteration, and automatic text correction. Transliteration is possible in two ways. The first way is to use the *AAConv*[18] web transliteration API, and the second possibility is to use the same application, but in a desktop version. A notable difference between these two versions is that the web version can accept only up to 1.2MB of text in a single process. An important module for the user is the spelling update, which, on request, considers the rules of writing the modern Romanian language. An example is the spelling with ***â*** (from *a*), included as an option in the transliteration process. According to the recommendations of the Romanian Academy, the letter "î" will always be written at the beginning or the end of the word ("început", "înger", "în", "întoarce", "a coborî", "a urî"). Inside the word, "â" is usually written ("cuvânt", "a mârâi"). However, there are

[18] https://translitera.cc/

a few exceptions to this rule. A G3 module for managing exception dictionaries is discussed next. The transliteration exception dictionaries keep words that cannot be correctly transliterated using only the transliteration rules. For example, the word “амязэ” according to transliteration rules becomes “amează”, the correct version being “amiază", which is found in the respective dictionary. This module allows the management of the exception list. Exceptions are handled after the transliteration of the text from Cyrillic to Latin according to the rules, but before viewing and checking the text in the text editor. Many exceptions come from the different spelling of words of foreign origin, especially proper nouns.

Additionally, the G3 group shares the same text editing module as the G2 group, and the virtual keyboard and word dictionaries for the spell checker are adapted to the transliterated text. Here it is considered that the virtual keyboard contains the letters of the modern Romanian alphabet, and the word dictionary is written with the modern Romanian alphabet. An experimental module described in G3 is the correction of transliterated text with an artificial intelligence system: *GPT-3*[19] developed by *OpenAI*[20]. In this module, experiments are conducted with the *text-davinci-003* model for correcting the recognized text.

In the following, the functional group G4 is described, with modules for managing and publishing documents, which allows saving recognized/transliterated texts in different formats, downloading processed images, and publishing digitized documents in digital libraries. An important module in G4 is saving the digitized document in the platform's database. In addition to storing texts and links to files, the digitized object is also stored, which represents a JavaScript object with the help of which the status of each step made through the digitization application described in the following section is preserved. The object includes preprocessing parameters, recognition and transliteration parameters, recognized and edited text, and transliterated and edited text. A set of modules included in G4 refers to the publication of the digitized document. A publishing module is based on the *eMoldova*[21] portal, based on a portlet called *Tezaurul Național Digital*[22].

In the last section of Chapter 3, a digitization application within the platform is described as a demonstrative instance of some modules implemented in the platform. The purpose of developing this application is to demonstrate the functionality of some modules in the platform. The integrated digitization application on the platform allows the digitization of the processed document in 7 steps,

[19] https://en.wikipedia.org/wiki/GPT-3
[20] https://en.wikipedia.org/wiki/OpenAI
[21] https://emoldova.org/
[22] https://digi.emoldova.org/

and the time duration for a complete digitization cycle varies between 2 and 15 minutes, depending on the volume of the document.

4. GENERAL CONCLUSIONS

The support of the process of revitalizing cultural and historical heritage remains an issue, the current relevance and importance of which constitutes a priority mentioned in several policy documents of European countries. By achieving the objectives set out in the thesis, certain contributions have been made to facilitate the digitization and transliteration of texts printed in Romanian with Cyrillic characters, covering a temporal segment of the last four centuries. Through the analysis and development of methods and tools used in image preprocessing, the development of OCR models, etc., integrated into the digitization platform, access to old Romanian Cyrillic documents is facilitated, opening new opportunities for research and valorization of cultural and historical resources.

The study of the obtained results allows the formulation of the following general conclusions:

- The analysis of tools and methods for digitizing historical documents reveals a multitude of methods, procedures, resources, and platforms available for preprocessing, recognition, postprocessing, and transliteration of historical documents, which differ in accuracy and operational efficiency [6-40].
- Recognition methods based on OCR training on images with text lines increase the speed of training OCR engines, thus accelerating the digitization process, giving it a mass character [18-20].
- As a result of adapting the FR 15 software system components for recognizing old Romanian prints, it was found that the accuracy of the model increases significantly with the increase of the number of training pages. The evaluation of the learning procedure within an iterative process showed that with the increase of training data, the accuracy of the model significantly increases, reaching acceptable values (0.96 when operating with a dictionary and 0.95 when operating without a dictionary) at the level of correct character recognition even after a not very large number of pages (5-7 pages). At the word level, the accuracy value is lower, indicating the need to use a larger number of pages for training.
- For processing images from old documents, we can use existing preprocessing tools, supplementing those incorporated in FR15 with the possibilities offered by Scan Tailor, especially for character thickening, Savitzky-Golay smoothing, and removal of broken edges.
- The font classification algorithm, developed by creating and training a multilayer neural network [51], demonstrated an accuracy of over 96%.

- Transliteration from the Romanian Cyrillic alphabet into the modern one using context-dependent rules is performed with an accuracy that exceeds 98%.
- The alignment tool developed performs the alignment of old texts to modern ones by evaluating and creating the similarity of diachronic parallel texts, based on character string similarity. This tool facilitates the creation of a diachronic parallel corpus [54].
- The digitization platform, the architecture, modules, and applications of which were developed within the thesis, including image preprocessing tools, OCR models, applications for transliteration from Cyrillic to Latin script, text editing modules of recognized/transliterated texts, allows the realization of the main tasks related to the digitization of old Romanian documents in an efficient and quick manner [60, 61].
- The digitization platform can be used as a web or desktop application and can be extended to include digitization modules for other languages. This platform is useful for libraries, publishers, and researchers who hold collections of documents in Romanian printed with Cyrillic characters. Along with these, the existence of such a platform, especially in the web version, facilitates access to the cultural-historical legacy also for the general public.

Recommendations for future research and developments in the field of digitizing Romanian Cyrillic documents could include:

- Continuous improvement of OCR models and transliteration algorithms, by integrating new and advanced techniques in the field of natural language processing and machine learning, to increase the accuracy and efficiency of the recognition and transliteration process.
- Development of a simple UI for mass training of OCR models to open access to as many users as possible. In this way, we will be able to build OCR models for short time intervals, as well as for most typographies.
- Expansion of the digitization platform to include other types of documents, such as manuscripts, maps, or illustrations, to allow access to a wider variety of cultural and historical resources.
- Integration of the digitization platform with other digital tools and resources, such as digital libraries, archives, and databases, to facilitate collaboration between researchers and provide additional information and resources.

REFERENCES

[1] Recomandarea Comisiei din 27 octombrie 2011 privind digitizarea și accesibilitatea online a materialului cultural și conservarea digitală (2011/711/UE) – In: *Jurnalul Oficial al Uniunii Europene*, 29.10.2011, https://eur-lex.europa.eu/legal-content/RO/TXT/PDF/?uri=CELEX:32011H0711&from=EN (Accesat 24.03.2023).

[2] Centru de Competență în Digitizare „IMPACT”, http://www.digitisation.eu/community/ map-of-the-digitisation-landscape/ (Accesat 5.08.2022).

[3] Proiectul „Gutenberg”, http://www.gutenberg.org/ (Accesat 23.03.2023).

[4] NIGGEMANN, E., DE DECKER, J., LÉVY, M. The new renaissance. In: *Raportul 'comité des sages'.* Grup de reflecție pentru aducerea online a patrimoniului cultural al Europei. Bruxelles, Comisia Europeană, 2011, p. 45.

[5] BALK, H., CONTEH, A. IMPACT: centre of competence in text digitisation. In: *Proceedings of the 2011 Workshop on Historical Document Imaging and Processing.* ACM, 2011, pp. 155–160.

[6] BOIAN, E., CIUBOTARU, C., COJOCARU, S., COLESNICOV, A., MALAHOV, L. DIGITIZAREA, RECUNOAŞTEREA ŞI CONSERVAREA PATRIMONIULUI CULTURAL-ISTORIC. *Revista Akademos*, nr. 1 (32), martie 2014, pp. 61-68.

[7] MORUZ, M., IFTENE, A., MORUZ, A., CRISTEA, D. Semi-automatic alignment of old Romanian words using lexicons. In: *Proceedings of the 8th International Conference „Linguistic resources and tools for processing ofthe Romanian language”*, Iași, Editura Universității „A.I. Cuza”, 2012, p. 119-125.

[8] HAUG, D. T. T., JØHNDAL, M. L. Creating a Parallel Treebank of the Old Indo-European Bible Translations. In: *Caroline Sporleder and Kiril Ribarov (eds.). Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008),* 2008, pp. 27-34.

[9] VITAS, D., KRSTEV, C., OBRADOVIĆ, I., POPOVIĆ, L., PAVLOVIĆ-LAŽETIĆ, G. Processing serbian written texts: An overview of resources and basic tools. In: *International Workshop on Balkan Language Resources and Tools,* Thessaloniki, Greece, 2003, pp. 97-104.

[10] PAVLOV, R., BOGDANOVA, G., PANEVA-MARINOVA, D., TODOROV, T., RANGOCHEV, K. Digital archive and multimedia library for bulgarian traditional culture and folklore. In: *International Journal "Information Theories and Applications".* Vol. 18, Number 3, 2011, pp. 276-288.

[11] Concept of Digital Heritage. In: *UNESCO*. https://en.unesco.org/themes/information-preservation/digital-heritage/concept-digital-heritage (Accesat: 1.04.2023).

[12] SPRINGMANN, U., LÜDELING, A. OCR of historical printings with an application to building diachronic corpora: a case study using the RIDGES herbal corpus. *arXiv preprint* arXiv:1608.02153 (2016)

[13] UWE, S., NAJOCK, D., MORGENROTH, H., SCHMID, H., GOTSCHAREK, A., FINK, F. OCR of historical printings of Latin texts: problems, prospects, progress. In: *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, 2014, pp. 57–61. DATeCH '14. New York, NY, USA: ACM. doi:10.1145/2595188.2595197.

[14] BREUEL, T. M., ADNAN, UL-H., MAYCE, A. AL-A., FAISAL, S. High-Performance OCR for Printed English and Fraktur Using LSTM Networks. In: 2th International Conference on Document Analysis and Recognition (ICDAR), 2013, 683–87. IEEE.

[15] Tesseract OCR project, https://github.com/tesseract-ocr (Accesat 7.06.2022).

[16] DUDCZAK, A., NOWAK, A., PARKOŁA, T. Creation of Custom Recognition Profiles for Historical Documents. In: *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, ACM, 2014, pp. 143–46.

[17] HELIŃSKI, M., KMIECIAK, M., PARKOŁA, T. Report on the comparison of Tesseract and ABBYY FineReader OCR engines. *PCSS*, 2012, 24 p.

[18] SPRINGMANN, U., NAJOCK, D., MORGENROTH, H., SCHMID, H., GOTSCHAREK, A., FINK, F. OCR of historical printings of latin texts: problems, prospects, progress. In: *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, ACM, 2014, pp. 71–75.

[19] WICK, C., REUL, C., PUPPE, F. Calamari—a high-performance TensorFlow-based deep learning package for optical character recognition. *arXiv preprint* arXiv:1807.02004, 2018.

[20] WICK, C., REUL, C., PUPPE, F. Comparison of OCR accuracy on early printed books using the open source engines Calamari and OCRopus. *JLCL 33*, 2018, pp. 79–96.

[21] LENC, L., MARTÍNEK, J., KRÁL, P., NICOLAOU, A., CHRISTLEIN, V. HDPA: historical document processing and analysis framework. *Evolving Systems*, 2021, pp. 177-190.

[22] CLAUSNER, C., PLETSCHACHER, S., ANTONACOPOULOS, A. Aletheia— an advanced document layout and text ground-truthing system for production environments. In: *Proceedings of the 11th International Conference on Document Analysis and Recognition* (ICDAR2011), Beijing, China, 2011, pp. 48–52.

[23] CLAUSNER, C., ANTONACOPOULOS, A., PLETSCHACHER, S. ICDAR2019 Competition on Recognition of Documents with Complex Layouts. In: *Proceedings of 2019 International Conference on Document Analysis and Recognition (ICDAR)*, 2019, pp.1521-1526.

[24] REUL, C., CHRIST, D., HARTELT, A., BALBACH, N., WEHNER, M., SPRINGMANN, U., WICK, C., GRUNDIG, C., BÜTTNER, A., PUPPE, F. Ocr4all— an open-source tool providing a (semi-) automatic OCR workflow for historical printings. *arXiv preprint* arXiv:1909.04032, 2019.

[25] CHERNYSHOVA, Y.S., GAYER, A.V., SHESHKUS, A.V. Generation method of synthetic training data for mobile OCR system. In: *Tenth international conference on machine vision 2017, ICMV*, vol. 10696, id. 106962G. SPIE, Vienna, 2018. 10.1117/12.2310119

[26] MARGNER, V., PECHWITZ, M. Synthetic data for Arabic ocr system development. In: *Proceedings of the sixth international conference on document analysis and recognition*, 2001. IEEE, 2001, pp 1159–1163.

[27] EGER, S., VOR DER BRÜCK, T., MEHLER, A.A comparison of four character-level string-to-string translation models for (OCR) spelling error correction. *Prague Bull. Math. Ling.* 105, 2016, pp. 77–99.

[28] LLOBET, R., CERDAN-NAVARRO, J.R., PEREZ-CORTES, J.C., ARLANDIS, J. OCR post-processing using weighted finite-state transducers. In: *2010 20th International Conference on Pattern Recognition*, 2010, pp. 2021–2024.

[29] CACHO, F., RAMON, J. Improving OCR Post Processing with Machine Learning Tools (2019). *UNLV Theses, Dissertations, Professional Papers, and Capstones.* 3722. Available: http://dx.doi.org/10.34917/16076262

[30] REUL, C., SPRINGMANN, U., WICK, C., AND PUPPE F. Improving OCR accuracy on early printed books by utilizing cross fold training and voting. In: *Proceedings of the 13th IAPR International Workshop on Document Analysis Systems* (DAS'18), 2018. IEEE, pp. 423–428.

[31] GÉNÉREUX, M., STEMLE, E.W., LYDING, V., NICOLAS, L. Correcting OCR errors for German in Fraktur font. In: *The First Italian Conference on Computational Linguistics CLiC-it 2014 Proceedings*, 2014, pp.186–190.

[32] VOBL, T., GOTSCHAREK, A., REFFLE, U., RINGLSTETTER, C., SCHULZ, K.U. Pocoto—an open source system for efficient interactive postcorrection of OCRed historical texts. In: *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, ACM, 2014, pp. 57–61.

[33] ENGLMEIER, T., FINK, F., SCHULZ, K.U. AI-PoCoTo—combining automated and interactive OCR postcorrection. In: *Proceedings of the Third International Conference on Digital Access to Textual Cultural Heritage*, ACM, 2019, pp.19-24.

[34] HÄMÄLÄINEN, M., HENGCHEN, S. From the Past to the Future: a fully automatic NMT and word embeddings method for OCR post-correction. In: *Recent Advances in Natural Language Processing*, INCOMA, 2019, pp. 432–437.

[35] REYNAERT, M. Ocr post-correction evaluation of early Dutch books online-revisited. In: *Proceedings of the tenth International Conference on Language Resources and Evaluation* LREC, 2016, pp. 967–974.

[36] CRISTEA, D., PĂDURARIU, C., REBEJA, P., ONOFREI, M. From Scan to Text. Methodology, Solutions, and Perspectives of Deciphering Old Cyrillic Romanian Documents into the Latin Script. In: *Knowledge, Language, Models*, Bulgaria, 2020, pp. 38-56.

[37] CRISTEA, D., REBEJA, P., PĂDURARIU, C., ONOFREI, M., SCUTELNICU, A. Data Structure and Acquisition in DeLORo – a Technology for Deciphering Old Cyrillic-Romanian Documents. In: *Proceedings of ConsILR* , Ed. Universității "Alexandru Ioan Cuza" din Iași, 2022, pp.115-122.

[38] MĂRĂNDUC, C., PEREZ, C. A. A Romanian dependency treebank. In: *The International Journal of Computational Linguistics and Applications* 6(2), 2015, pp.25-40.

[39] IONESCU R.T., POPESCU M., CAHILL A. String kernels for native language identification: Insights from behind the curtains. In: *Computational Linguistics*, 42(3), 2016, pp. 491-525.

[40] BOIAN, E., CIUBOTARU, C., COJOCARU, S., COLESNICOV, A., MALAHOV, L. Digitizarea, recunoaşterea şi conservarea patrimoniului cultural-istoric. *Revista Akademos*, nr. 1 (32), 2014, pp.61-68.

[41] CERETEU, I. Cartea Românească Veche în Basarabia: Istorie, Circulație, Valoare Documentară. *Editura Academiei Române*, București, 2019, pp. 25-47, pp.81-150.

[42] Valori Bibliofile, Rev. *Gazeta bibliotecarului*, Iunie-Iulie 2008, nr. 6-7, p.1.

[43] LU, S. J., TAN, C. L. Binarization of Badly Illuminated Document Images through Shading Estimation and Compensation. *Ninth International Conference on Document Analysis and Recognition*, 2007, pp. 312-316, doi: 10.1109/ICDAR.2007.4378723.

[44] HELIŃSKI, M., KMIECIAK, M., PARKOLA, T. Report on the comparison of Tesseract and ABBYY FineReader OCR engines. *IMPACT Project Report,* 2012, 13 p. https://www.digitisation.eu/fileadmin/Tool_Training_Materials/Abbyy/PSNC_Tesseract-FineReader-report.pdf

[45] DESA, I., MORĂRESCU, D., PATRICHE, I., RALIADE, A., SULICĂ, I. Publicațiile periodice românești (ziare, gazete, reviste). Vol. III: Catalog alfabetic 1919–1924, București, *Editura Academiei*, 1987, pp. 235–236, 264, 368, 374, 575, 708, 1024.

[46] COJOCARU, S.; BURTSEVA, L.; CIUBOTARU, C.; COLESNICOV, A.; DEMIDOVA, V.; MALAHOV, L.; PETIC, M.; BUMBU, T.; UNGUR, S. On Technology for Digitization of Romanian Historical Heritage Printed in the Cyrillic Script. In: Conference on Mathematical Foundations of Informatics. 25-30 iulie 2016, Chişinău. Republica Moldova: "VALINEX" SRL, 2016, pp. 160-176.

[47] BOROȘ, T., ZAFIU, A. Transliterare automată din engleză în română. Aplicaţii şi rezultate. *Romanian Journal of Human - Computer Interaction*, Vol. 5, Iss. 3, 2012, pp. 1-14.

[48] ZHANG, M. HAIZHOU, L. JIAN, S. Direct Orthographical Mapping for Machine Transliteration. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, 2004, pp. 716–722.

[49] VINTILĂ-RĂDULESCU, I. Dicționar normativ al limbii române ortografic, ortoepic, morfologic și practic, *Editura Corint*, București, 2009, p. 817.

AUTHOR PUBLICATIONS ON THESIS TOPIC

[50] **BUMBU, T.**, COJOCARU, S., COLESNICOV, A., MALAHOV, L., UNGUR, S. User Interface to Access Old Romanian Documents. In: *Proceedings of the 4th Conference of Mathematical Society of Moldova CMSM4-2017*, June 25-July 2, 2017, pp. 479–482.

[51] **BUMBU, T**. Towards a Font Classification Model for Romanian Cyrillic Documents. *Computer Science Journal of Moldova*, v.29, n.3 (87), 2021, pp.291-298.

[52] COJOCARU, S., COLESNICOV, A., MALAHOV, L., **BUMBU, T.** Optical Character Recognition Applied to Romanian Printed Texts of the 18th-20th Century. In: Computer Science Journal of Moldova. 2016, nr. 1(70), pp. 106-117. ISSN 1561-4042

[53] **BUMBU, T.** On classification of 17th century fonts using neural networks. In: *Mathematics and IT: Research and Education*. 1-3 iulie 2021, Chişinău. Chișinău, Republica Moldova: 2021, pp. 95-96.

[54] **BUMBU, T.** Building a Diachronic Parallel Corpus for the Alignment of the Old Romanian Texts. In: *Proceedings of the of the Conference on Mathematical Foundations of Informatics MFOI-2019*, July 3-6, 2019, Iasi, Romania, pp. 263–269.

[55] **BUMBU, T.** Evaluarea Corpusului Diacronic Paralel cu Texte Românești din Noul Testament din 1648 & 1990. În materialele conferinței științifice a doctoranzilor *„Tendințe contemporane ale dezvoltării științei: viziuni ale tinerilor cercetători”*, ediția a 9-a, vol., 10 iunie 2020, Chișinău, pp.6-12.

[56] **BUMBU, T.** On Alignment of Textual Elements in a Parallel Diachronic Corpus. In: *Computer Science Journal of Moldova*. 2020, nr. 3(84), pp. 241-248. ISSN 1561-4042.

[57] DRUGUS, I., **BUMBU, T.**, BOBICEV, V., DIDIC, V., BURDUJA, A., PETRACHI, A., ALEXEI, V. Punctilog: A New Method of Sentence Structure Representation. In: *Proceedings of the Workshop on Intelligent Information Systems WIIS2021*, October 14-15, 2021, Chisinau, Republic of Moldova. pp. 118-129.

[58] BOBICEV, V., **BUMBU, T.**, DIDIC, V., PRIJILEVSCHI, D., MORARI, G. Punctilog Compared to Dependency Grammar and Constituency Grammar. In: *Logic and Artificial Intelligence*, Chisinau, 2023, pp. 92-106.

[59] COJOCARU, S., COLESNICOV, A., MALAHOV, L., **BUMBU, T.**, UNGUR, Ș. On Digitization of Romanian Cyrillic Printings of the 17th-18th Centuries. CSJM, vol.25, no.2 (74), 2017, pp.217-225.

[60] **BUMBU, T.**, BURTSEVA, L., COJOCARU, S., COLESNICOV, A., MALAHOV, L. Platform for Digitization of Heterogeneous Documents. In: *Conference on Applied and Industrial Mathematics CAIM 2022*. Ediţia a 29 (R), 25-27 august 2022, Chişinău. Chișinău, Republica Moldova: Bons Offices, 2022, pp. 170-171. ISBN 978-9975-81-074-6.

[61] COLESNICOV, A., MALAHOV, L., COJOCARU, S., BURTSEVA, L., **BUMBU, T.** Development of a platform for heterogeneous document recognition using convergent technology. In: *Workshop on Intelligent Information Systems. 6-8 octombrie 2022*, Chişinău: Valnex, 2022, pp. 104-107. ISBN 978-9975-68-461-3.

[62] **BUMBU, T.**, CAFTANATOV, O., MALAHOV, L. Revitalization of the RM Folkloric Texts from the Second Half of the 20th Century and their Diachronic Analysis. *ROMAI J.*, v.14, no.2 (2018), pp. 33–40.

[63] CIUBOTARU, C., DEMIDOVA, V., **BUMBU, T.** Generation of the Romanian Cyrillic lexicon for the period 1967 – 1989. In: *Proceedings IMCS-55 The Fifth Conference of Mathematical Society of the Republic of Moldova*. Chișinău. Chișinău, Republica Moldova: Tipografia Valinex, 2019, pp. 309-316. ISBN 978-9975-68-378-4.

[64] CAFTANATOV, O., **BUMBU**, **T.**, ERHAN, L., CERNEI, I., IAMANDI, V., LUPAN, V., CAGANOVSCHI, D., CURMEI, M. Discover the Moldovan Cultural Heritage through e-Moldova Portal by Using Crowdsourcing Concept. In: *Proceedings of the Workshop on Intelligent Information Systems WIIS2021*, October 14-15, 2021, Chisinau, Republic of Moldova, pp. 65-75.

[65] BOBICEV, V., **BUMBU, T.**, LAZU, V., MAXIM, V., ISTRATI, D. Folk Poetry for Computers: Moldovan Codri's Ballads Parsing. In: *PROCEEDINGS OF THE 12 TH INTERNATIONAL CONFERENCE "LINGUISTIC RESOURCES AND TOOLS FOR PROCESSING THE ROMANIAN LANGUAGE" MĂLINI,* 27-29 OCTOBER 2016, pp. 39-50.

[66] BUMBU, T., BURTSEVA, L., COJOCARU, S., COLESNICOV, A., MALAHOV, L. A Platform for Processing Heterogeneous Documents. In: *Proceedings of the the 17th International Conference "Linguistic Resources and Tools for Processing The Romanian Language"*, 10-12 November 2022, ISSN 1843-911X, pp. 141-151.

ADNOTARE

Bumbu Tudor: “Tehnologii și resurse informaționale pentru digitizarea și procesarea textelor din patrimoniul istorico-cultural”.

Teză de doctor în informatică, Chișinău, 2023.

Structura tezei: teza este scrisă în limba română și constă din introducere, 3 capitole, concluzii generale și recomandări, bibliografie din 140 de titluri. Teza conține 120 de pagini cu text de bază, 59 figuri și 10 tabele. Rezultatele obținute sunt publicate în 17 lucrări științifice.

Cuvinte-cheie: digitizare, patrimoniul de limbă română, documente chirilice, rețele neurale, modele OCR, transliterare, platformă de digitizare.

Scopul lucrării: elaborarea instrumentelor informatice pentru procesarea patrimoniului de limbă română tipărit în secolele 17-20.

Obiectivele cercetării: crearea unei colecții de resurse scanate pentru antrenarea modelelor OCR și elaborarea dicționarelor OCR; elaborarea unei tehnologii OCR pentru documentele românești tipărite în secolele 17-20; dezvoltarea algoritmilor de transliterare din grafie chirilică în cea latină pentru o varietate de alfabete; dezvoltarea unei platforme de digitizare pentru procesarea documentelor chirilice românești.

Noutatea și originalitatea științifică: constau în cercetarea și elaborarea tehnologiei pentru soluționarea problemei de recunoaștere și transliterare a documentelor chirilice românești tipărite în secolele 17-20.

Rezultatul obținut care contribuie la soluționarea unei probleme științifice importante îl constituie dezvoltarea tehnologiei de recunoaștere optică a caracterelor și transliterare din grafia chirilică în cea latină a documentelor chirilice românești tipărite în secolele 17-20, în condițiile existenței unei varietăți mari de alfabete și fonturi.

Semnificația teoretică a lucrării: este determinată de obținerea unei tehnologii care permite conversia documentelor românești din alfabetul chirilic în cel latin, cu aplicarea și dezvoltarea metodelor bazate pe rețele neurale.

Valoarea aplicativă a lucrării: constă în elaborarea unei platforme de digitizare, care aduce un aport substanțial la automatizarea reeditării documentelor vechi, fiind un instrument util pentru un cerc larg de utilizatori.

Implementarea rezultatelor lucrării: Instrumentele de digitizare au fost utilizate pentru recunoașterea parțială sau completă a unor cărți românești tipărite în alfabetul chirilic. Instrumentarul de recunoaștere și transliterare a fost instalat pentru utilizare în cadrul Bibliotecii Academiei Române și a Bibliotecii Științifice „Andrei Lupan”.

ANNOTATION

Bumbu Tudor: “Technologies and Information Resources for the Digitization and Processing of Texts from the Cultural Heritage.”

Doctoral thesis in Computer Science, Chisinau, 2023.

The structure of the thesis: the thesis is written in Romanian and consists of an introduction, 3 chapters, general conclusions and recommendations, and a bibliography of 140 titles. The thesis contains 120 pages of basic text, 59 figures, and 10 tables. The obtained results are published in 17 scientific papers.

Keywords: digitization, Romanian language heritage, Cyrillic documents, neural networks, OCR models, transliteration, digitization platform.

The aim of the paper: development of computer tools for processing printed Romanian language heritage of the 17th-20th centuries.

Research objectives: Creation of a collection of scanned resources for training OCR models and developing OCR dictionaries; Development of OCR technology for Romanian printed documents from the 17th to 20th centuries; Development of transliteration algorithms from Cyrillic to Latin script for a variety of alphabets; Development of a digitization platform for processing Romanian Cyrillic documents.

The novelty and scientific originality: are based on the research and development of technology for solving the problem of recognition and transliteration of Romanian Cyrillic documents printed in the 17th to 20th centuries.

The result obtained that contributes to solving an important scientific problem: is the development of optical character recognition technology and transliteration from Cyrillic to Latin script of Romanian Cyrillic documents printed in the 17th to 20th centuries, under the conditions of a wide variety of alphabets and fonts.

The theoretical significance of the paper: is determined by obtaining a technology that allows the conversion of Romanian documents from the Cyrillic alphabet to Latin, with the application and development of methods based on neural networks.

The practical value of the paper: is based on the development of a digitization platform, which brings a substantial contribution to the automation of republishing old documents, being a useful tool for a wide range of users.

Implementation of the paper results: Digitization tools have been used for partial or complete recognition of some Romanian books printed in the Cyrillic alphabet. The recognition and transliteration tools have been installed for use at the Romanian Academy Library and the “Andrei Lupan” Scientific Library.

АННОТАЦИЯ

Тудор Бумбу: “Технологии и информационные ресурсы для оцифровки и обработки текстов из культурного наследия”

Докторская диссертация по информатике, Кишинёв, 2023 год.

Структура диссертации: диссертация написана на румынском языке и состоит из введения, 3 глав, общих выводов и рекомендаций, библиографии из 140 наименований. Диссертация содержит 120 страниц основного текста, 59 иллюстраций и 10 таблиц. Полученные результаты опубликованы в 17 научных работах.

Ключевые слова: оцифровка, наследие румынского языка, кириллические документы, нейронные сети, модели OCR, транслитерация, платформа для оцифровки.

Цель работы: разработка компьютерных инструментов для обработки печатного наследия румынского языка XVII-XX веков.

Задачи исследования: создание коллекции сканированных ресурсов для обучения моделей OCR и разработка словарей OCR; разработка технологии OCR для румынских печатных документов XVII-XX веков; разработка алгоритмов транслитерации с кириллицы на латиницу; разработка платформы оцифровки для обработки румынских кириллических документов.

Научная новизна и оригинальность работы: основываются на исследовании и разработке технологии для решения проблемы распознавания и транслитерации румынских кириллических документов, напечатанных в XVII-XX веках в различных алфавитах.

Полученный результат, который способствует решению важной научной проблемы: разработка технологии оптического распознавания символов и транслитерации с кириллицы на латиницу румынских кириллических документов, напечатанных в XVII- XX веках, в условиях большого разнообразия алфавитов и шрифтов.

Теоретическая значимость работы: определяется получением технологии, которая позволяет конвертировать румынские документы из кириллического алфавита в латинский, с применением и разработкой методов, основанных на нейронных сетях.

Прикладная ценность работы: заключается в разработке платформы для оцифровки, которая вносит существенный вклад в автоматизацию переиздания старых документов, являясь полезным инструментом для широкого круга пользователей.

Реализация результатов работы: Инструменты оцифровки были использованы для частичного или полного распознавания ряда румынских книг, напечатанных кириллицей, инструменты распознавания и транслитерации были установлены для использования в Библиотеке Румынской Академии и научной библиотеке “Andrei Lupan”.

BUMBU TUDOR

TECHNOLOGIES AND INFORMATION RESOURCES FOR THE DIGITIZATION AND PROCESSING OF TEXTS FROM THE CULTURAL HERITAGE

121.03 − COMPUTER PROGRAMMING

Summary of the Ph.D. thesis in Computer Science

Approved for printing: 11.07.2023	Paper size 60x84 1/16
Offset paper. Offset printing.	Copies 30 ex.
Print sheets: 2,1	Order No. 65/23

The editorial-polygraphic center of the State University of Moldova,
Chisinau, 60 Alexei Mateevici str., MD-2009, Republic of Moldova