UTILIZAREA RESURSELOR REUTIZABILE ALE LIMBII ROMÂNE ÎN PROCESUL DE RECUNOAŞTERE A TEXTELOR

E. Boian, C. Ciubotaru, S. Cojocaru, A. Colesnicov, L. Malahov
Institutul de Matematică şi Informatică al AŞM, Moldova
lena@math.md

Abstract. *The paper describes the technology for text digitalization and converting of scanned Romanian texts written with the Cyrillic letters. Full-text digitalization is performed using lexicon Reusable Romanian Linguistic Resources (RRLR) and spell checker RomSP. This technology can be applied for other kind of resources like some specific Romanian records written in the Cyrillic script.*

Cuvinte cheie: *digitalizarea textului, dechirilizarea textului, recunoaşterea textului, lexicon lingvistic.*

I. Introducere

Pentru a avea acces la tezaurul cultural, opere literare vechi, materialele vechi din arhive, creaţii ale autorilor români şi moldoveni, resurse folclorice specifice, este necesar să se creeze biblioteci virtuale care să includă aceste surse. O mare parte din aceste surse pot fi manuscrise sau texte tipărite cu litere chirilice. Există mai multe versiuni de alfabet chirilic folosit la scrierea textelor româneşti [1]. Vom menţiona doar alfabetul folosit pe tot teritoriul românesc în secolele XIV-XV (până în anul 1862) şi cel folosit în Republica Sovietică Socialistă Autonomă Moldovenească începând cu anul 1924 (cu o întrerupere între 1932-1938) şi, ulterior, în Republica Sovietică Socialistă Moldovenească. În figurile 1 şi 2 vedem exemple de texte scrise cu aceste alfabete.

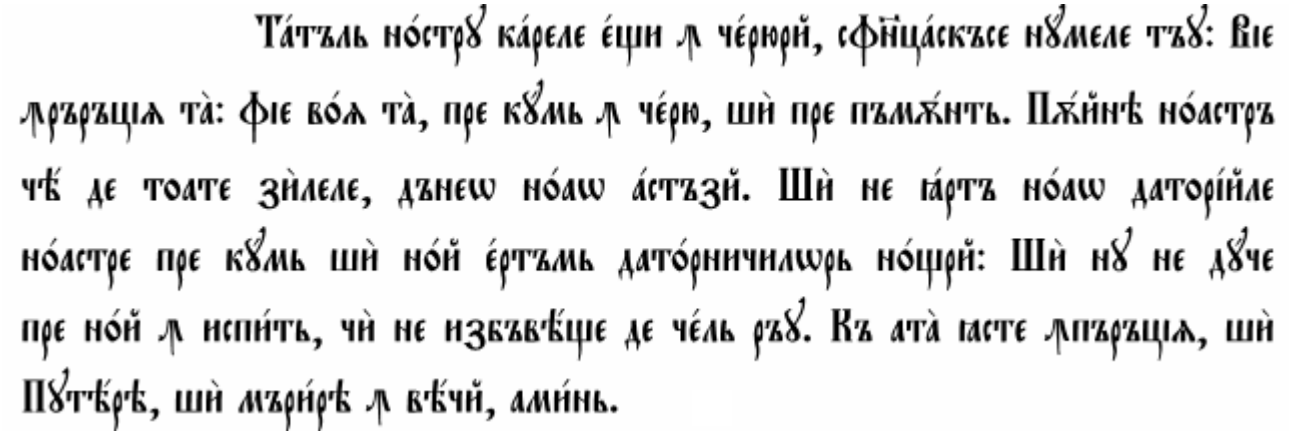

Fig.1. Rugăciunea „Tatăl nostru".

Ш'ачел реже-ал поезией, вечник тынэр ши фериче,
Че дин фрунзе ыць дойнеште, че ку флуерул ыць зиче…

Fig. 2. Mihai Eminescu, „Epigonii".

Pentru a le include în biblioteci virtuale ele trebuie sa fie digitalizate (scanate, recunoscute şi reprezentate ca texte cu litere latine). În cele ce urmează acest proces îl vom numi dechirilizare. În

lucrare se propune o tehnologie pentru dechirilizarea textelor, inclusiv a manuscriselor. Efectuarea acestor lucrări va permite unificarea, omogenizarea şi integrarea mediului cultural național în circuitul internațional, va consolida statutul limbii române ca limbă de comunicare pe continentul european.

II. Tehnologia de dechirilizare a textelor

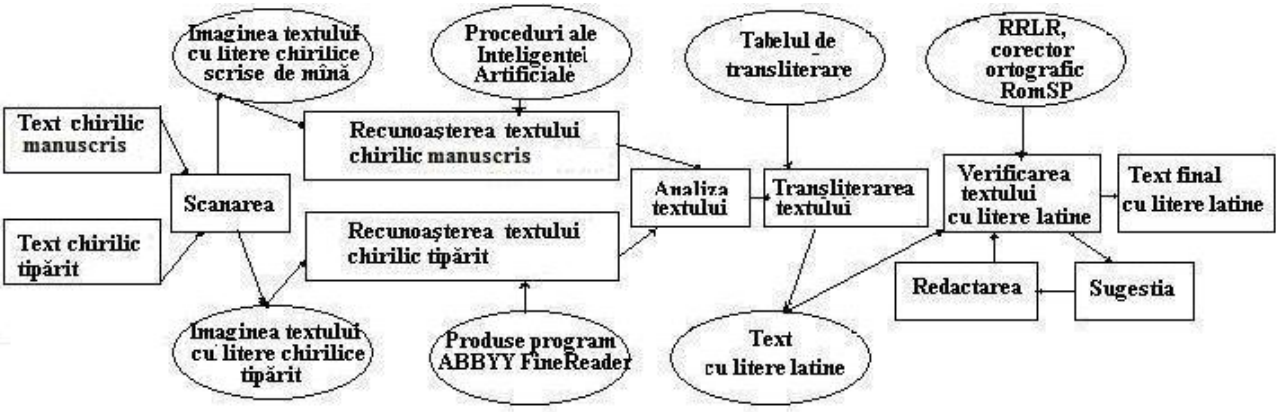

Fig.3. Procesul de dechirilizare a textului.

Procesul de dechirilizare a unui text (manuscris sau tipărit) este prezentat în Fig.3 şi este alcătuit din următoarele etape:

- scanarea, clasificarea şi indexarea textului;
- segmentarea şi recunoaşterea textului chirilic;
- transliterarea textului;
- verificarea textului cu litere latine utilizând corectorul ortografic RomSP [2] şi resursele reutilizabile ale limbii române (RRLR) [2, 3].

II.1. Scanarea, clasificarea şi indexarea textelor

Conversia documentelor în format electronic constă din mai multe etape. La prima etapă documentele sunt scanate. Următoarea etapă este clasificarea, care permite regruparea documentelor cu ajutorul unor şabloane predefinite (de exemplu, scrisori, articolele de ziar, contracte sau facturi, etc.). După scanare şi clasificare se va face indexarea documentului electronic, care constă în generarea unor câmpuri cu informații referitor la text: data, numele autorului, titlul, etc., Clasificarea şi indexarea pot fi efectuate manual sau semiautomat. Regimul semiautomat presupune utilizarea în regim interactiv a unor reguli speciale, care contribuie la sporirea acurateței şi accelerează procesul de completare.

Exportul datelor extrase din diverse aplicații, de exemplu, business, baze de date, arhive electronice, se poate face în diverse formate electronice. Pentru arhivele electronice, de regulă, se utilizează formatul pdf.

Următorul pas este recunoaşterea imaginii şi transformarea ei în format textual.

II.2. Recunoaşterea textului chirilic

Imaginea textului chirilic manuscris poate fi transformată în format textual utilizînd reţelele neuronale pentru recunoaşterea literelor chirilice. Aceste reţele se bazează pe metoda de auto-organizare dezvoltată de T. Kohonen [4, 5]. Abilitatea de recunoaştere a reţelei este rezultatul unui proces de învăţare pornind de la exemple de recunoaştere corectă. Ca date de intrare se utilizează descrierea sintetică a unui obiect (de exemplu, descrierea grafică a unei litere, sau ansamblul caracteristicilor acesteia). În majoritatea situaţiilor descrierea este un vector de valori numerice obţinute printr-o prelucrare prealabilă (preprocesare) a informaţiilor brute. În calitate de date de ieşire se obţine un indicator al clasei căreia îi aparţine obiectul (de exemplu, numărul de ordine al literei în cadrul alfabetului).

Imaginea textului chirilic tipărit poate fi recunoscută cu ajutorul produsului program ABBYY FineReader [6], care converteşte orice fişier în format editabil, obţinut prin scanarea documentelor. Acestea pot fi transformate în formate editabile ale pachetului Microsoft Office (Microsoft Word, Excel), *rtf, html, txt* etc. Recunoaşterea optică a caracterelor (OCR) transformă imaginile în text. OCR face posibilă editarea şi reutilizarea textului inserat în imaginile scanate. OCR utilizează o formă de recunoaştere a imaginii, cunoscută sub numele de recunoaştere a modelului pentru identificare individuală a caracterelor într-un text dintr-o pagină, inclusiv semnele de punctuaţie, spaţiile şi sfârşitul de linie.

II.3. Dechirilizarea textului

Procesul de dechirilizare presupune recunoaşterea caracterelor chirilice, segmentarea cuvintelor, transcrierea cu caractere latine şi recunoaşterea ulterioară utilizînd resursele lingvistice existente şi algoritmi pentru sugestii (corectarea semiautomată).

Pentru transliterare se va utiliza tabelul cu reguli de transliterare a textelor cu litere chirilice în cele cu litere latine. Pentru verificarea textului obţinut se vor utiliza RRLR şi corectorul ortografic RomSP. RRLR conţin o bază de date cu informaţie lingvistică pentru limba română la nivel de cuvînt şi un set de programe de gestionare a bazei de date. Ca volum RRLR conţine circa un milion de cuvinte.

Verificarea textului se efectuează la nivel de cuvânt cu RomSP folosind RRLR. Textul se verifică la nivel de cuvânt în modul următor. Dacă cuvântul selectat este găsit în RRLR, rezultă că el este corect şi se selectează următorul cuvânt din text. În caz contrar (cuvântul este greşit sau nu este inclus în RRLR) se iniţiază un dialog, care propune următoarele opţiuni:

- a corecta cuvântul manual,
- a solicita prin intermediul sugestiei o listă de cuvinte din RRLR similare cuvântului greşit şi a înlocui cuvântul greşit cu unul corect din lista de sugestii;
- a cere eliminarea cuvîntului sau marcarea acestuia pentru a-l prelucra ulterior consultând un dicţionar de specialitate sau un expert lingvist;
- pentru verificarea corectitudinii unui cuvînt care lipseşte în dicţionar, cum ar fi nume, prenume, abreviere, denumire geografică etc. acest cuvînt se va declara ca unul corect pentru restul sesiunii de lucru cu textul;
- totodată există posibilitatea includerii cuvântului nou în RRLR (folosind un set de programe de gestionare cu funcţia de flexionare).

Vom menţiona, că aplicarea acestei tehnologii pentru procesarea textelor din surse vechi necesită şi completarea lexiconului cu termenii specifici perioadei şi domeniului respectiv.

III. Concluzii

Tehnologia descrisă permite digitalizarea şi recunoaşterea resurselor specifice cu utilizarea lexiconului RRLR şi corectorului ortografic RomSP. Această tehnologie poate fi propusă pentru digitalizarea resurselor de altă natură, cum ar fi resursele folclorice specifice, scrise cu litere chirilice. Resursele digitalizate vor reprezenta manuscrise şi înregistrări specifice stocate într-o bază de date şi protejate prin metode originale special elaborate. Pentru facilitarea accesului utilizatorilor la aceste resurse va fi elaborată o interfață specială şi un set de instrumente care vor permite recunoaşterea textelor. Crearea unei platforme de tipul e-learning ar permite utilizarea resurselor digitalizate ca materiale didactice de instruire.

Crearea resurselor specifice, plasarea acestor resurse pe Internet pentru acces public vor extinde aria şi posibilitățile de cercetare, inclusiv şi în domeniul ştiinţelor umanitare, prin modificarea mediului comunicativ informațional. Aceasta se manifestă prin posibilitatea aplicării tehnologiilor propuse pentru alte limbi, de exemplu, limba sîrbă.

Această tehnologie va putea fi utilizată şi la completarea resurselor lingvistice RRLR cu cuvinte noi, extrase din materialele digitalizate, atestate de experți lingvişti.

IV. Referințe

1. Bărbulescu Ilie. Fonetica alfabetului cirilic în textele române din vécul XVI şi XVII. Anul 1904, Bucureşti.
2. 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.
3. http://www.math.md/elrr/
4. Egmont-Petersen, M., de Ridder, D., Handels, H. (2002). "Image processing with neural networks - a review". Pattern Recognition 35 (10), pp. 2279–2301.
5. Kohonen T. 1988. Self-organization and associative memory. 2d ed. New-York, Springer-Verlag.
6. http://www.abbyy.com/