UNIVERSITATEA DE STAT DIN MOLDOVA
ȘCOALA DOCTORALĂ ȘTIINȚE FIZICE, MATEMATICE, ALE INFORMAȚIEI ȘI INGINEREȘTI

Cu titlu de manuscris
C.Z.U.: 004:[94(478):008]

BUMBU TUDOR

TEHNOLOGII ȘI RESURSE INFORMAȚIONALE PENTRU DIGITIZAREA ȘI PROCESAREA TEXTELOR DIN PATRIMONIUL ISTORICO-CULTURAL

Rezumatul tezei de doctor în informatică

121.03 PROGRAMAREA CALCULATOARELOR

Autor: Bumbu Tudor

Conducător științific: Cojocaru Svetlana, mem. cor., prof. cerc., dr. hab. în informatică.

Comisia de îndrumare: Gaindric Constantin, mem. cor., prof. cerc., dr. hab. în informatică;
Țițchiev Inga, dr. în informatică, conf. univ.;
Burțeva Liudmila, dr. în informatică, conf. cerc.

CHIȘINĂU, 2023

Teza a fost elaborată în Școala Doctorală Științe Fizice, Matematice, ale Informației și Inginerești, Universitatea de Stat din Moldova, Chișinău.

Componența comisiei de doctorat:

Președinte: LOZOVANU Dmitrii, doctor habilitat în științe fizico-matematice, profesor universitar, m.c. AȘM, Institutul de Matematică și Informatică „V. Andrunachievici", USM.

Conducător științific: COJOCARU Svetlana, doctor habilitat în informatică, profesor cercetător, m.c. AȘM, Academia de Științe a Moldovei.

Referenți oficiali:

GAINDRIC Constantin, doctor habilitat în informatică, profesor universitar, m.c. AȘM, Institutul de Matematică și Informatică „V. Andrunachievici", USM.

IFTENE Adrian, doctor, profesor universitar, Universitatea „Al. I. Cuza", Iași, România.

PETIC Mircea, doctor în informatică, conferențiar universitar, Universitatea de Stat „Alecu Russo" din Bălți.

Secretar științific: NOVAC Ludmila, doctor în științe fizico-matematice, conferențiar universitar, Universitatea de Stat din Moldova.

Susținerea va avea loc la 19 septembrie 2023, ora 14:00, la Institutul de Matematică și Informatică „Vladimir Andrunachievici", str. Academiei 5, bir. 340.

Teza de doctor şi rezumatul pot fi consultate la Biblioteca Universității de Stat din Moldova şi la pagina web a Agenției naționale de Asigurare a Calității în Educație și cercetare (www.cnaa.md)

Rezumatul a fost expediat la _____

Secretar științific al Comisiei de doctorat _____ Novac Ludmila, doctor în științe fizico-matematice, conferențiar universitar

Autor _____ Bumbu Tudor

© Bumbu Tudor, 2023

CUPRINS

CUVINTE CHEIE

Cuvinte cheie: Digitizare, documente chirilice românești, patrimoniul istorico-cultural, tehnologie OCR, transliterare, alinierea textelor, clasificarea fonturilor vechi, rețele neurale, platformă de digitizare.

1. SCOPUL ȘI OBIECTIVELE CERCETĂRII

Actualitatea și importanța temei de cercetare. Digitizarea ocupă un loc de frunte în tehnologiile secolului XXI. Încă în anul 2011 Comisia Europeană a venit cu un document de recomandări privind digitizarea și accesibilitatea online a materialului cultural și conservarea digitală [1], în care menționa că „dezvoltarea procesului de digitizare a materialului aflat în biblioteci, arhive și muzee ar trebui să fie încurajată în continuare, pentru a se garanta faptul că Europa își menține poziția de actor principal pe plan internațional în domeniul culturii și al conținutului creativ și că își utilizează bogăția materialului cultural, în cel mai bun mod cu putință", îndemnând statele membre să își intensifice investițiile în acest domeniu.

Recomandarea a fost inclusă drept acțiune de politici în mai multe țări (nu doar cele din UE), dezvoltându-se o întreagă industrie ce oferă servicii de scanare, recunoaștere și alte activități adiacente, problema digitizării și conservării tezaurului cultural reprezentând un domeniu prioritar din agenda digitală pentru Europa [2].

Digitizarea pe scară largă, care inițial se reducea la scanarea și stocarea imaginilor, a început odată cu proiectul Gutenberg [3], inițiat în anii '70, iar mai târziu, colecția de un milion de cărți[1] și proiectul de digitizare Google Books[2]. Cu toate că aceste proiecte soluționează problema conservării patrimoniului tipărit, digitizarea prin scanare a materialelor tipărite poate fi considerată doar punctul de plecare în ceea ce privește conservarea cunoștințelor incluse în ele și facilitarea accesului către acestea.

În pofida faptului că mai multe documente pot fi găsite și citite online, ele nu pot fi procesate automat, deoarece în cele mai dese cazuri sunt expuse doar în format de imagine, și nu cel de text lizibil (sau editabil) automat, de către mașină. Prin urmare, provocarea automatizării procesului de transformare a documentelor în text lizibil pentru calculator, deci editabil, revine aplicațiilor de învățare automată și viziune computerizată, și anume celor de recunoaștere optică a caracterelor (OCR – Optical Character Recognition). Vom demonstra în această lucrare că sarcina respectivă nu poate fi întotdeauna considerată drept una trivială, deoarece diapazonul de variație a materialului-sursă (calitatea documentului și volumul lui, perioada editării, condițiile de păstrare etc.) este extrem de mare. Cu toate acestea, punerea la dispoziție a documentelor de patrimoniu cultural digitizat sub formă editabilă a fost și este considerată în continuare o necesitate. Acest lucru este subliniat în mod deosebit în raportul de referință [4], care avertizează că Europa se află în pericol de a intra într-o nouă eră întunecată dacă nu se creează mijloace suficiente de conservare și facilitare a accesului la

[1] Proiectul „Million Book Collection", http://ulib.isri.cmu.edu/ (Accesat 23.03.2023).
[2] Proiectul de Digitizare „Google Books", https://books.google.com/ (Accesat 23.03.2023).

materialul de patrimoniu cultural. Ca urmare, au fost finanțate mai multe proiecte la scară largă care se ocupă de OCR-ul tipăriturilor istorice în contextul digitizării în masă, cel mai important fiind proiectul IMPACT [2, 5] care presupune îmbunătățirea accesului la text și proiectul OCR pentru tipăriturile moderne timpurii – eMOP[3].

Soluționarea acestei probleme pentru patrimoniul românesc se confruntă cu dificultăți și aspecte specifice: un număr mare de perioade în evoluția limbii, un număr relativ mic și foarte dispersat de resurse depozitate, o mare diversitate de alfabete folosite la tipărirea lor. Obstacolele întâmpinate la digitizarea și conservarea acestui tezaur țin de recunoașterea corectă a literelor chirilice, dar și de inexistența unui lexicon adecvat perioadei de tipărire a resurselor vechi [6]. În particular, problema creării resurselor lingvistice, digitizarea și procesarea textelor ce fac parte din patrimoniul cultural din diverse perioade istorice este actuală pentru mai multe țări europene [7-10].

Prin Hotărârea Guvernului Republicii Moldova nr. 857 din 31 octombrie 2013 a fost aprobată Strategia națională de dezvoltare a societății informaționale „Moldova Digitală 2020", precum și planul de acțiuni pentru punerea în aplicare a acesteia. Acest act normativ, în pofida faptului că prevederile lui nu au fost realizate integral, a impulsionat activitățile de digitizare a colecțiilor de documente din bibliotecile și arhivele țării. Cu toate acestea, rămâne actuală soluționarea problemei digitizării patrimoniului cultural tipărit al Republicii Moldova, care ar oferi instrumente informatice capabile să proceseze documente din diferite perioade istorice, cu diferite alfabete, cu diverse vocabulare, păstrate în diverse condiții etc. – instrumente inteligente, de care ar putea beneficia atât cercetătorii, cât și publicul larg, oferindu-li-se posibilitatea operării cu colecții mari de date indexate.

Scopul și obiectivele cercetării. Scopul cercetării constă în fundamentarea și elaborarea instrumentelor informatice pentru procesarea patrimoniului de limbă română tipărit în secolele 17-20. Scopul propus a determinat necesitatea formulării următoarelor obiective:

- analiza și determinarea metodelor principale de preprocesare a documentelor vechi;
- crearea unei colecții de resurse scanate pentru antrenarea modelelor OCR și elaborarea dicționarelor OCR;
- elaborarea unei tehnologii OCR a documentelor românești tipărite în secolele 17-20;
- dezvoltarea algoritmilor de transliterare din grafie chirilică în cea latină;
- cercetarea și elaborarea metodelor de aliniere a textelor vechi vocabularului contemporan, elaborarea unui suport pentru aliniere;
- dezvoltarea unei platforme de digitizare pentru procesarea documentelor chirilice românești.

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

Realizarea obiectivelor propuse a contribuit la obținerea unor rezultate aplicative importante, încorporate în platforma de digitizare, utilizarea căreia facilitează accesul la patrimoniul cultural românesc tipărit în grafie chirilică.

2. METODOLOGIA CERCETĂRII ȘTIINȚIFICE

Pe parcursul cercetărilor efectuate în cadrul tezei au fost folosite metode din domeniul prelucrării limbajului natural și învățării automate. Procesul de cercetare a fost unul complet, urmărindu-se parcurgerea riguroasă a fiecărei etape: definirea problemei, faza de documentare, emiterea ipotezelor de lucru, faza de testare, analiza rezultatelor și diseminarea lor.

Faza de documentare și emiterea ipotezelor de lucru se bazează pe scopurile și obiectivele cercetării. Unul din principalele obiective a fost elaborarea tehnologiei OCR a documentelor românești tipărite în secolele 17-20. Atingerea acestui obiectiv a necesitat investigarea în profunzime a tehnicilor OCR existente și adaptarea lor la specificul textelor românești vechi. S-a ținut cont de faptul că aceste texte cuprind o varietate de fonturi (mai ales cele din secolul XVII), de formatul de tipărire și calitatea documentului scanat, de specificul lingvistic, cu o ortografie și o sintaxă diferite de cele contemporane.

În faza de testare, s-au efectuat experimente folosind diverse seturi de antrenare și dicționare de cuvinte la învățarea modelelor OCR cu ajutorul a două versiuni ale softului ABBYY FineReader pentru a identifica cea mai potrivită abordare pentru OCR. Pe lângă testarea OCR, au fost testate și regulile din algoritmul de transliterare din grafia chirilică în cea latină pe baza lexicoanelor chirilice generate.

Analiza rezultatelor a constat în evaluarea performanței modelelor OCR antrenate și a comparației rezultatelor cu seturi de testare. De asemenea, au fost efectuate comparații între performanțele modelelor cu și fără dicționare de cuvinte.

Rezultatele cercetării au fost diseminate printr-o serie de publicații în reviste de specialitate și prezentări la conferințe naționale și internaționale. Acest lucru a permis un schimb de idei cu alți cercetători și a deschis calea spre îmbunătățiri ulterioare ale metodelor și tehnicilor folosite.

Noutatea și originalitatea științifică a lucrării constau în cercetarea și elaborarea tehnologiei pentru soluționarea problemei de recunoaștere și transliterare a documentelor chirilice românești tipărite în secolele 17-20, care permite procesarea eficientă și rapidă a documentelor menționate. Gradul de noutate și originalitate este reprezentat de:

- elaborarea tehnologiei OCR a documentelor românești tipărite în secolele 17-20;

- dezvoltarea algoritmilor de transliterare din alfabetul chirilic românesc în alfabetul modern (latin) românesc;
- elaborarea unei metode de clasificare a fonturilor utilizate la tipărirea textelor vechi;
- elaborarea unei metode de aliniere a textelor vechi la vocabularul modern utilizând tehnici de similaritate ale secvențelor.
- dezvoltarea unei platforme web de digitizare pentru procesarea documentelor chirilice.

Problema științifică importantă rezolvată în domeniul de cercetare este 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 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 biblioteci și arhive în crearea conținutului digital, pentru cercetătorii din domeniul istoriei, filologiei, etc., dar și pentru un cerc larg de utilizatori, oferindu-le asistență la etapele de preprocesare, recunoaștere și postprocesare a documentelor.

Aprobarea rezultatelor cercetării. Rezultatele științifice obținute de autor în această teză au fost prezentate la conferințe științifice naționale și internaționale și au fost publicate în reviste recenzate. Principalele rezultate incluse în teză au fost prezentate la următoarele conferințe științifice:

- *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.* Proceedings of 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.* Proceedings IMCS-55 The Fifth Conference of Mathematical Society of the Republic of Moldova. 28 septembrie - 1 octombrie 2019, Chișinău. Chișinău, Republica 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*. Proceedings of the 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*. Conferința științifică a doctoranzilor „Tendințe contemporane ale dezvoltării științei: viziuni ale tinerilor cercetători", ediția a 9-a, vol., 10 iunie 2020, Chișinău;
- *Building a Diachronic Parallel Corpus for the Alignment of the Old Romanian Texts.* Proceedings of the Conference on Mathematical Foundations of Informatics MFOI-2019, July 3-6, 2019, Iasi, Romania.

Publicații pe tema cercetării. Rezultatele obținute în teză sunt publicate în 17 lucrări științifice (a se vedea [50-66]): 5 articole în reviste științifice (a se vedea [51, 52, 56, 59, 62]) dintre care 2 articole de un singur autor (a se vedea [51, 56]); 12 lucrări la conferințe internaționale (a se vedea [50, 53-55, 57, 58, 60, 61, 63, 64, 65, 66]).

Volumul și structura tezei. Teza este scrisă în limba română, tehnoredactată la calculator, cu titlu de manuscris. Lucrarea are următoarea structură: introducere, trei capitole, concluzii generale și recomandări, adnotări în limba română, rusă și engleză, bibliografie ce cuprinde 140 de titluri. Volumul total al tezei este de 136 de pagini, dintre care 115 pagini text de bază.

3. SINTEZA CAPITOLELOR

În compartimentul de **Introducere** se evidențiază relevanța și importanța temei de cercetare, prezentându-se informații concise și actualizate despre stadiul recent al digitizării patrimoniului istorico-cultural. Sunt definite scopul și obiectivele tezei, noutatea științifică a rezultatelor obținute, precum și valoarea teoretică și aplicativă a tezei, acestea fiind însoțite de demonstrarea și validarea rezultatelor.

Primul capitol, „Instrumente și metode de procesare a documentelor istorice", are un caracter introductiv și conține o analiză a studiilor științifice referitoare la metodele, instrumentele și resursele pentru digitizarea documentelor din patrimoniul istorico-cultural. Se definesc conceptele de

patrimoniu digital [11], digitizare a documentelor vechi, preprocesare a imaginii din documente vechi tipărite, recunoaștere optică a caracterelor (OCR), adevăr de bază, metrici de evaluare a acurateței OCR, postprocesare.

Ulterior se descriu metodele și instrumentele pentru recunoașterea optică a caracterelor (OCR, în continuare și *recunoaşterea*) în documentele istorice. Multe documente de acest fel au fost scanate și stocate în baze de date și portaluri, iar recunoașterea lor este esențială pentru a le face accesibile. Se menționează faptul că recunoașterea documentelor moderne tipărite este foarte eficientă, cu o acuratețe de peste 99%, datorită asemănării dintre caracterele învățate și cele recunoscute, separării clare a caracterelor de fundal și ortografiei moderne a cuvintelor. Documentele istorice, însă, constituie o provocare serioasă pentru OCR. Există două metode principale de antrenare a modelelor OCR: antrenarea pe date sintetice (imagini generate din text electronic și fonturi disponibile pe computer) și antrenarea pe date reale (perechi de forme de glifă sau imagine a caracterului și transcripția acestuia - caracterul Unicode) [12]. În timp ce antrenarea pe date sintetice este mai eficientă, calitatea recunoașterii este mai scăzută pentru documentele istorice comparativ cu antrenarea pe date reale [13]. Se identifică două probleme majore în aplicarea tehnologiei de recunoaștere ale documentelor istorice tipărite: necesitatea de a antrena un model specific pentru fiecare carte și dificultatea de a transfera acuratețea modelului de la o carte la alta. Pentru a soluționa aceste probleme, se experimentează algoritmi de recunoaștere care se bazează pe rețele neurale recurente [14]. Modelele individuale oferă o acuratețe excelentă pe cărțile pe care au fost instruite, dar nu pot fi generalizate cu succes pe alte documente. O soluție o constituie antrenarea modelelor mixte, care folosesc pentru instruire o varietate de documente tipărite la diferite tipografii cu scopul de a obține o generalizare mai bună. Antrenarea modelelor mixte poate depăși bariera tipografică, astfel că rezultatele OCR pot fi folosite pentru a antrena modele OCR mai eficiente din punct de vedere al acurateței. Multe lucrări axate pe recunoașterea documentelor istorice se bazează pe Tesseract [15, 16]. Se demonstrează că motorul OCR FineReader poate oferi o acuratețe mai bună (la nivel de caractere) decât Tesseract [17]. Din aceste considerente s-a decis folosirea motorului FineReader la recunoașterea documentelor chirilice românești. Se ilustrează utilizarea Ocropy [18], o metodă OCR pentru procesarea documentelor istorice, care posedă o acuratețe considerabilă. Se prezintă softul Calamari [19], un set de instrumente OCR care depășește performanța Ocropy prin

utilizarea unei arhitecturi de rețele neurale CNN[4]-LSTM[5] bazate pe TensorFlow[6]. Calamari îmbunătățește performanța de calcul, în special pe un GPU, și prezintă funcții suplimentare, cum ar fi oprirea timpurie a procesului de antrenare, validarea încrucișată și preantrenarea. Comparativ cu Ocropy, Calamari se dovedește a fi mai rapid și mai eficient [20]. În teste, Calamari a fost antrenat pe un corpus de ziare istorice din Finlanda[7], oferind o acuratețe a caracterelor între 87% și 92%. Seturile de date pe care se antrenează rețelele neurale din Ocropy și Calamari constau din *linii de text*, ci nu din glife individuale, așa cum sunt formate seturile de date pentru FineReader și Tesseract. Se descriu performanțele prin validare încrucișată cu rata de eroare a caracterelor (CER) și rata de eroare a cuvintelor (WER) pentru a măsura eficiența metodei. Rezultatele experimentelor au indicat că modelele mixte au realizat o eroare medie de 2,6% CER și 10% WER. Aplicarea mecanismului de votare conduce la rezultate îmbunătățite, iar corectarea post-recunoaștere a erorilor îmbunătățește și mai mult acuratețea. Noile versiuni ale FineReader și Tesseract, de asemenea, folosesc învățarea profundă, cu rețele multistrat de tip CNN și LSTM.

În continuare sunt analizate o serie de platforme sau cadre de digitizare și procesare a documentelor istorice. Un exemplu important este *Historical Document Processing and Analysis Framework (HDPA[8])*, descris în lucrarea [21], un cadru web complex pentru gestionarea și analiza documentelor istorice, cu accent pe OCR (Optical Character Recognition). HDPA este gratuit pentru cercetare și eficient în pregătirea seturilor de date pentru OCR. HDPA are opt module, care facilitează preprocesarea și segmentarea imaginii, crearea setului de date pentru antrenarea modelului OCR și recunoașterea propriu-zisă. Cadrul este dezvoltat în Django, permițând elaborarea modulelor Python individuale. HDPA nu include un modul de postprocesare, dar admite o integrare simplu de efectuat a modulelor noi, permițând personalizarea sistemului pentru nevoile specifice ale utilizatorului. Modulul OCR din HDPA se bazează pe tehnologii de învățare automată, folosind rețele CNN pentru extragerea caracteristicilor și o rețea neurală recurentă LSTM bidirecțională pentru recunoașterea secvențială a liniilor de text. Una dintre aceste platforme este și Aletheia [22], cu o focalizare pe analiza aspectului paginii documentului și segmentarea paginii, identificând și clasificând regiunile de interes într-o imagine scanată a unui document text. Procesul include detectarea și etichetarea diferitelor blocuri, cum ar fi blocuri de text, ilustrații, simboluri matematice și tabele. Aletheia poate

[4] O rețea neuronală convoluțională (CNN sau ConvNet) este o clasă de rețele neuronale artificiale, cel mai frecvent aplicată pentru analiza și recunoașterea imaginilor.
[5] Long short-term memory (LSTM) este o rețea neuronală artificială recurentă utilizată în domeniile inteligenței artificiale și învățării profunde. O astfel de rețea neuronală (recurentă) poate procesa nu numai puncte de date individuale (cum ar fi imagini), ci și secvențe întregi de date (cum ar fi audio sau video).
[6] https://www.tensorflow.org/learn
[7] „Digital Materials of Finland: The newspaper collection.”, https://digi.kansalliskirjasto.fi/search?formats=NEWSPAPER (Accesat 15.06.2022).
[8] Cadrul HDPA este disponibil gratuit la adresa http://ocr-corpus.kiv.zcu.cz/ (Accesat 20.06.2022).

detecta automat obiecte pe patru niveluri: regiuni de interes, linii de text, cuvinte și glife. Crearea seturilor de date cu adevăr de bază, stocate în formatul PAGE XML, este o altă caracteristică a platformei [23]. Transkribus[9], o altă platformă, a fost dezvoltată la Universitatea din Innsbruck și include instrumente pentru recunoașterea, transcrierea și căutarea documentelor istorice. Totuși, acesta nu oferă un suport pentru generarea de date sintetice, o funcție disponibilă în HDPA. Se menționează aici și proiectul *OCR-D*[10] dezvoltat în Germania, care include 8 module specializate pentru diferite etape ale OCR. În cadrul acestui proiect a fost creată și platforma OCR4all, un instrument open-source pentru procesarea semi-automată a documentelor istorice [24]. Există și alte instrumente mai specializate, cum ar fi cele pentru generarea de seturi de date OCR artificiale pentru limbile rusă, arabă și română [25, 26, 52]. Cu toate acestea, ele sunt limitate la anumite sarcini și nu iau în considerare procesul integral de digitizare. Prin urmare, valoarea platformelor de digitizare, care oferă un spectru mai larg de funcționalități, este mai mare, ceea ce ne-a convins să lucrăm în capitolul 3 asupra unei astfel de platforme.

De asemenea se analizează postprocesarea OCR – etapă vitală în verificarea și îmbunătățirea textului recunoscut de un motor OCR, adăugând valoare sistemului prin creșterea robusteții și utilității pentru digitizarea documentelor istorice. Abordările diferă, unele considerând postprocesarea ca pe o sarcină de corectare ortografică, în timp ce altele, cum ar fi metoda secvență la secvență, pot utiliza un dicționar sau lexicon pentru a detecta și corecta erorile OCR [27-30]. Majoritatea metodelor de postprocesare includ cel puțin doi pași: generarea candidaților pentru înlocuirea erorilor și luarea deciziei de a accepta corecturile. Alte abordări adaugă pași suplimentari, cum ar fi extinderea dicționarului de cuvinte și clasarea candidaților pe baza regulilor analitice, precum în lucrarea [31] în care se descrie postprocesarea documentelor în limba germană. Astfel de abordări dau rezultate bune atunci când distanța Levenshtein între tokenul candidat și rezultatul corect nu era mai mare de 2. Postprocesarea OCR manuală poate oferi rezultate de înaltă calitate, dar necesită timp, efort și cunoștințe de specialitate, în particular pentru documentele istorice cu alfabete vechi. Abordări semiautomate, cum ar fi PoCoTo [32], un instrument pentru corectarea semiautomată a textului OCR, fac această sarcină mai ușoară și mai eficientă. O versiune îmbunătățită și complet automatizată, A-PoCoTo [33], a fost dezvoltată în 2019. Alte abordări includ gruparea erorilor OCR în spațiul vectorial, cum se prezintă în [34, 35], unde se folosește un model Word2Vec pentru a obține grupuri de erori și sinonime ale cuvintelor.

În continuare se analizează proiectul DeLORo [36, 37], dedicat procesării textelor istorice tipărite în limba română cu caractere chirilice și transliterării acestora în caractere latine. În cadrul

[9] Proiectul „Transkribus”, https://readcoop.eu/transkribus (Accesat 26.05.2023).
[10] „DFG-funded Initiative for Optical Character Recognition Development”, https://ocr-d.de/ (Accesat 25.06.2022).

acestui proiect a fost dezvoltat un instrument online de adnotare a imaginilor din documente chirilice românești, numit OOCIAT, și a fost creat un corpus important, numit ROCC, care include 367 de pagini de documente istorice scanate, adnotate cu text transcris. Corpusul ROCC conține documente din secolele XVI-XIX, organizate în funcție de dificultate, tipul de scriere și nivelul de adnotare. Se recunoaște că pentru antrenarea rețelelor neurale, sunt necesare seturi mari de date. Pentru aceasta se folosesc atât adnotările efectuate prin intermediul interfeței OOCIAT, inclusiv și corpusul UAIC-RoDia Treebank [38]. Experimentele de antrenare a modelului OCR au folosit o combinație între un model statistic pentru extragerea de caracteristici și o rețea neurală cu arhitectura CNN, pentru a descoperi obiectele și a le atribui etichete. Totuși, în pofida progreselor realizate, rezultatele pentru colecțiile de manuscrise au fost mai puțin satisfăcătoare, ceea ce constituie o problemă încă neabordată de autori. Se propun și metode de separare a cuvintelor [39], folosind o abordare secvență la secvență, și se intenționează să aplice string kernels și clusterizarea spectrală pentru a grupa forme vechi de cuvinte aparținând aceleiași leme și aceleiași părți de vorbire. De asemenea, în DeLORo se planifică integrarea instrumentelor dezvoltate într-o platformă cu acces gratuit pentru cercetători.

În ultima secțiune din acest capitol se descrie necesitatea instrumentelor de digitizare și procesare a documentelor românești tipărite cu alfabete chirilice, luându-se în considerare numărul mare și diversitatea acestora. În evoluția limbii române vom distinge două epoci: veche și modernă, prima durând până în 1650 [40]. Pe teritoriul românesc prima carte a fost tipărită în 1508, iar prima în limba română - în 1535 [41]. Importante colecții de documente românești cu alfabete chirilice sunt găsite atât în bibliotecile din Republica Moldova și România, cât și în bibliotecile din alte țări, precum cele din Sankt Petersburg (Rusia). Spre exemplu, Biblioteca Academiei Române are peste 1960 de tipărituri din 1508-1830, 79% fiind în grafie chirilică [42]. Toate aceste documente necesită digitizare și procesare pentru valorificarea lor în patrimoniul românesc.

În capitolul 2 „Tehnologii de procesare a documentelor românești din sec. XVII-XX”, sunt fundamentate abordările noastre în proiectarea tehnologiei de procesare a textelor istorice (tipărite în limba română cu caractere chirilice, începând cu secolul al XVII-lea), descriindu-se metodele elaborate și argumentându-se utilizarea anumitor module din categoria celor existente. Acestea includ: module de preprocesare a imaginilor; modele OCR; modele de rețele neurale pentru clasificarea fonturilor; tehnologia de transliterare din alfabetul chirilic românesc în cel modern; suport pentru alinierea textelor vechi la cele moderne.

Inițial sunt descrise acțiunile de bază efectuate la recunoașterea textelor din secolul XVII. Procesul este organizat pe principiul utilizării tehnologiilor convergente, adică cel de interconectare în cadrul unei platforme a aplicațiilor din anumite domenii, de rând cu elaborarea componentelor proprii. Se analizează utilizarea programului ABBYY FineReader Professional (în continuare FR)

pentru recunoașterea optică a caracterelor chirilice românești. Următoarele acțiuni țin de testarea și adaptarea versiunii FR 12, dar și a unor versiuni mai noi, cum ar fi FineReader 14 și FineReader 15. Deoarece aceste versiuni nu sunt inițial orientate spre procesarea textelor vechi românești, este nevoie de extinderea capacităților acestora pentru a le adapta la soluționarea problemelor menționate. Operarea cu documente dintr-o anumită perioadă istorică a impus crearea de componente noi (cum ar fi alfabete și dicționare) și antrenarea software-ului pe seturi de date adiționale, pentru a asigura o calitate cât mai înaltă a rezultatelor. Aceste acțiuni duc la elaborarea unuia sau mai multor modele orientate spre recunoașterea textelor dintr-o anumită perioadă istorică. Procesul de recunoaștere optică a caracterelor pentru o limbă nouă în FR constă din următoarele etape: preprocesarea imaginii, care implică editarea, curățarea și ajustarea rezoluției imaginii pentru a optimiza rezultatele; crearea limbii și a alfabetului, ce va conține toate caracterele limbii noi; pregătirea și crearea unui dicționar de cuvinte pentru limba nouă; crearea și antrenarea șabloanelor, unde are loc învățarea supervizată a fiecărui caracter pentru a permite segmentarea corespunzătoare acestora din imagine.

În compartimentul dedicat procesării imaginii din documentele vechi sunt descrise particularitățile aplicării a două instrumente pentru procesarea imaginilor – cel incorporat în FR și Scan Tailor[11], elucidându-se specificul aplicării acestora la procesarea textelor vechi. Aici se conclude faptul că FR asigură unele opțiuni necesare pentru preprocesarea textelor vechi, însă nu oferă tot spectrul util, fiind necesară implicarea unor instrumente adiționale. Unul dintre modulele esențiale de preprocesare a documentelor vechi care lipsesc în FR se referă la îngroșarea caracterelor. Problema menționată apare din faptul că unele metode de binarizare a imaginii pot subția liniile din glife, iar pentru a le îngroșa din nou în etapa de preprocesare a imaginii se folosește un modul special din Scan Tailor, instrument descris în continuare în acest compartiment din teză. Convertirea imaginii în alb-negru (binarizarea) este mai performantă în Scan Tailor decât în FR. Implementarea acestui funcțional în ScanTailor se bazează pe normalizarea iluminării [43], netezirea Savitzky-Golay[12], și binarizarea propriu-zisă bazată pe Metoda lui Otsu și, la final, eliminarea marginilor întrerupte. Se menționează că este oportun să se salveze documentele în format „alb-negru”, deoarece aceasta demonstrează o acuratețe mai bună la OCR. Totuși, aplicarea acestei opțiuni cere o atenție deosebită, deoarece decolorarea poate duce la pierderea unor elemente de text. Acest lucru poate fi compensat într-o anumită măsură prin îngroșarea caracterelor. O altă particularitate este configurarea rezoluției pentru ca motorul OCR să poată detecta corect liniile de text din imagine, mai ales dacă în text persistă diacritice, accente sau alte elemente deasupra liniei de text.

[11] https://scantailor.org
[12] https://en.wikipedia.org/wiki/Savitzky–Golay_filter

Ulterior, în compartimentul despre crearea limbii utilizatorului și adăugarea dicționarului de cuvinte se menționează că unele caractere, precum ѧ și ѳ din alfabetul chirilic românesc nu există în sistemul de adăugare a alfabetului în FR și nici nu pot fi afișate de fonturile din sistemul acestuia. Prin urmare, a fost necesară identificarea și adaptarea acestora din *BabelMap*[13]. În continuare sunt descrise trei metode de extindere a dicționarului de cuvinte în procesul OCR din FR. Prima metodă implică transliterarea vocabularelor existente din alfabetul modern (latin) în cel chirilic și adăugarea acestora în dicționarul din FR; a doua metodă implică crearea dicționarului din textul unui document deja recunoscut; a treia metodă se bazează pe includerea cuvintelor din porțiuni recunoscute ale documentului din interfața grafică a FR.

În continuare sunt descrise particularitățile modelelor OCR aplicate pe texte tipărite în secolul XVII. Studiul de caz al recunoașterii optice a cărților din secolul XVII, descris în teză, a fost făcut pe cartea "Noul Testament" tipărită în anul 1648 de unde se creează seturi de date preluate din primele 257 de pagini. Seturile de date constau din caracterul tăiat din pagina și caracterul UNICODE corespunzător. Un aspect comun al cărților tipărite în secolul XVII este scrierea anumitor litere deasupra altora. De asemenea, sunt folosite abrevieri care folosesc tilde sau alte semne diacritice. Luându-se în considerare acest aspect, se propune mărirea rezoluției considerabil (peste 1200 DPI) astfel încât să se includă toate elementele unui caracter în procesul de instruire.

În continuare se discută procesul și rezultatele evaluării OCR pentru documente chirilice românești din secolul XVII. În acest scop se creează un set de date din 15 pagini ale unor cărți din această perioadă. O singură pagină din setul de pagini conține în medie 1400 de caractere și 260 de cuvinte. Criteriile de evaluare luate în considerare au fost acuratețea OCR atât la nivel de caracter, cât și la nivel de cuvânt. Acuratețea generală se calculează după metoda descrisă în [44]. În experimentele demonstrate, acuratețea OCR se calculează împreună și fără dicționarul de cuvinte. Se prezintă patru experimente cu mărirea numărului setului de date pentru fiecare experiment, unde ultimul experiment prezintă modelul OCR antrenat cu 7 pagini din setul de antrenare și numărul de glife din set este peste 3600. În acest experiment a fost găsită și cea mai bună acuratețe la nivel de caractere inclusiv cu dicționar de cuvinte, și anume 96%. Se conclude că mărirea setului de antrenare și a dicționarelor de cuvinte conduc la mărirea acurateței generale.

În continuare se abordează problema clasificării fonturilor din secolul XVII. Tipografiile din secolul XVII fonturi diferite, dintre care, se disting două fonturi total diferite, atât după stilul scrierii/tipăriturii, cât și după utilizarea caracterelor [50, 51]. Această problemă nu poate fi soluționată în FR prin antrenarea unor modele mixte, de aceea pentru fiecare font sunt antrenate modele

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

individuale. În continuare se propun unele soluții de clasificare a fonturilor. O soluție este un program de selectare a modelului OCR potrivit în funcție de tipografie [50], iar o altă soluție constă în clasificarea fonturilor utilizând rețele neurale [53], soluție care se descrie detaliat în teză. Setul de date se creează din 10 cărți scanate, selectate din Biblioteca Digitală a României[14]. La crearea setului de date se folosesc metode de clusterizare precum PCA și K-Means. Setul de date obținut constă din peste 21.200 de exemple de antrenare și peste 9 mii de exemple de testare. În continuare se antrenează rețea neurală multistrat (RNM) bazată Keras și TensorFlow pentru a clasifica caracterele în două fonturi diferite. Construirea rețelei neurale se începe cu o transformare a datelor de intrare dintr-o matrice *x* pe *y* într-un vector de lungimea $x * y$. Se adaugă un strat ascuns cu 128 de neuroni cu funcția de activare *ReLU*[15] care este complet conectat la ultimul strat. Fiind vorba de o clasificare binară, stratul de ieșire conține un singur neuron și o funcție de activare *sigmoidală*[16]. În urma antrenării se constată o acuratețe de 96.7%, acuratețe care poate fi îmbunătățită de o arhitectură mai complexă, precum CNN-LSTM unde se ia în considerare și ordinea caracterelor.

Ulterior se descrie procesul de transliterare, definindu-l ca o conversie a unui text dintr-un alfabet în altul, care implică schimbarea literelor în moduri previzibile [45]. În ceea ce privește limba română, se subliniază schimbările care au avut loc în sistemele de scriere pe parcursul istoriei [46]. Apoi se introduc mai multe metode de transliterare, inclusiv tehnici utilizate la transliterarea numelor proprii englezești în chineză, japoneză, coreeană sau arabă [47]. Se menționează, de asemenea, tehnicile de „mapare ortografică directă" care utilizează modele bazate pe n-grame pentru transliterare [48]. Standardizarea procedurilor de transliterare este esențială pentru a asigura o conversie riguroasă, univocă și complet reversibilă [49]. Procesul de transliterare în Republica Moldova a început în anul 1989, odată cu adoptarea Legii cu privire la funcționarea limbilor vorbite pe teritoriul RSS Moldovenești.

În continuare sunt descrise unele dificultăți specifice transliterării textelor tipărite cu alfabetul chirilic român, fiind pusă în evidență, în particular, problema reprezentării corecte a textului chirilic în calculator, în special ale celor din documentele din sec. XVII. Majoritatea literelor (37 din 43) sunt transliterate folosind reguli simple, independente de context, iar restul se transliterează folosind reguli dependente de context. Cea mai problematică literă este "ѧ" care poate fi transliterată ca și "a", "e", "ea", "ia". Cu toate că au fost stabilite unele reguli de dependențe contextuale la nivel de caractere, totuși sunt cazuri de excepție în care transliterarea acesteia iese din tipare. De exemplu: "чѣѧ" și "кърꙋѧ" se vor translitera în ambele cazuri cu aceeași regulă ("ѧ" => "ia"), deși cuvântul

[14] http://digitool.bibnat.ro/ (Carte românească veche şi bibliofilă/Sec. XVII)
[15] https://keras.io/api/layers/activations/#relu-function
[16] https://keras.io/api/layers/activations/#sigmoid-function

“чѣа” poate trece în “ceia” și în “ceea”. În mod ideal am avea nevoie și de reguli de dependență la nivel de cuvinte vecine. Cu toate acestea acuratețea de transliterare întrece 98%. Pe lângă reguli, se mai folosesc dicționarele de excepții pentru a putea translitera cuvintele care nu se reprezintă corect doar în baza regulilor. La sfârșitul compartimentului se discută două aplicații de transliterare din grafia chirilică în latină pentru secolele XVII-XX. O aplicație desktop este dezvoltată în Java, iar alta - cu o interfață web, dar cu funcționalități limitate.

La finalul acestui capitol se discută o serie de lucrări referitoare la alinierea textelor vechi la texte moderne. Se introduce un corpus diacronic paralel [54] și instrumente de aliniere. Alinierea unui text vechi la reprezentarea sa modernă implică traducerea acestuia într-un limbaj contemporan, înlocuindu-se variantele lexicale învechite cu expresii moderne. Textele paralele, cum ar fi acestea, sunt resurse valoroase pentru traducerea automată și analiza diacronică a limbilor naturale. Un prim pas propus în această direcție, este elaborarea unui corpus paralel diacronic de circa 8400 de propoziții, bazat pe Noul Testament tipărit în 1648 la Bălgrad aliniat la varianta sa electronică modernă din 1990. În continuare se discută instrumente de aliniere a cuvintelor [55, 56], programe software care ajută la alinierea cuvintelor dintr-un text sursă cu cele dintr-un text țintă. Instrumentele analizate sunt *Berkeley Word Aligner*[17] (BWA), un program Java care folosește modele Markov ascunse (HMM) pentru alinierea cuvintelor într-un corpus paralel la nivel de propoziții și *GIZA* ++[18], un instrument care folosește modelele HMM pentru alinierea textelor. Luând în considerare că obiectul nostru de studiu sunt texte paralele diacronice, a fost decis să se creeze un instrument special de aliniere, care include funcționalități specifice precum calculul scorului BLEU între texte, propoziții, expresii și cuvinte, vizualizarea interactivă de *n*-grame și altele. Instrumentul dezvoltat este o aplicație web bazată pe Django, cu trei module principale: modulul de editare a textului paralel și de formare a corpusului paralel, modulul de procesare a textului și modulul de învățare automată. Se planifică extinderea acestui instrument cu o componentă de adnotare a textului folosind metodologia Punctilog [57, 58] și funcționalități care folosesc scorul BLEU pentru a evalua și îmbunătăți similitudinea dintre textele paralele diacronice.

Capitolul 3, „Platformă pentru digitizarea documentelor chirilice românești”, este capitolul final al tezei – dedicat proiectării și descrierii unei platformei care include instrumentarul de digitizare pentru documentele în limba română tipărite în grafie chirilică [71, 73, 80-82]. Platforma se prezintă ca fiind principalul rezultat aplicativ al tezei; ea permite accesul la instrumente și resurse pentru digitizarea acestor documente printr-o interfață interactivă bazată pe tehnologii web. Se descrie

[17] https://github.com/mhajiloo/berkeleyaligner
[18] https://github.com/moses-smt/giza-pp

arhitectura platformei care include patru grupuri funcționale (*G1-G4*), anume: prelucrarea imaginilor, recunoașterea optică a documentelor, transliterarea textului, salvarea și publicarea rezultatelor.

În continuare este descris fiecare grup funcțional. Un grup funcțional este alcătuit din module integrate care conțin soft-uri elaborate de autor și cele terțe. Primul grup funcțional descris se referă la preprocesarea imaginii și anume la pașii întreprinși pentru a pregăti o imagine cu text pentru ca motorul OCR să o poată analiza. Motorul OCR poate avea uneori dificultăți în interpretarea corectă a imaginilor care sunt încețoșate, distorsionate sau au un contrast scăzut. Preprocesarea ajută la îmbunătățirea preciziei OCR prin pregătirea imaginii pentru a fi mai potrivită pentru recunoaștere. Unele module importante incluse în acest grup funcțional sunt: ajustarea contrastului sau a luminozității imaginii pentru a îmbunătăți citirea textului; binarizarea, care implică convertirea imaginii în varianta alb si negru, pentru îmbunătățirea contrastului; îndepărtarea zgomotului prin îndepărtarea pixelilor negri suplimentari din imagine care pot duce la recunoașterea unor caractere inutile, precum unele semne de punctuație; corectarea distorsiunii care implică rotirea imaginii pentru a o alinia corect. Prin preprocesarea imaginii înainte de a o trimite motorului OCR, de regulă, se poate îmbunătăți precizia si fiabilitatea procesului OCR. În acest grup funcțional sunt integrate module de preprocesare din soft-urile Scan Tailor, FineReader 15, și pachetul Python – OpenCV. Scopul grupului funcțional *G1* este de a pregăti documentul pentru OCR.

În continuare se descrie grupul funcțional G2, care se ocupă cu recunoașterea optică a documentelor. Acest grup include module de selectare a modelului OCR în dependență de perioada istorică; folosire a unui dicționar de cuvinte la recunoaștere; editare a textului recunoscut, folosire a unui dicționar de excepții OCR. Motorul OCR se bazează pe FineReader 15. Acest grup este inițializat cu 8 modele OCR antrenate cu seturi de date colectate din documente tipărite în secolele XVII, XVIII, XIX și XX. Utilizatorii pot adăuga, de asemenea, modele OCR noi. Aceste modele au formatul FineReader XML (fișiere *.fbt*) care conțin configurațiile modelului OCR, setul de date de antrenare, alfabetul necesar, precum și dicționarele de cuvinte. Acțiunile din grupul G2 încep cu selectarea perioadei documentului. Deci, în mod obișnuit, utilizatorul cunoaște perioada istorică de tipărire a documentului care urmează să fie supus digitizării, mai mult ca atât, poate indica chiar și anul când a fost tipărit. Totuși, nu este exclus și cazul, când utilizatorul are câteva imagini dintr-un document aleatoriu despre care nu știe nimic mai mult decât faptul că acest document este în grafie chirilică. Pentru astfel de cazuri ar fi util un modul de detectare automată a perioadei. O abordare ce poate fi utilă în soluționarea acestei probleme este experiența de detectare a fonturilor din documentele chirilice tipărite în secolul XVII, unde anumite modele de rețele neurale au fost antrenate să recunoască automat fontul documentului. Un modul care se ocupă de detectarea fonturilor din secolul XVII, este inclus în G2. Se discută în continuare procesul de recunoaștere a documentelor care poate

fi împărțit în mai multe părți ce pot fi executate în paralel pentru a mări viteza de procesare. Acest proces se gestionează prin ABBYY Hot Folder[19] (în continuare Hot Folder). Acest lucru poate fi realizat prin utilizarea mai multor instanțe pentru fiecare model OCR, împărțind astfel imaginile preprocesate în mai multe dosare, ceea ce poate îmbunătăți eficiența atunci când mai mulți utilizatori lucrează simultan pe platformă. Recunoașterea unei singure pagini cu text în medie durează 30 de secunde, cu toate că uneori recunoașterea unei astfel de pagini ar putea lua până la două minute. Aceasta se datorează faptului că instanțele create în Hot Folder verifică la fiecare minut (aceasta este opțiunea minimă de timp în Hot Folder) dacă au apărut imagini prelucrate noi în dosarele cu imagini prelucrate. Un document PDF cu 50 pagini de text se va recunoaște în circa 90 de secunde; un PDF cu 100 pagini de text – în 150 de secunde; PDF cu 360 pagini de text - peste 385 de secunde (mai mult de 6 minute). Documentele-text în format PDF atestă durata de aproximativ 1.2 secunde per pagină. La PDF-urile cu imagini nu a fost observată o durată stabilă. Criteriul acurateței OCR la nivel de caractere și la nivel de cuvinte este analizat în capitolul 2 al tezei. De exemplu, modelul OCR pentru secolul XX ne dă o acuratețe la nivel de caractere de peste 98%; modele din secolul XVIII oferă peste 92% la nivel de cuvinte; iar modelul pentru secolul XVII oferă o acuratețe de peste 95% la nivel de caractere și dicționare de cuvinte, luând în considerare preprocesarea adecvată a imaginii, calitatea de scanare a documentului, uzura acestuia etc. În continuare se demonstrează utilizarea dicționarelor de cuvinte folosite în interiorul motorului OCR și a unor dicționare de excepții OCR care constau din tupluri formate dintr-o expresie care conține „ambiguități” de recunoaștere și varianta corectă a acestei expresii. Am utilizat sintagma “ambiguități de recunoaștere” din simplu motiv că unele litere au o similaritate grafică extrem de apropiată, iar uneori motorul OCR recunoaște cu o probabilitate foarte mare varianta greșită. În așa caz, dicționarul intern nu poate propune candidatul corect chiar dacă varianta corectă s-ar fi aflat în dicționar. De exemplu, litera **и** este confundată cu litera **н** în expresia „сърачїн”, respectiv dicționarul de excepții OCR ar putea conține excepția: (сърачїн, сърачїи). Pentru a trata astfel de situații se include în G2 o componentă de postprocesare OCR prin folosirea dicționarului de excepții. Dicționare de excepții sunt folosite și la transliterare, iar un modul similar avem și în G3. În cadrul grupului funcțional G2 este inclus și un modul de editare a textului recunoscut. Acest editor de texte dispune de o tastatură virtuală web care își adaptează compoziția caracterelor în funcție de perioada documentului. Există, de asemenea, un modul dedicat gestionării tastaturilor virtuale pentru desktop.

În următoarea secțiune sunt prezentate modulele de transliterare incluse în grupul funcțional G3. Grupul de transliterare a textului, de rând cu operația propriu-zisă, oferă actualizarea ortografiei,

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

gestionarea dicționarelor de excepții pentru transliterarea și corectarea automată a textului. Transliterarea este posibilă prin două căi. Prima cale este folosirea aplicației web de transliterare *AAConv*[20], iar cea de a doua posibilitate este folosirea aceleiași aplicații doar că în variantă desktop. O diferență notabilă între aceste două variante este că varianta web poate accepta doar până la 1.2MB de text la o singură procesare. Un modul important pentru utilizator îl constituie actualizarea ortografiei, unde la solicitare se iau în considerare normele scrierii limbii române moderne. Un exemplu este scrierea cu **â** (din **a)**, inclusă ca opțiune în procesul de transliterare. Potrivit recomandărilor Academiei Române, litera "î" va fi întotdeauna scrisă la începutul și sfârșitul cuvântului ("început", "înger", "în", "întoarce", "a coborî", "a urî"). În interiorul cuvântului, de obicei este scris "â" ("cuvânt", "a mârâi"). Totuși, există câteva excepții pentru această regulă. În continuare se discută un modul din G3 pentru folosirea dicționarului de excepții. Dicționarul cu excepții de transliterare păstrează cuvinte care nu pot fi transliterate corect utilizând doar regulile de transliterare. De exemplu, cuvântul "амязэ" conform regulilor de transliterare trece in "amează", varianta corectă fiind "amiază", aceasta regăsindu-se în dicționarul respectiv. În acest modul este posibilă gestionarea listei de excepții. Excepțiile se tratează după transliterarea textului din grafia chirilică în cea latină conform regulilor, dar înainte de vizualizarea și verificarea textului în editorul de texte. Mai multe excepții provin de la scrierea diferită a cuvintelor de origine străină, în special a substantivelor proprii.

De asemenea, grupul *G3* utilizează același modul de editare a textului precum și grupul G2, iar tastatura virtuală și dicționarele de cuvinte pentru verificatorul ortografic sunt adaptate la textul transliterat. Aici se are în vedere că tastatura virtuală conține literele alfabetului românesc modern, iar dicționarul de cuvinte este scris cu alfabetul românesc modern. Un modul experimental descris în G3 este corectarea textului transliterat cu un sistem de inteligență artificială de top: GPT-3[21] dezvoltat de OpenAI[22]. În acest modul se experimentează cu modelul *text-davinci-003* (în continuare *davinci*) pentru corectarea textului recunoscut.

În continuare se descrie grupul funcțional G4, cu module pentru gestionarea și publicarea documentelor, care permite salvarea textelor recunoscute/transliterate în diferite formate, descărcarea imaginilor prelucrate și publicarea documentelor digitizate în bibliotecile digitale. Un modul important din G4 este salvarea documentului digitizat în baza de date a platformei. Pe lângă stocarea textelor și a link-urilor către fișiere, se mai stochează și obiectul digitizat, care reprezintă un obiect[23]

[20] https://translitera.cc/
[21] https://en.wikipedia.org/wiki/GPT-3
[22] https://en.wikipedia.org/wiki/OpenAI
[23] https://developer.mozilla.org/en-US/docs/Web/JavaScript/Reference/Global_Objects/Object

JavaScript cu ajutorul căruia se păstrează starea fiecărui pas făcut prin aplicația de digitizare descrisă în următoarea secțiune. Obiectul include parametrii de preprocesare, parametrii de recunoaștere și transliterare, textul recunoscut și editat, textul transliterat și editat. Un set de module incluse în G4 se referă la publicarea documentului digitizat. Un modul de publicare este bazat pe portalul eMoldova[24], în particular bazat pe un portlet[25] numit Tezaurul Național Digital[26].

În ultimul compartiment din capitolul 3, este descrisă o aplicație de digitizare din cadrul platformei ca fiind o instanță demonstrativă a unor module implementate în platformă. Scopul elaborării acestei aplicații este demonstrarea funcționalului unor module din platformă. Aplicația de digitizare integrată pe platformă permite digitizarea în 7 pași a documentului procesat, iar durata de timp pentru un ciclu complet de digitizare variază între 2 și 15 minute, în funcție de volumul documentului.

[24] https://emoldova.org/
[25] https://fr.wikipedia.org/wiki/Portlet
[26] https://digi.emoldova.org/

CONCLUZII GENERALE

Suportul procesului de revitalizare a patrimoniului cultural-istoric rămâne o problemă, actualitatea și importanța căreia constituie o prioritate menționată în mai multe documente de politici ale țărilor europene. Prin realizarea obiectivelor stabilite în cadrul prezentei teze au fost aduse anumite contribuții pentru facilitarea digitizării și transliterării textelor tipărite în limba română cu caractere chirilice, fiind acoperit un segment temporal al ultimelor patru secole. Prin analiza și dezvoltarea metodelor și instrumentelor folosite în preprocesarea imaginilor, elaborarea modelelor OCR etc., înglobate în platforma de digitizare este înlesnit accesul la documentele vechi chirilice românești, deschizând noi oportunități pentru cercetarea și valorificarea resurselor culturale și istorice.

Studiul rezultatelor obținute permite formularea următoarelor concluzii generale:

- Analiza instrumentelor și metodelor de digitizare a documentelor istorice relevă o multitudine de metode, procedee, resurse și platforme disponibile pentru preprocesarea, recunoașterea, postprocesarea și transliterarea documentelor istorice, care se deosebesc prin acuratețea și eficiența operațională [6-40].
- Metodele de recunoaștere bazate pe antrenarea OCR pe imagini cu linii de text sporesc viteza de antrenare a motoarelor OCR, contribuind astfel la accelerarea procesului de digitizare, acordându-i acestuia un caracter de masă [18-20].
- În rezultatul adaptării componentelor sistemului software FR15 pentru recunoașterea tipăriturilor vechi românești s-a constatat că acuratețea modelului crește semnificativ odată cu creșterea numărului de pagini de antrenare. Evaluarea procedurii de învățare în cadrul unui proces iterativ a demonstrat că odată cu creșterea numărului datelor de antrenare crește semnificativ și acuratețea modelului, atingând valori acceptabile (0.96 în cazul operării cu dicționar și 0.95 în cazul operării fără dicționar) la nivel de recunoaștere corectă a caracterelor chiar și după un număr nu prea mare de pagini (5-7 pagini). La nivel de cuvinte valoarea acurateței este mai mică, fapt ce denotă necesitatea utilizării unui număr mai mare de pagini pentru instruire.
- Pentru procesarea imaginilor din documentele vechi putem utiliza instrumente existente de preprocesare, suplimentându-le pe cele incorporate în FR15 cu posibilitățile oferite de Scan Tailor, în special pentru îngroșarea caracterelor, netezirea Savitzky-Golay și eliminarea marginilor întrerupte.
- Algoritmul de clasificare a fonturilor, dezvoltat prin crearea și antrenarea unei rețele neurale multistrat [51], a demonstrat o acuratețe de peste 96%.

- Transliterarea din alfabetul chirilic român în cel modern cu utilizarea regulilor dependente de context se efectuează cu o acuratețe care întrece 98%.
- Instrumentarul de aliniere elaborat efectuează alinierea textelor vechi la cele moderne prin evaluarea și crearea similitudinii textelor paralele diacronice, bazându-se pe similaritatea șirurilor de caractere. Acest instrumentar facilitează crearea unui corpus paralel diacronic [54].
- Platforma de digitizare, arhitectura, modulele și aplicațiile căreia au fost elaborate în cadrul tezei, incluzând instrumente de preprocesare a imaginilor, modele OCR, aplicații pentru transliterarea din grafie chirilică în cea latină, module de editare a textelor recunoscute/transliterate, permite realizarea sarcinilor principale referitoare la digitizarea documentelor vechi românești într-un mod eficient și rapid [60, 61].
- Platforma de digitizare poate fi folosită ca aplicație web sau desktop și poate fi extinsă pentru a include module de digitizare pentru alte limbi. Această platformă este utilă pentru biblioteci, edituri și cercetători care dețin colecții de documente în limba română tipărite cu caractere chirilice. De rând cu acestea, existența unei astfel de platforme, în special în versiunea web, facilitează accesul la tezaurul literar-istoric și pentru publicul larg.

Recomandările pentru viitoarele cercetări și dezvoltări în domeniul digitizării documentelor chirilice românești ar putea include:

- Îmbunătățirea continuă a modelelor OCR și a algoritmilor de transliterare, prin integrarea unor tehnici noi și avansate în domeniul prelucrării limbajului natural și învățării automate, pentru a crește precizia și eficiența procesului de recunoaștere și transliterare.
- Elaborarea unei interfețe simple de antrenare a modelelor OCR în masă pentru a deschide accesul pentru cât mai mulți utilizatori. Altfel, vom putea construi modele OCR pentru intervale de timp scurte, precum și pentru majoritatea tipografiilor.
- Extinderea platformei de digitizare pentru a include și alte tipuri de documente, cum ar fi manuscrisele, hărțile sau ilustrațiile, pentru a permite accesul la o varietate mai mare de resurse culturale și istorice.
- Integrarea platformei de digitizare cu alte instrumente și resurse digitale, cum ar fi biblioteci digitale, arhive și baze de date, pentru a facilita colaborarea între cercetători și pentru a pune la dispoziție informații și resurse adiționale.

BIBLIOGRAFIE

[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, RECUNOAȘTEREA ȘI 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.

PUBLICAȚIILE AUTORULUI LA TEMA TEZEI

[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*, Chișinău: 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 17^{th}-20^{th} 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.

Theoretical significance of the paper: is determined by obtaining a technology that allows the conversion of Romanian documents from 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 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 в XVII-XX веках; 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 на 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

TEHNOLOGII ȘI RESURSE INFORMAȚIONALE PENTRU DIGITIZAREA ȘI PROCESAREA TEXTELOR DIN PATRIMONIUL ISTORICO-CULTURAL

121.03 PROGRAMAREA CALCULATOARELOR

Rezumatul tezei de doctor în informatică

Aprobat spre tipar: 11.07.2023 | Formatul hârtiei 60x84 1/16
Hârtie ofset. Tipar ofset. | Tiraj 30 ex.
Coli de tipar.: 2,1 | Comanda nr. 65/23

Centrul editorial-poligrafic al Universității de Stat din Moldova,
str. Alexei Mateevici 60, Chișinău, MD-2009, Republica Moldova