DIGITIZAREA, RECUNOAŞTEREA ŞI CONSERVAREA PATRIMONIULUI CULTURAL-ISTORIC

Dr. ***Elena BOIAN***
Dr. ***Constantin CIUBOTARU***
Dr. hab. ***Svetlana COJOCARU***
Dr. ***Alexandru COLESNICOV***
Ludmila MALAHOV
Institutul de Matematică şi Informatică al AŞM

CULTURAL AND HISTORICAL HERITAGE DIGITIZATION, RECOGNITION AND CONSERVATION

Summary. This article describes digitization of old Romanian texts, problems at their recognition, and motivates the necessity to create specific electronic resources mirroring the history of the standard Romanian language. We provide also statistics of results at recognizing a Romanian text of the 19th century by modern software, and we propose a technology for creation of linguistic lexicon for Moldavian Cyrillic script of 1967–1989, starting from modern (standard) Romanian lexicon. This technology is based on transliteration and parallel texts alignment.

Keywords: digitization, Romanian linguistic resources, text recognition, language technology, Cyrillic script, transliteration, text aligning.

Rezumat. În lucrare se abordează problemele ce apar în procesul de digitizare şi recunoaştere a textelor vechi româneşti, se argumentează necesitatea creării resurselor electronice specifice care caracterizează evoluţia limbii române moderne. Se prezintă rezultate statistice obţinute la recunoaşterea unui text românesc din secolul al XIX-lea, utilizându-se produse program moderne. Se propune o tehnologie în vederea creării lexiconului lingvistic pentru patrimoniul moldovenesc tipărit cu alfabet chirilic în perioada 1967-1989, pornind de la lexiconul românesc modern. Această tehnologie se bazează pe transliterare şi pe aliniere paralelă a textelor.

Cuvinte-cheie: digitizare, resurse lingvistice româneşti, recunoaşterea textului, tehnologia limbajului, alfabet chirilic, transliterare, alinierea textelor.

Introducere

Problema digitizării şi conservării patrimoniului istorico-lingvistic (cultural) reprezintă un domeniu prioritar din agenda digitală pentru Europa. UE evidenţiază necesitatea unui efort coordonat în domeniu şi întreprinde vaste acţiuni în vederea impulsionării acestui proces, printre care dezvoltarea bibliotecii virtuale *Europeana*, susţinută prin rezoluţia Parlamentului European din 5 mai 2010 şi adoptarea Programului de lucru pentru activităţi culturale 2011-2014. Menţionăm şi recomandările Comisiei Europene „Privind digitizarea şi accesibilitatea online a materialului cultural şi conservarea digitală" din 27 octombrie 2011.

Dezideratele principale ale politicii culturale pentru zonele unde se vorbeşte limba română ţin de studierea, valorificarea şi digitizarea patrimoniului cultural-istoric. Procesul de digitizare a patrimoniului necesită soluţionarea unui şir de probleme legate de recunoaşterea, editarea, traducerea, interpretarea, circularea şi recepţionarea textelor tipărite atât în limba română, cât şi în alte limbi moderne. Soluţionarea acestor probleme pentru patrimoniul istorico-lingvistic românesc se confruntă cu dificultăţi şi aspecte specifice: un număr mare de perioade în evoluţia limbii, un număr relativ mic şi foarte dispersat de resurse depozitate, o mare diversitate de alfabete folosite la tipărirea lor, în particular câteva „alfabete de tranziţie" chirilico-latine. Dificultăţile în digitizarea şi conservarea acestui tezaur ţin de recunoaşterea corectă a literelor chirilico-latine, dar şi de inexistenţa unui lexicon adecvat perioadei de tipărire a resursei. O soluţie pentru problema lexiconului ar fi alinierea la normele lingvistice contemporane ale textelor vechi [1].

Istoric, limba română a parcurs o cale lungă şi bogată de dezvoltare. Există studii care explică apariţia foneticii şi ortografierii caracteristice etapelor concrete de evoluţie a limbii, care sunt necesare atât pentru determinarea alfabetului, cât şi a literelor specifice [2,3]. Cunoaşterea acestor legităţi ne permite să construim resurse lingvistice utilizând un instrumentar special elaborat pentru o perioadă istorică concretă.

Prima carte tipărită pe teritoriul românesc a fost *Liturghierul slavon*, îngrijit de către ieromonahul Macarie în anul 1508, iar prima carte tipărită în limba română a fost *Catehismul Românesc* al diaconului Coresi, apărut la Braşov în anul 1535 [3].

Biblioteca Naţională a Republicii Moldova deţine o colecţie de aproximativ 21 000 cărţi vechi şi rare. Circa 20 de cărţi din această colecţie sunt tipărite în limba română, în Basarabia (Chişinău şi

Dubăsari), utilizând alfabetele chirilic şi tranziţional [4,5]. Bibliotecile publice din Sankt Petersburg deţin importante mostre de carte românească veche (secolele XVI-XIX). Dintre cele 66 de titluri incluse, spre exemplu, în *Catalogul ediţiilor chirilice ale slavilor de sud şi ale românilor*, 45 de volume revin slavilor de sud, iar 21 de volume – ţărilor româneşti [6].

Studiile existente explică aspectele legate de dezvoltarea componentelor principale ale limbii: alfabet, lexicon, ortografie cu referire la etapele specifice de evoluţie a limbii. Această informaţie este utilă pentru a crea resurse şi instrumente lingvistice racordate la anumite perioade din istoria limbii. Ţinând cont de particularităţile fiecărei perioade, vom propune o tehnologie pentru crearea acestor componente. În particular, vom studia problema de digitizare a textelor tipărite cu caractere chirilice în Republica Sovietică Socialistă Moldovenească (RSSM) în perioada 1967 -1989.

Lucrarea prezintă un proiect pe termen lung, care abia începe. Pe parcurs ne vom conduce de principiul „din prezent în adâncul secolelor”.

Perioade de evoluţie a limbii române

Istoria limbii române cunoaşte două epoci în dezvoltarea sa. Prima se referă la formarea dialectului dacoromân, începând cu căderea Sarmisegetuzei (106 A.D.) până în secolul al XV-lea [2]. Se utiliza alfabetul chirilic graţie influenţei masive a Bisericii Ortodoxe.

Epoca a doua de dezvoltare a limbii române literare (sec. XVI-XX) începe cu apariţia primelor texte scrise în limba română şi constituie rezultatul unei îndelungate şi complexe evoluţii [3]. Procesul de unificare lingvistică este marcat de apariţia Bibliei de la Bucureşti (1688), care a condus ulterior la stabilirea a două mari etape în evoluţia lingvistică [7].

Etapa întâi începe cu apariţia primelor texte literare româneşti şi se încheie la începutul secolului al XVIII-lea. În cadrul acestei etape pot fi distinse 3 perioade:

- Anii 1532 şi 1588, prima fază a limbii literare;
- Anii 1588-1656, faza consolidării principalelor variante ale limbii române literare (muntenească, moldovenească şi sud-vest-ardelenească);
- Anii 1656-1715, faza influenţei reciproce dintre variantele literare.

A doua etapă se întinde pe un interval între 1715 şi 1960. Este epoca de consolidare a limbii unice supradialectale. Procesul de unificare a limbii române literare a cunoscut o evoluţie lungă, în cursul a 4 perioade:

1. Anii 1715-1780, momentul primei unificări, aproximativ în 1750;
2. Anii 1780-1836, diversificarea lingvistică;
3. Anii 1836-1881, constituirea principalelor norme ale limbii literare de astăzi;
4. Anii 1881-1960, definitivarea formării normelor limbii române literare contemporane.

Ultima perioada ne descrie consolidarea stilurilor limbii române literare. În 1904, prin modificările aduse ortografiei, se stabilesc definitiv bazele scrierii fonetice, păstrate, cu unele retuşări ulterioare, până în prezent. Vom arăta în Fig.1-8 exemple de texte tipărite în diverse perioade de evoluţie a limbii.

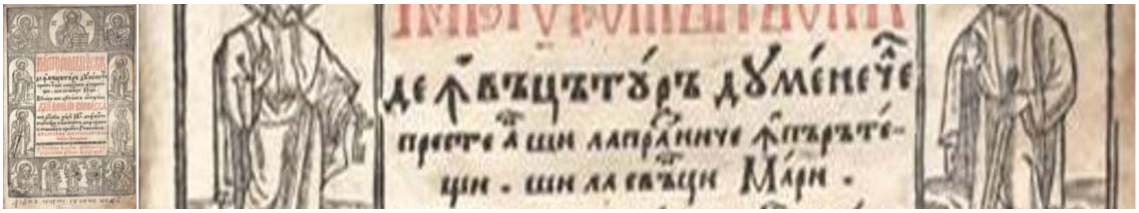


Figura 1. *Cazania* lui Varlaam, Iaşi, 1643
http://tiparituriromanesti.wordpress.com/2011/12/04/cazania-lui-varlaam-iasi-1643/

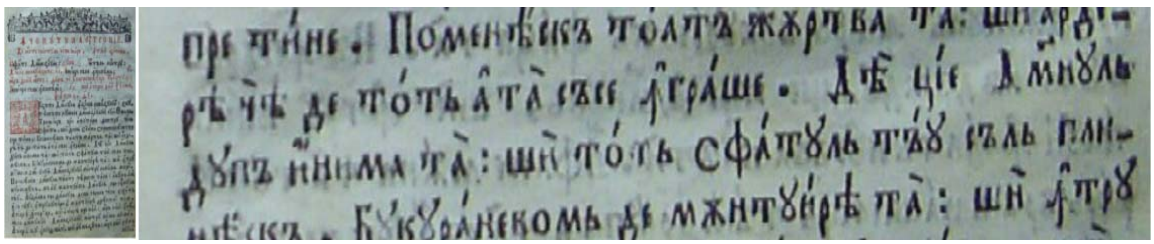


Figura 2. *Ceaslov*, 1748. *http://muzeu.reintregirea.ro/index.php?cid=carte-53*

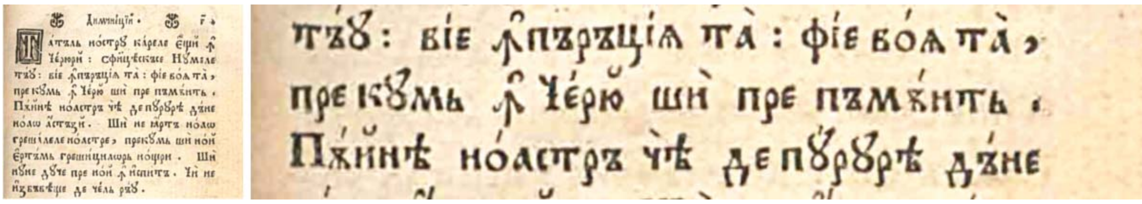


Figura 3. *Татъл ностру*. Acathist cu multe alease rugăciuni pentru evlaviia fieştecăruia creştin. Acum a treia oară tipărit. Blaj: Tipografiia Seminariului, 1786. http://documente.bcucluj.ro/web/bibdigit/patrimoniu/BCUCLUJ_FCS_BRV497.pdf

Figura 4. *Letopisiţile Ţării Moldovii publicate pentru întâiaşi dată de Mihail Kogălniceanu.* Tom I. Iaşii. La toate libreriile. 1852. *http://tiparituriromanesti.wordpress.com/2012/03/24/miron-costin-cartea-pentru-descalecatul-dintaiu-a-tarii-moldovii-si-neamul-moldovinesc/*

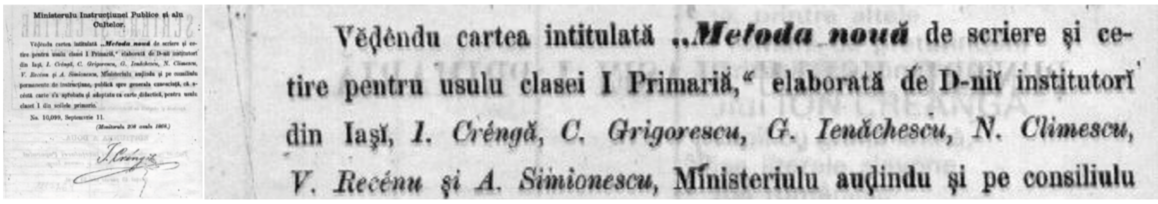


Figura 5. *Metodă nouă de scriere şi cetire: Pentru usulu clasei I*, Primariă / I. Créngă , C.Grigorescu, G. Ienăchescu, Ed. a II-a. – Iassy: Tipografia H. Goldner, 1868. – 71 p. http://www.scribd.com/doc/70357520/Carte-rară-in-limba-romană-din-colecţiile-bibliotecii-Contribuţii-bibliografice-Fascicula-2

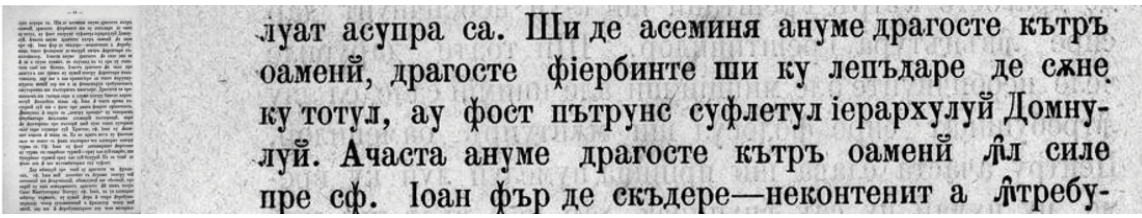


Figura 6. O pagină din revista „Луминъторюл”, 1908, Nr. 1. http://upload.wikimedia.org/wikipedia/ru/b/b1/Rumynskaja_Kirillica_Grazhdanskij_Shrift.jpeg

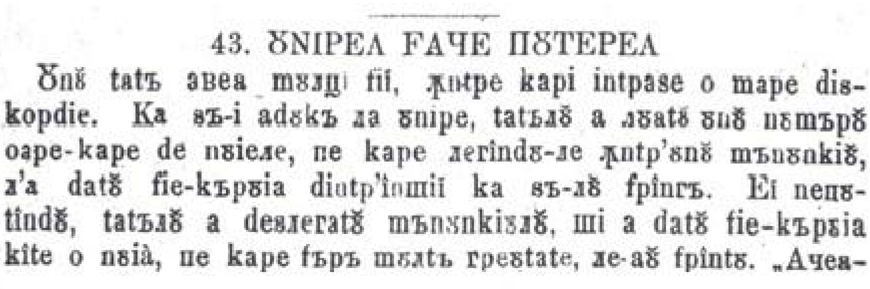


Figura 7. Una din variantele alfabetului tranziţional utilizat în alfabetul lui I. Creangă.

„Ш’ачел реӂе-ал поезией, вечник тынэр ши фериче,
Че дин фрунзе ыць дойнеште, че ку флуерул ыць зиче…”

Figura 8: Text tipărit cu alfabet chirilic, perioada 1967-1989. Utilizat în prezent în Transnistria (M. Eminescu, „Epigonii”)

Perioade de evoluție a alfabetului român

În secolul al XVII-lea, tiparul românesc utiliza un alfabet chirilic cu 47 de litere, majoritatea dintre ele fiind împrumutate din alfabetul bisericii slavone. S-au adăugat câteva litere greceşti, în mare parte pentru redarea adecvată a numelor proprii, dar şi litere originale româneşti. De exemplu, litera  utilizată pentru a reda prefixul (prepoziția) **în**, **îm**, sau litera **î** modernă la începutul cuvântului. Acest alfabet a fost utilizat la tipărirea *Cazaniei* lui Varlaam la Iaşi în anul 1643 (Fig.1). Primul abecedar românesc a fost tipărit în 1699 la Bălgrad (Alba Iulia), iar prima gramatică românească a fost tipărită în 1757 de Dimitrie Eustatievici.

Începând cu anul 1830 şi până la adoptarea oficială a alfabetului latin român în 1862, nu exista un alfabet stabil, astfel în această perioadă au fost utilizate cel puțin şapte modificări ale aşa-numitului „alfabet de tranziție", chirilico-latin, care conținea atât litere latine, cât şi litere chirilice (Fig. 4, 7). De exemplu, **е - є** (1830) - **ε** (1846); **к - k**; **щ - шт**; **s - дз - dz - ʒ** (1846).

Utilizarea grafiei latine în România nu a influențat activitatea tipografică din Basarabia. După alipirea Basarabiei la Imperiul Rus în 1812, limba oficială la Chişinău devine limba rusă. În anul 1833 limba română a fost exclusă din circuitul oficial, dar a continuat să fie utilizată în activitățile eparhiale. Astfel, pe parcursul anilor 1867-1871 apărea versiunea română a monitorului eparhiei Chişinău tipărit cu caractere chirilice. Tipografia bisericească din Chişinău a fost sistată pe perioada 1883-1890, procesul fiind reluat la începutul secolului XX.

Spre deosebire de alfabetul chirilic utilizat pentru scrierea limbii române din secolele XIV-XV până în anul 1862, alfabetul chirilic folosit în Republica Autonomă Sovietică Socialistă Moldovenească (RASSM) începând cu anii 1930 şi, ulterior, în Republica Sovietică Socialistă Moldovenească (RSSM) şi Transnistria în prezent, este de fapt o adaptare a alfabetului chirilic rusesc. De menționat că în perioada 1932-1938 în RASSM a fost utilizat alfabetul latin. În Republica Moldova alfabetul chirilic a fost utilizat până în 1989.

Vom prezenta mai jos (Tab.1) perioadele de evoluție a alfabetului român începând cu *Cazania* lui Varlaam. Pe lângă alfabet, există şi alți factori care caracterizează evoluția limbii, precum ortografierea şi lexiconul.

Recunoaşterea textelor tipărite

Procesul de digitizare şi de recunoaştere pentru manuscrise este destul de complicat, deoarece necesită efectuarea unor operații suplimentare, de exemplu, ajustarea contrastului, „curățirea imaginii', segmentarea textului. De asemenea, trebuie elaborați algoritmi speciali de recunoaştere şi lexicoane specializate. Procesul de digitizare şi recunoaştere e constituit din următoarele etape (Figura 9):

- Digitizarea textului pentru obținerea copiei electronice grafice;
- Recunoaşterea cu metode standardizate, adică utilizarea nemijlocită a OCR (Optical Character Recognition) [8], sau prin instruirea lui. În caz contrar, se vor folosi proceduri ale Inteligenței Artificiale, aşa-numitul proces de conversie. Transliterarea textului se va efectua ținând cont de literele specifice utilizate în textul inițial.
- Verificarea textului recunoscut se produce utilizând resursele lingvistice reutilizabile specializate pentru perioada de timp respectivă.

Digitizarea textelor constă în scanarea lor şi obținerea variantei electronice în formă de imagine. Pentru recunoaşterea textelor din imagine se aplică OCR. Sistemele standard OCR utilizează diferite metode de recunoaştere a textelor. Am cercetat posi-

Tabelul 1

Evoluția alfabetului român începând cu anul 1642

România	**Basarabia**
1642 – 1710 (alfabet chirilic)	
1710 – 1830 (alfabet chirilic modificat)	1710 – 1814 (alfabet chirilic modificat)
1830 – 1862 (alfabet tranzițional, mixt chirilico-latin)	1814 – 1880 (alfabet chirilic bazat pe alfabetul rus şi cel slavonic bisericesc, ocazional alfabet tranzițional şi latin)
1862 – 1904 (alfabet latin)	1880 – 1905 (n-a existat tipar românesc) 1905 – 1918 (alfabet chirilic bazat pe alfabetul civil rus)
1904 – 1960 (alfabet latin modificat)	1919 – 1940, 1941 – 1944 (alfabet latin modificat) 1940 – 1941 (alfabet chirilic bazat pe alfabetul rus) [Vezi mai sus în text situația din Transnistria]
1960 – 1993 (alfabet latin modificat)	1944 – 1989 (alfabet chirilic bazat pe alfabetul rus; din 1967 apare litera ӂ)
1993 – prezent (alfabet modern român bazat pe alfabetul latin)	1989 – prezent (alfabet modern român bazat pe alfabetul latin) [Vezi mai sus în text situația din Transnistria]

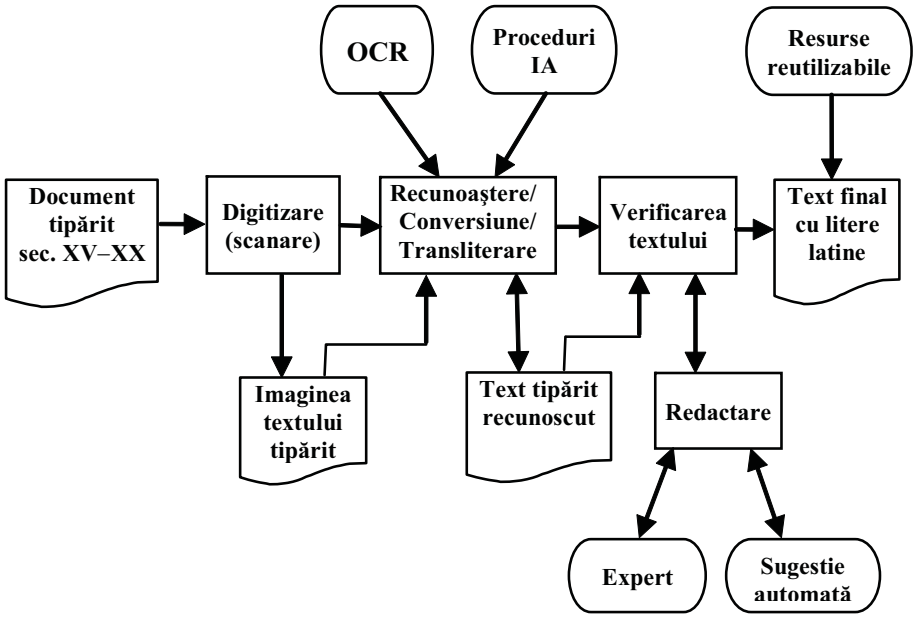


Figura 9. Etapele tehnologice de recunoaştere a textelor tipărite

bilităţile a două sisteme: IRIS şi ABBY FineReader. Rezultatele experienţelor de recunoaştere a textului tipărit în sec. al XIX-lea sunt expuse în continuare. S-a determinat că sistemul IRIS, în procesul de instruire, nu poate selecta orice fragment din imaginea textului şi de aceea acest sistem nu satisface scopurile noastre de recunoaştere a textului tipărit cu scrisul vechi român.

În continuare se vor folosi tehnici de recunoaştere a formelor pentru identificarea individuală a caracterelor unei pagini de text, inclusiv semnele de punctuaţie, spaţiile şi sfârşitul de linie. Textul recunoscut se va prezenta ca un fişier editabil.

Transliterarea este un proces strict individual ce depinde de perioada examinată. În funcţie de textul iniţial, se vor utiliza programe care conţin informaţie despre caracterele specifice întâlnite în text. Transliterarea presupune stabilirea unei relaţii bidirecţionale univoce între două sisteme de scriere astfel, încât un cunoscător să poată reconstitui textul original din varianta transliterată. Procesul de transliterare se va folosi doar la necesitate.

Verificarea textului se efectuează cu aplicaţii special elaborate [9], care utilizează resursele reutilizabile specifice pentru perioada istorică a textului tipărit. Totodată, cuvintele noi obţinute se vor introduce în lexiconul corespunzător.

Procesarea textelor tipărite cu alfabet chirilic în RASSM şi RSSM

Perioada iniţială de utilizare a alfabetului chirilic în Republica Autonomă Sovietică Socialistă Moldovenească (RASSM) se referă la anii 1924-1940 şi se asociază cu utilizarea unui lexicon foarte specific, caracterizat de:

- utilizarea cuvintelor ruseşti (de exemplu, **совет**, **указ**, **словарь)** în locul echivalentelor româneşti (**consiliu**, **decret**, **dicţionar**);
- excluderea neologismelor româneşti, ele fiind considerate „burghezisme”;
- utilizarea lexiconului local (Transnistrean);
- introducerea unor neologisme auto-inventate pentru unele noţiuni abstracte neatestate în limbajul basarabean. De exemplu, **амувремник** (amuvremnic) în loc de **contemporan**;
- utilizarea particularităţilor accentului local (Transnistrean), de exemplu, **ди (di)** în loc de **de**, **мержи** (**merji**) în loc de **merge**, **сунити** (**suniti**) în loc de **sunete** etc.

Ne vom referi în continuare la perioada 1967-1989 de utilizare a alfabetului chirilic în Republica Sovietică Socialistă Moldovenească (RSSM). Pentru procesarea OCR a textelor apărute în această perioadă, este necesar să fie instruit sistemul OCR pentru a recunoaşte litera adiţională **ж** şi pentru a crea lexiconul respectiv. Existenţa unui lexicon caracteristic acestei perioade ar permite automatizarea procesului de verificare şi validare a cuvintelor recunoscute prin procedeele expuse mai sus. Acest lexicon poate fi creat: a) manual, b) prin transliterarea cuvintelor româneşti scrise cu caractere latine în varianta corectă scrisă cu caractere chirilice şi c) prin alinierea variantelor de text tipărite în paralel cu caractere latine şi caractere chirilice.

Prin transliterare vom înţelege transcrierea unui cuvânt din limba română în forma echivalentă scrisă

cu caractere chirilice conform normelor lingvistice acceptate pe perioada 1967-1989 în RSSM. Metoda transliterării s-ar potrivi ideal în cazul dacă se reuşeşte formalizarea tuturor regulilor de transcriere. Un studiu prealabil arată că acest proces este anevoios şi nu poate fi automatizat în totalitate din cauza iregularităţilor legate de discordanţa dintre fonologia, morfologia şi sintaxa limbii române şi normele lingvistice acceptate în RSSM. Parţial acest proces poate fi automatizat implicând elemente de formalizare a regulilor de transcriere, de intervenţie manuală şi de aliniere.

Dificultăţi evidente apar la transliterarea cuvintelor de proveniență străină. Dacă în limba română aceste cuvinte se scriu, de regulă, ca şi în limba originală, atunci transcrierea lor cu caractere chirilice se face conform pronunţării. De exemplu, ***design – дизайн, cowboy – ковбой, watt – ват, charleston – чарлстон***. Aceste cuvinte pot fi transliterate doar în regim manual.

Pentru lexiconul original românesc procesul respectiv poate fi parţial automatizat. În acest scop s-au stabilit reguli de transcriere a literelor şi îmbinărilor de litere. Inserăm mai jos câteva astfel de reguli.

1) Reguli de transcriere „literă → literă”. De exemplu, ***a →a, ă→ э, b →б, d→ д, f→ ф, l→л, m→ м, n→ н, r→ р, ş→ ш, t→ т, ţ → ц, v→ в, z→з (bardă – бардэ, zarvă – зарвэ, măr→мэр).***

2) Reguli de transcriere pentru literele ***î*** şi ***â***. Îmbinările ***âi***, sau ***îi*** se vor transcrie în ***ы*** pentru cuvintele ***mâine, pâine, câine*** şi derivatele lor (***mâine→мыне, pâine→пыне, câine→кыне, mîine→мыне, pîine→пыне, cîine→кыне***). În alte situaţii se va aplica regula „literă → literă”: ***â →ы, î →ы*** (***român→ромын, întâi→ынтый***).

3) Reguli pentru **ea** şi **ia.** Se transcriu în **я**, cu o singură excepţie: pronumele **ea** se transcrie **ea**; în acelaşi timp, verbul **ia** se transcrie ca **я**;

4) Transcrierea lui **i** prin trei litere diferite: **и**, **й**, **ь**.Concomitent, menţionăm existenţa cazurilor când litera **i** este omisă ***(iepure→enype),*** sau trecută în ***ы (introducere→ынтродучере).*** Reguli de transcriere pentru litera ***c***.

a) c→к, dacă după ***c*** urmează una din vocalele ***a, â, î, o, u***, sau o consoană diferită de ***h*** (***încreţit→ынкрецит, clocot→клокот, casă→касэ, cucoş→кукош; câmp→кымп***).

b) Combinaţiile ***che***, ***chi*** se vor transcrie în ***ке*** şi, respectiv, ***ки*** (***cheltuială→келтуялэ, chihlimbar→кихлимбар, chibzui→кибзуи***).

c) Dacă după îmbinarea ***ce*** nu urmează ***a***, atunci se aplică regula ***ce→че*** (***cercel →черчел, cep→чеп***).

d) cea → ча (***ceară→чарэ, ceas→час, ceaţă→чацэ, ceaşcă→чашкэ***). Excepţie pentru articolul demonstrativ ***cea (acea) →чя (ачя).***

e) Dacă după îmbinarea ***ci*** nu urmează una din vocalele ***a***, ***o, u***, atunci se aplică regula ***ci→чи*** (***ciment→ чимент, ciclu→чиклу, cimbrişor→ чимбришор***). Dacă cuvântul se termină în ***ci***, atunci poate fi aplicată una din regulile: ca excepţie ***ci→ч*** (***arici→арич, beci→беч, prichici→прикич***); ***ci→чь***, pentru plural (arici***→аричь***, saci***→сачь,*** maci***→мачь***); ***ci→чи***, alte situaţii (***aci→ачи, răci→рэчи, înveşnici→ынвешничи***).

f) cio→чо (***ciorbă→чорбэ, ciocârlie→ чокырлие, cioban→чобан***, ***cocioabă →кочоабэ***).

g) ciu→чу (***ciuperci→чуперчь, ciubotă→ чуботэ, bucium→бучум***).

Utilizând astfel de reguli (lista cărora poate fi prelungită), procesul de transliterare se transformă într-o acţiune de trecere prin „ciur şi prin dârmon”. Pornind de la lexiconul contemporan al limbii române [11,12] se stabileşte un set de filtre, fiecare filtru având un coeficient de prioritate, care depinde de probabilitatea obţinerii unui rezultat corect la aplicarea regulilor acestui filtru. Mai întâi, se vor aplica acele filtre care exclud, sau minimizează, intervenţia manuală. Cuvintele filtrate se exclud din lexicon şi asupra lexiconului rămas se aplica alte filtre. Din păcate, toate aceste etape de filtrare necesită un anumit grad de intervenţie manuală.

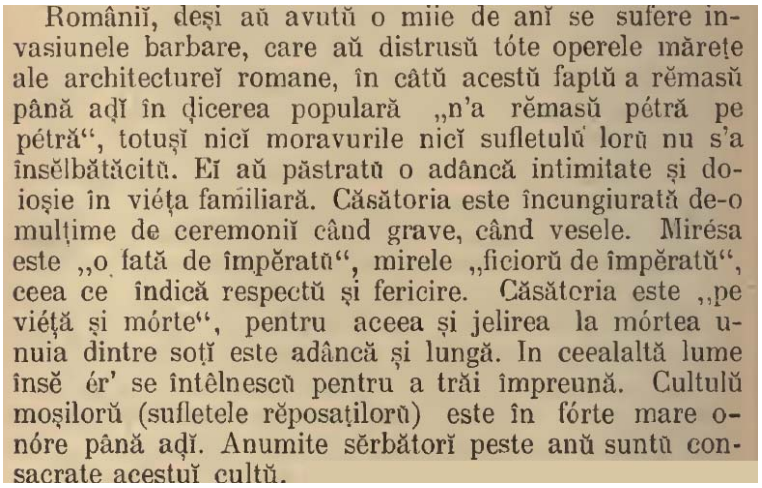
Procesarea textelor tipărite cu alfabet latin şi litere adiţionale

Pentru ilustrarea tehnologiei descrise vom cerceta procesul de recunoaştere şi verificare a unui text digitizat din cartea [10], tipărită în anul 1894, (Fig.10).

Textul din Fig.10 a fost recunoscut cu sistemul OCR IRIS. Ca urmare au rămas nerecunoscute cuvintele ortografiate cu litere specifice secolului al XIX-lea. De exemplu, se obţine **tnsălbătăcitu** în loc de **însélbătăcitŭ**.

Acest rezultat nu poate fi îmbunătăţit, deoarece IRIS nu posedă capacitatea de a selecta fragmente arbitrare din imagine. Utilizarea unui lexicon modern permite să se recunoască **avutŭ** ca **avută, v**arianta corectă pentru acest context fiind **avut.** Cuvintele specifice lexiconului secolului al XIX-lea nu pot fi recunoscute corect, deoarece pentru aceasta sunt necesare dicţionare corespunzătoare perioadei date care, în cazul nostru, ar conţine cuvintele **remasŭ**, **viéţa**, **împêratŭ** etc.

Dacă în textul recunoscut se vor restabili literele specifice şi textul obţinut se va verifica cu ajutorul



Românii, deşi aŭ avutŭ o miie de anĭ se sufere invasiunele barbare, care aŭ distrusŭ tóte operele mărețe ale architectureĭ romane, în câtŭ acestŭ faptŭ a rĕmasŭ până adi în ḑicerea populară „n'a rĕmasŭ pétră pe pétră", totuşĭ nicĭ moravurile nicĭ sufletulŭ lorŭ nu s'a însĕlbătăcitŭ. Eĭ aŭ păstratŭ o adâncă intimitate şi doioşie în viéța familiară. Căsătoria este încungiurată de-o mulțime de ceremoniĭ când grave, când vesele. Mirésa este „o fată de împĕratŭ", mirele „ficiorŭ de împĕratŭ", ceea ce indică respectŭ şi fericire. Căsătoria este „pe viéță şi mórte", pentru aceea şi jelirea la mórtea unuia dintre soțĭ este adâncă şi lungă. In ceealaltă lume însĕ ér' se întâlnescŭ pentru a trăi împreună. Cultulŭ moşilorŭ (sufletele rĕposaților) este în fórte mare onóre până adi. Anumite sĕrbătorĭ peste anŭ suntŭ consacrate acestuĭ cultŭ.

Figura 10. Text digitizat, 1894 (Densuşianu, 1984, p. 130)

corectorului ortografic RomSp [9], care posedă un lexicon al limbii române moderne de circa un milion de cuvinte, vom constata că 57 la sută din cuvintele textului sunt recunoscute drept corecte. Acestea sunt cuvintele, ortografierea cărora a rămas intactă față de perioada secolului al XIX-lea, de exemplu, **sufere**, **acesta**, **fericire**. Cuvinte „suspicioase" sunt cele afectate de modificările ortografice, de exemplu, **ceealaltă** (***cealaltă***), **doioşie** (***duioşie***), **miie** (***mie***), **avutŭ** (***avut***), **adi** (***azi***).

Pentru recunoaşterea corectă a textului trebuie de instruit sistemul OCR ca să recunoască literele şi să completeze lexiconul cu cuvinte noi, specifice secolului al XIX-lea. De exemplu: **avutŭ, o miie, invasiunele, pétră, nicĭ, sufletulŭ, lorŭ, însălbătăcitŭ, doioşie, viéța, ficiorŭ, împăratŭ, miresa.**

Ținând cont de faptul că sistemul OCR ABBYY Fine Reader este înzestrat cu facilități de instruire, am mai efectuat un experiment. Sistemul a fost instruit în mod special ca să poată recunoaşte literele specifice secolului al XIX-lea. Iată câteva astfel de litere:

- **ŭ** (literă finală, mută sau citită),
- **é** (é se pronunță ca diftongul ea),
- **ó** (ó se pronunță ca diftongul oa),
- **ḑ** (se citea z sau dz),
- **ê (**se folosea ca â**).**

Unele rezultate ale experimentelor sunt relatate în Tabelul 2.

Pentru a obține rezultate mai performante la verificarea textelor tipărite este necesar ca pentru perioada istorică corespunzătoare:

- să fie instruit scanerul pentru a recunoaşte caracterele specifice;
- să fie elaborat un lexicon cu cuvinte şi fraze uzuale specifice perioadei;
- să fie extinse facilitățile corectorului ortografic (spellchecker) pentru a utiliza alfabetul şi lexiconul elaborat.

Procesarea textelor tipărite cu alfabete tranziționale

Există cel puțin şapte versiuni ale alfabetului tranzițional (mixt chirilico-latin). Majoritatea literelor acestor alfabete pot fi recunoscute de ABBYY Fine Reader prin evidențierea codurilor respective din setul Unicode. O singură literă specifică pentru aceste alfabete lipseşte în Unicode - Ѫ . În acest caz urmează să fie inclusă o variantă de literă echivalentă (de exemplu, o săgeată ↑, sau slavonica „*yus*" Ѧ) şi instruit sistemul pentru recunoaşterea acestei variante grafice.

Tabelul 2

Rezultatele experimentelor OCR cu texte din secolul al XIX-lea

Mod de recunoaştere	Cuvinte corecte	Cuvinte suspecte
IRIS	57%	43%
ABBYY FR, fără instruire	63%	37%
ABBYY FR, cu instruire şi dicționar pentru o pagină	98%	2%
ABBYY FR, cu instruire, mai multe pagini, aceeaşi carte	95%	5%
ABBYY FR, cu instruire, pagini din altă carte	95.4%	4.6%

Concluzii

Resursele digitizate sunt înregistrări specifice stocate într-o bază de date postată pe Internet. Tehnologia propusă se axează pe soluționarea cu succes, pentru fiecare perioadă din evoluția limbii, a două probleme majore: 1. Elaborarea (dezvoltarea) algoritmilor pentru recunoaşterea literelor specifice perioadei; 2. Elaborarea instrumentarului şi interfe-

ţelor necesare pentru crearea resurselor lingvistice (lexiconului) corespunzătoare perioadei în scopul eficientizării procesului de recunoaştere a cuvintelor şi de aliniere la normele lingvistice contemporane.

La trecerea de la o perioadă la alta, în limitele posibilităţilor, se vor utiliza instrumentarul şi resursele deja elaborate, materializând astfel principiul „din prezent în adâncul secolelor".

Resursele electronice create pot fi plasate în Internet pentru acces public, contribuind la dezvoltarea mediului de comunicare informaţională pentru limba română. În plus, aceste resurse ar constitui un suport esenţial pentru cercetători, iar convertite în text literar contemporan ar putea fi utilizate ca materiale didactice în procesul de instruire.

Bibliografie

1. M. Moruz, A. Iftene, A. Moruz, D. Cristea, *Semi-automatic alignment of old Romanian words using lexicons*, In: Proceedings of the 8-th International Conference „Linguistic resources and tools for processing of the Romanian language", Iaşi, Editura Universităţii „A.I. Cuza", 2012, p. 119-125.

2. G. Ivănescu, *Istoria limbii romane*, Iaşi, 1980. [G. Ivănescu, History of the Romanian language, Iaşi, 1980.

3. Ştefan Munteanu şi Vasile Ţâra, *Istoria limbii române literare*, Editura Didactică şi Pedagogică, Bucureşti, 1978.

4. *Cartea Moldovei (sec XVII – înc. sec XX). Ediţii cu caractere chirilice (sec XVII – înc. sec XX)*,Catalog general, Chişinău, 1992.

5. Zamfira Mihail, *155 cărţi într-o carte*, Editura Prometeu, Chişinău, 2010, 532 p.

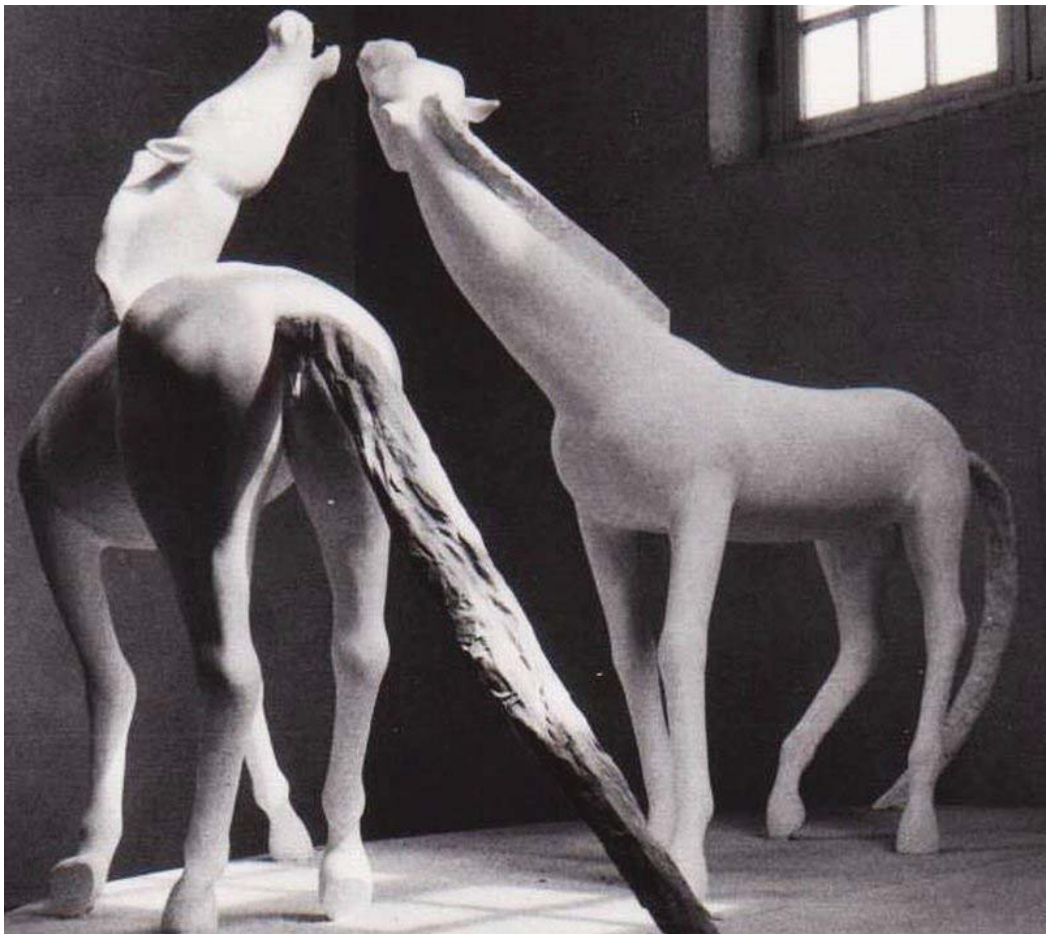
6. Valori Bibliofile-2008, Rev. Gazeta bibliotecarului, Iunie-Iulie 2008, nr. 6-7, p.1 http://87.248.191.115/bnrm/publicatii/files/3/93.pdf

7. Gheţie I., *Istoria limbii române literare*, Bucureşti, 1978.

8. Optical Character Recognition (OCR) Technology.

9. Burlaca O., Ciubotaru C., Cojocaru S., Colesnicov A., Magariu G., Malahov L., Petic M., Verlan T., *Applications based on reusable linguistic resources. In Multilinguality and interoperability in language processing with emphasis on Romanian*, Editors: D. Tufiş, C. Forăscu, Bucureşti, 2010, p.461-476.

10. Densuşianu, A., *Istoria limbii şi literaturii române*, Iaşi, 1894, http://ru.scribd.com/doc/123035210/Istoria-limbii-si-literaturii-romane.



Idel Ianchelevici. *Perennis perdurat poeta*, 1972, bronză