INSTRUMENTAR PENTRU DIGITIZAREA ŞI TRANSLITERAREA TEXTELOR TIPĂRITE ÎN LIMBA ROMÂNĂ CU CARACTERE CHIRILICE

Svetlana COJOCARU, Constantin CIUBOTARU,
Alexandru COLESNICOV, Ludmila MALAHOV,
Tudor BUMBU

Abstract: The paper discusses some of the problems regarding the digitization of Romanian Cyrillic printings of the 17th – 20th centuries. An application was created to accompany ABBYY Fine Reader OCR engine and facilitate the process of digitization. The proposed tools and solutions involved were already successfully used at the re-edition of a 20th century book in mathematics in the modern Latin Romanian script, and at research in Romanian philology.

Keywords: cultural heritage, digitization, OCR, Romanian language, Romanian Cyrillic script, 17th – 20th centuries.

Introducere

Problema digitizării şi conservării patrimoniului istorico-lingvistic reprezintă un domeniu prioritar din agenda digitală pentru Europa. Dezideratele principale ale politicii culturale pentru spaţiile unde se vorbeşte limba română ţin de studierea, valorificarea şi digitizarea acestui patrimoniu. Procesul de digitizare 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 volum mic de resurse depozitate foarte dispersat, o mare diversitate de alfabete folosite la tipărirea lor, în particular câteva

„alfabete de tranziţie” chirilico-latine, lipsa instrumentarului pentru recunoaşterea corectă a literelor chirilice din diferite perioade istorice, precum şi inexistenţa unui lexicon adecvat perioadei de tipărire a resursei.

Instrumentarul descris mai jos integrează o serie de componente software, atât existente, cât şi dezvoltate de autori, care formează o platformă pentru procesarea preliminară a imaginii, recunoaşterea textului şi transliterarea lui în grafie latină modernă. Este prezentată evoluţia alfabetelor utilizate pentru tipăriturile de limbă română în România şi pe teritoriul actual al Republicii Moldova, descrise componentele instrumentarului şi aplicarea lui pentru recunoaşterea şi transliterarea textelor din diferite perioade istorice.

Evoluţia alfabetelor româneşti

Utilizarea scrisului chirilic pentru tipăriturile româneşti a cunoscut o evoluţie atât în timp, cât şi în spaţiul geografic. Pe teritoriul României aceasta înregistrează variaţii de alfabete (sau elemente ale alfabetelor) chirilice vechi (alfabete chirilice româneşti – ACR) de la primele texte tipărite până în 1830, când începe perioada de tranziţie, caracterizată prin utilizarea alfabetelor mixte (alfabete de tranziţie – AT). Sunt înregistrate 17 astfel de variante, perioada respectivă finalizându-se în 1862 prin trecerea completă şi definitivă la grafia latină (alfabet latin pentru limba română – ALR).

Pe teritoriul actual al Republicii Moldova situaţia a fost diferită. În 1924, odată cu formarea Republicii Autonome Sovietice Socialiste Moldoveneşti (în componenţa Republicii Sovietice Socialiste Ucrainene) a fost impusă grafia chirilică, care prelua, de fapt, alfabetul limbii ruse cu excluderea a trei litere (ё, щ şi ъ). În anii 1932-1938 s-a revenit la grafia latină, care a fost din nou înlocuită cu cea chirilică (alfabetul chirilic moldovenesc), ultima fiind stabilită şi pe întregul teritoriu al Republicii Sovietice Socialiste Moldoveneşti (RSSM), formate în anul 1940. Alfabetul din 30 de litere ruseşti a fost completat în 1967 prin introducerea literei ӝ pentru redarea sunetului ʤ. Grafia chirilică fost înlocuită cu cea latină prin votarea de către Sovietul Suprem al RSSM pe 31 august 1989 a celor mai importante compartimente din Legea cu privire la funcţionarea limbilor vorbite pe teritoriul RSS Moldoveneşti, această zi fiind declarată ulterior sărbătoare naţională – Ziua limbii. În Tab. 1 este prezentată o schemă comparativă a etapelor principale ale evoluţiei alfabetelor pe teritoriul României şi Basarabiei.

Tab. 1. Etapele principale ale evoluţiei alfabetelor utilizate în tipar pe teritoriul României şi Basarabiei

România	**Basarabia**
1642 – 1797 (alfabet chirilic, până la 47 litere)	
1797 – 1830 (alfabet chirilic, 43 litere)	1710 – 1814 (alfabet chirilic, 43 litere)
1830 – 1862 (alfabete de tranziţie, mixt chirilic-latin)	1814 – 1880 (alfabet chirilic bazat pe alfabetul rus şi cel slavon bisericesc; ocazional alfabete de tranziţie)
1862 – 1904 (alfabet român bazat pe alfabetul latin, versiunea întâia)	1880 – 1905 (nu a existat tipar românesc) 1905 – 1918 (alfabet chirilic bazat pe alfabetul civil rus)
1904 – prezent (alfabet modern român bazat pe alfabetul latin)	1919 – 1940, 1941 – 1944 (alfabet modern român bazat pe alfabetul latin) 1940 – 1941 (alfabet chirilic bazat pe alfabetul rus)
	1944 – 1989 (alfabet chirilic bazat pe alfabetul rus; în 1967 apare litera ѫ)
	1989 – prezent (alfabet modern român bazat pe alfabetul latin)

Astfel, problema digitizării şi transliterării textelor româneşti scrise cu caractere chirilice poate fi divizată în trei compartimente majore: alfabet chirilic românesc vechi, alfabete mixte şi alfabet chirilic contemporan moldovenesc.

O particularitate importantă a alfabetelor menţionate o constituie existenţa aplicaţiei univoce într-o direcţie, anume: ACR→AT→ACM→ALR. Aceasta permite aducerea tuturor cuvintelor la alfabetul modern, folosindu-l drept o reprezentare generală a literelor din diferite epoci, precum şi utilizarea mijloacelor existente de procesare a limbajului natural şi (într-o anumită măsură) a resurselor lingvistice moderne.

Etapele principale ale procesării textelor

Principalele componente ale procesului de digitizare şi transliterare sunt următoarele:

- Recunoaşterea optică a caracterelor din tipărituri româneşti din sec. XVII-XX;
- Transliterarea rezultatelor obţinute în grafia modernă latină;
- Crearea modelelor (pattern-urilor) pentru reprezentarea caracterelor şi îmbinărilor de caractere;
- Crearea alfabetelor şi dicţionarelor specifice pentru anumite perioade, spaţii geografice, tipografii;
- Transliterarea inversă (din grafie latină în chirilică).

În linii generale abordarea noastră este ilustrată în Fig. 1.

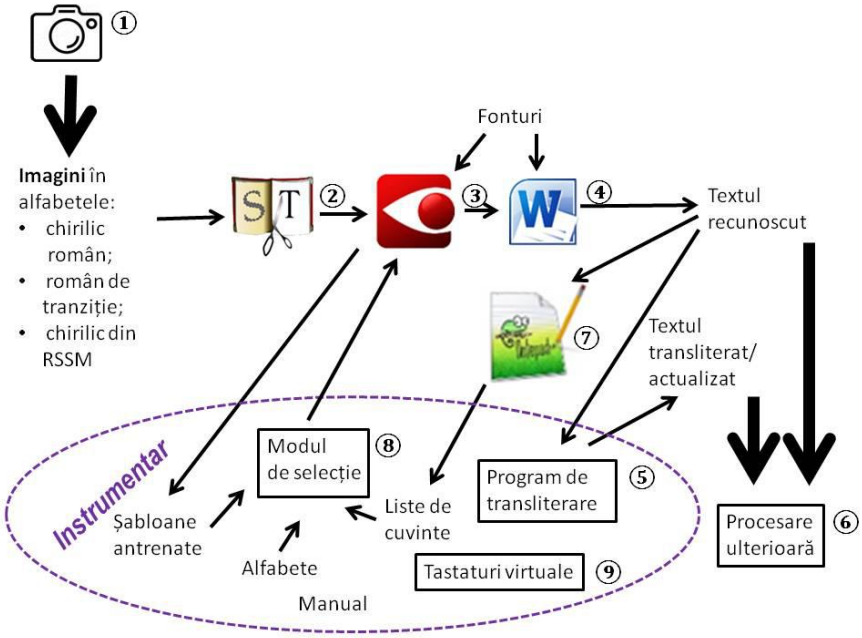

Fig. 1. Schema digitizării şi transliterării textelor româneşti tipărite cu caractere chirilice

Operaţiunile principale (notate cu cifrele 1-6) includ următoarele procedee:

1. Obţinerea imaginii prin scanare, cu utilizarea softului din dotarea scanerului. Calitatea dorită este de 600 dpi sau mai mult.
2. Pregătirea imaginilor pentru OCR. În cazul nostru am utilizat ScanTailor – un soft utilitar gratuit, care efectuează corecţii automate masive ale defectelor imaginii, spre exemplu, corectează unghiul de înclinare a paginii sau curăţă unele pete mici. Desigur, există mai multe instrumente de acest fel. ABBYY Finereader (AFR), pe care noi îl aplicăm la pasul următor, conţine şi el un propriu editor de imagini, chiar mai performant, capabil să alinieze linii ondulate sau să corecteze distorsiuni trapezoidale. Însă aceste corecţii ar trebui aplicate manual pentru fiecare imagine (pagină) în parte, ceea ce este destul de laborios la procesarea volumelor mari de text.
3. Recunoaşterea optică a caracterelor (OCR) este efectuată cu ABBYY Finereader (AFR). Acest program operează cu texte din circa 140 de limbi, ultimele versiuni incluzând şi caractere din alfabetul chirilic vechi. Funcţionează pe reţele neurale, graţie modelării intrinseci a limbajului dă dovadă de performanţă şi acurateţe.
4. Textul recunoscut este salvat ca un document Microsoft Word, care ne oferă o colecţie bogată de fonturi pentru a reda caracterele cât mai aproape de cele originale, precum şi vaste posibilităţi de formatare şi editare.
5. În funcţie de scopurile urmărite de utilizator, textul obţinut în una din cele trei clase de alfabete chirilice, descrise mai sus, poate fi transliterat în grafie latină modernă. Pentru alfabetul chirilic românesc vechi, precum şi pentru alfabetele de tranziţie există o corespondenţă destul de constantă în raport cu alfabetul latin românesc. În cazul alfabetului chirilic moldovenesc ne confruntăm cu mai multe iregularităţi cauzate de anumite litere, dar şi de scrierea cuvintelor de origine străină şi a substantivelor proprii. Modul de soluţionare a acestor probleme va fi expus în secţiunile ce urmează.
6. Procesarea manuală sau automatizată a rezultatului obţinut pentru corectarea finală a textului.

Modulele adiţionale (7-9) oferă nişte servicii pentru facilitarea operării şi pentru îmbunătăţirea rezultatului. AFR utilizează liste de cuvinte pentru soluţionarea ambiguităţilor şi eliminarea cratimelor. Aceste liste trebuie să

conţină câte un cuvânt per linie utilizând codurile UTF-8. În cadrul abordării noastre folosim, de regulă, editorul Notepad++, care se distribuie gratuit şi operează atât cu UTF-8, cât şi cu alte codificări. Utilizând pluginul TextFX, aplicaţia Notepad++ poate crea liste de cuvinte sortate unice, adică fără repetări, adică anume în formatul solicitat de Finereader. De asemenea, Notepad++ permite selecţia fonturilor necesare pentru editarea cuvintelor româneşti în caractere chirilice, inclusiv din alfabetele vechi.

Instrumentarul conţine un modul de selectare a grupului de date pregătite a priori pentru AFR (alfabet, listă de cuvinte, set de şabloane pentru recunoaştere). Selecţia se efectuează în funcţie de perioada de timp, regiune geografică şi tipografie. Acest modul apelează programul AFR incluzând datele selectate. Alfabetele şi listele de cuvinte sunt pregătite manual, patternurile pentru antrenare sunt stocate de către AFR. Pentru secolul XVII au fost incluse circa 3500 de patternuri şi o listă de peste 2600 de cuvinte, pentru secolul XVIII – circa 1800 cuvinte şi peste 4000 de patternuri. Fig. 2 ilustrează şabloanele introduse în AFR pentru antrenarea recunoaşterii literei „a” în tipărituri din secolul XVII. Pentru culegerea de pe ecran a caracterelor româneşti vechi în scopul introducerii sau redactării textelor respective este creată o tastatură virtuală.

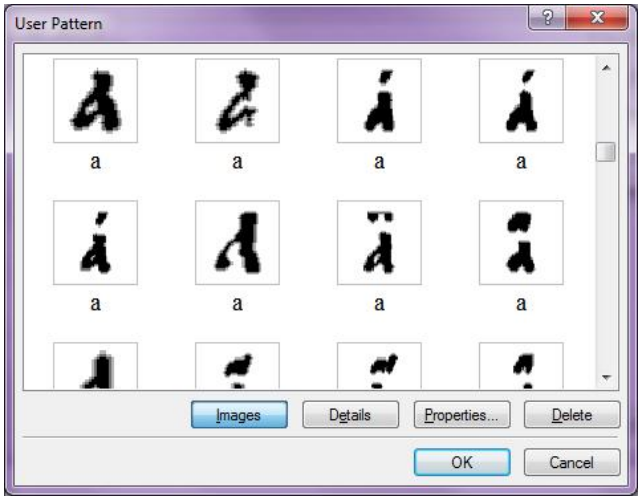

Fig. 2. Şabloane AFR pentru recunoaşterea literei „a”

Evident că evoluţia unei limbi nu se reduce doar la modificarea scrisului, ci se exprimă, în primul rând, în dezvoltarea lexiconului şi a ortografiei. Aceste subiecte nu sunt tratate în studiul nostru, exceptând posibilitatea de adaptare la ortografierea contemporană a unor cuvinte transliterate din alfabetul chirilic moldovenesc, efectuată în cazul unei solicitări venite de la utilizator.

Procesarea textelor din secolul XVII

Fiecare perioadă din cele trei enumerate mai sus îşi are specificul său de procesare. Pentru textele din secolul XVII este caracteristic un nivel mai avansat de degradare, zgomote, caractere alipite sau rupte. În recunoaşterea lor ne confruntăm cu o serie de probleme, dintre care vom evidenţia următoarele:

1. Suprascrierea literelor (slovelor), având drept scop economia de spaţiu şi de eforturi la scriere / citire;
2. Omiterea unor litere;
3. Multitudinea diverselor semne (title, diacritice etc.) plasate deasupra liniei;
4. Prescurtări (în denumiri de funcţii, luni ale anului etc.);
5. Scrierea numeralelor atât cu cifre, cât şi cu litere;
6. Diversitatea fonturilor.

Suprascrierea şi prezenţa semnelor deasupra rândului conduc spre separarea eronată a unei linii în două. Pentru soluţionarea acestei probleme, la sugestia autorilor Finereader-ului, s-a recurs la majorarea formală a densităţii imaginii.

În urma procesării cu instrumentarul nostru literele suprascrise sunt incluse în cuvânt, iar celelalte semne supralinie sunt omise.

Tipografiile din secolul XVII foloseau o varietate de fonturi pentru tipărirea cărţilor şi a documentelor, dar dintre acestea putem distinge, în general, două seturi substanţial diferite, atât după stil, cât şi după utilizarea caracterelor. În Fig. 3 sunt prezentate fragmente a două pagini din două cărţi tipărite în perioade destul de apropiate (1648 şi 1679), care au fonturi distincte.

Fig. 3. Două pagini tipărite cu utilizarea diferitelor fonturi;
caracterele atipice sunt evidenţiate sub text

Este evident, că aceste două texte sunt diferite după stil, având fonturi accentuat distincte în cazul literelor “t” şi “z”. Litera “t” din primul text este tipărită ca “**т**” standard, iar în cel de al doilea apare în forma scrierii de mână, adică “***m***”. Acelaşi lucru îl observăm şi în cazul literei “z”, al cărei mod de scriere diferă substanţial. Dacă am aplica procedura de recunoaştere a unui text de al doilea tip utilizând modele din primul, rezultatul ar avea o rată de eroare destul de mare. Deci este necesar să avem un instrument cu ajutorul căruia utilizatorul va putea alege cel mai potrivit set de modele pentru cartea sau documentul din secolul XVII tipărit în grafie chirilică românească.

Utilizatorul trebuie să fie familiarizat cu modelele existente, el va analiza vizual o pagină din cartea care urmează să fie recunoscută şi va alege modelul cel mai potrivit, orientându-se după anumite caractere distincte (precum ar fi scrierea diferită a literelor “**т**” şi “**з**” în exemplele de mai sus). În caz că modelul cel mai potrivit nu poate fi ales vizual, se recomandă recunoaşterea unei pagini prin câteva modele diferite, iar cel care va prezenta un rezultat mai bun va fi utilizat pentru recunoaşterea întregii lucrări.

O interfaţă specială este creată pentru selecţia regiunii geografice, unde s-a tipărit textul. Putem alege una din următoarele variante: Iaşi, Bucureşti, Târgovişte, Bălgrad (Alba Iulia), Uniev (Cernăuţi), Sas Sebeş, Snagov sau Buzău. În cadrul unei regiuni avem posibilitatea selectării tipografiei, spre exemplu, pentru Bucureşti sistemul este antrenat în recunoaşterea fonturilor din Tipografia Domnească şi cea a Scaunului Mitropoliei Bucureştilor (Fig. 4).

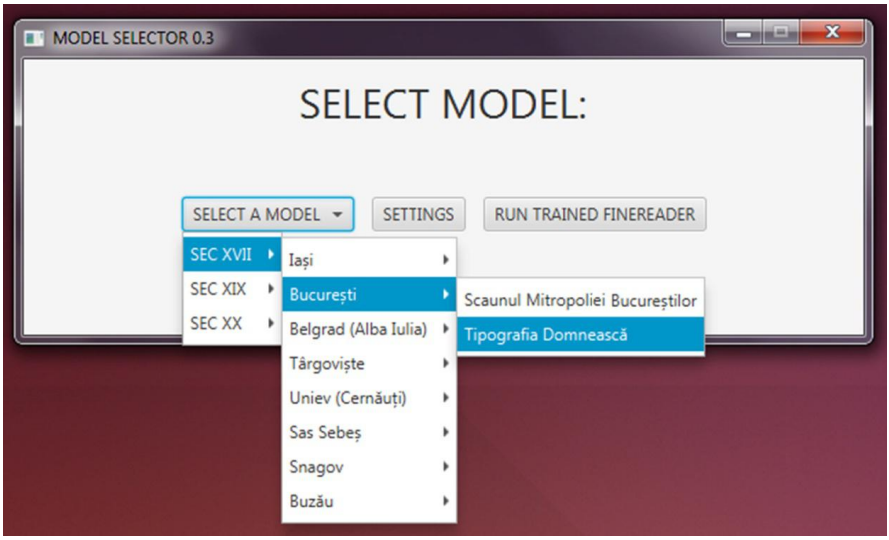

Fig. 4. Interfaţa de selectare a modelului OCR

Transliterarea în majoritatea cazurilor reprezintă o aplicaţie de tip „literă → literă", excepţie făcând şapte cazuri (**г, к, ч, ц, ѣ, ѧ, ѧ**) când este necesară o analiză simplă a contextului de dreapta cu lungimea de 1-2 simboluri. Spre exemplu, litera **ѧ** trece în **a** la începutul cuvântului şi după **ї, ц,** în **e** - după litera **ч,** şi în **ea** după orice consoană, la sfârşitul cuvântului.

Corectitudinea recunoaşterii depinde de calitatea imaginii, de cantitatea şabloanelor introduse şi de mărimea dicţionarului. Acurateţea la recunoaştere fără antrenare este destul de mică – 35%, în urma învăţării supervizate se poate ajunge la circa 70% din cuvinte recunoscute corect. Oricum, o corectare manuală rămâne necesară.

Procesarea textelor din secolele XIX - XX

Diverse surse atestă circa 17 versiuni de alfabete de tranziţie. Mai mult, varietatea scrierii poate fi întâlnită chiar şi în cadrul uneia şi aceleiaşi lucrări, unde unele pagini sunt tipărite doar cu caractere chirilice, altele mixte sau pur latine.

În recunoaşterea textelor tipărite cu alfabete mixte am utilizat două abordări. În cadrul primei textul scanat este reprodus în urma digitizării în glifele sale originale. Acest lucru este posibil prin configurarea şi antrenarea AFR, precum şi prin dotarea lui cu un dicţionar de epocă. Corectitudinea recunoaşterii este de circa 93%. Cea de a doua abordare a fost propusă pentru a soluţiona

problema varietăţii alfabetelor. AFR permite atât obţinerea rezultatului în glife originale, cât şi substituirea oricărei glife printr-o secvenţă de litere din alfabetul selectat. Finereader propune această metodă pentru ligaturi, dar ea poate fi utilizată în mod mai general pentru o substituţie arbitrară. În cazul alfabetelor de tranziţie a fost construită o versiune generalizată de alfabet de intrare în care se stabileşte o corespondenţă univocă cu un singur alfabet de ieşire, indiferent de modul de prezentare a literei de intrare. Spre exemplu, atât **т** (chirilic) cât şi **t** (latin) vor fi recunoscute drept **t**. S-a constatat, că în abordarea bazată pe ligaturi rata de acurateţe creşte până la 97% din cuvinte sau 0.6% caractere eronate faţă de 1.5% în cazul primei abordări.

Ca şi în cazul alfabetului chirilic vechi pentru transliterarea textelor din alfabete de tranziţie sunt suficiente două tipuri de reguli: „literă → literă” sau substituţie dependentă de context.

Digitizarea textelor tipărite cu caractere chirilice moderne (alfabet chirilic moldovenesc) se caracterizează printr-un grad mai înalt de acurateţe, atât graţie calităţii mai bune a imaginii, cât şi faptului că AFR este antrenat din start pentru recunoaşterea alfabetului rusesc, fiind necesară doar adăugarea literei **ӂ.** Corectitudinea recunoaşterii după antrenare depăşeşte 98%.

Însă o problemă mai dificilă pentru textele din această perioadă o constituie transliterarea. Ca şi în cazul alfabetului chirilic vechi şi al celor de tranziţie, pentru mai multe caractere există corespondenţa univocă „literă → literă”, pentru câteva litere problema se rezolvă prin analiza contextului, însă avem şi cazuri, când o soluţie exactă nu poate fi identificată. Din ele face parte transliterarea caracterului **я**, care poate fi reprezentat prin **ia, ea** sau **a**, spre exemplu: тр**я**з→tr**ea**z, ам**я**зэ→ am**ia**ză, абревиа→abrevi**a**. Circa 20 de reguli pentru această literă, atât euristice cât şi statistice, au permis să fie rezolvate mai multe cazuri, dar nu s-a ajuns la o soluţie completă.

O altă problemă parvine de la cuvintele de origine străină, care în scrierea cu caractere latine respectă forma din limba originală, pe când în alfabetul chirilic sunt aplicate principiile fonetice, spre exemplu, cuvântul **дизайн,** în transliterare directă ar fi **dizain,** pe când ortografierea corectă este **design.** Aceeaşi situaţie apare şi în cazul substantivelor proprii preluate din alte limbi. Numele **Шекспир (**Shakespeare) în transliterare directă apare ca **Şexpir**, etc. Toate aceste cazuri formează o listă de excepţii, care sunt procesate prioritar, urmând apoi aplicarea regulilor.

Programul de transliterare funcţionează în două moduri, direct şi actualizat. În cel de al doilea caz cuvintele sunt aduse la normele moderne de scriere. Spre exemplu, la scrierea cu grafie chirilică într-o serie de cuvinte litera ***i*** lipseşte sau este utilizată scrierea cu ***î*** în loc de ***i***: **требуе (trebue), ынтродучере (introducere)** etc. În funcţie de cererea utilizatorului este păstrată ortografia originală sau efectuată actualizarea, în exemplele de mai sus cuvintele respective fiind transliterate în **trebuie** şi **introducere**.

Resurse chirilice pentru OCR

Calitatea recunoaşterii textelor ar putea fi îmbunătăţită având la dispoziţie vocabulare cu caractere chirilice din perioada respectivă. În cazul secolului XX aceste vocabulare pot fi obţinute prin transliterarea inversă, adică din grafie latină în cea chirilică. Există mai multe resurse lingvistice, care pot fi supuse acestei transformări, de exemplu, DEX online[1] sau ELRR[2]. În urma analizei lor am considerat mai potrivit lexiconul elaborat la Universitatea „A. I. Cuza", Iaşi[3], care conţine circa 1 milion de intrări, este bine structurat şi însoţit de tag-uri morfologice. Transliterarea inversă se confruntăncu propriile dificultăţi, în particular, la procesarea literei **i**, care în grafie chirilică poate avea trei reprezentări: **и**, **й**, **ь** sau poate fi omisă. Exemple: arici → арич (la singular, litera **i** este omisă), arici →аричь (la plural, substituţia **i** → **ь**), [a] cheltui → келтуи (infinitiv, substituţia **i** → **и**), [eu] cheltui → келтуй (prezent, singular, persoana I, substituţia **i** → **й**). O problemă similară de lipsă a aplicaţiei univoce apare şi la transliterarea inversă a diftongilor şi triftongilor, spre exemplu, există trei variante posibile pentru diftongul **ia: я**, **ия** şi **иа.** Ca şi în cazul transliterării directe**,** rămâne problema cuvintelor de origine străină. Nu în toate cazurile a fost posibilă stabilirea de reguli formale, care ar permite automatizarea completă a procesului, pentru soluţionarea ambiguităţilor fiind necesare şi intervenţii manuale.

Procesarea începe cu selectarea cuvintelor incluse în dicţionarul de excepţii, asupra celor rămase se aplică o serie de filtre, care transliterează separat prefixele, sufixele, diftongii şi triftongii, apoi se efectuează filtrul final, care se reduce la aplicaţia „literă → literă".

[1] https://dexonline.ro/
[2] http://www.math.md/elrr/
[3] http://nlptools.info.uaic.ro/WebPosRo/resources/posDictRoDiacr.txt

Concluzii

Instrumentarul propus a demonstrat posibilitatea aplicării lui la digitizarea şi transliterarea textelor chirilice româneşti din diferite perioade istorice. După cum e şi firesc, cele mai dificile pentru recunoaştere sunt tipăriturile vechi, unde gradul de corectitudine este mai redus şi intervenţia manuală este mai solicitată. Mai puţin laborioase sunt transliterările textelor din secolul XX. Cu ajutorul tehnologiei propuse a fost digitizată şi transliterată cartea de matematică[4]. Procesarea unui volum de 224 de pagini a servit drept test pentru validarea instrumentarului. S-a constatat o calitate foarte bună a recunoaşterii, dar au fost semnalate mai multe erori de transliterare, majoritatea din ele fiind cauzate de ambiguităţi la interpretarea literei **я.** Nu în toate cazurile s-a reuşit actualizarea ortografiei, fiind necesare intervenţii manuale.

Svetlana COJOCARU
E-mail: svetlana.cojocaru@math.md
Constantin CIUBOTARU
Alexandru COLESNICOV
Ludmila MALAHOV
Tudor BUMBU

Institutul de Matematică şi Informatică, Academia de Ştiinţe
din Republica Moldova, Chişinău

[4] V. Andrunachievici, I. Chitoroagă. *Numere şi ideale*, Chişinău, „Lumina”, 1979.