DIGITIZAREA ȘI VIZUALIZAREA ALFABETULUI CHIRILIC ROMÂNESC DE TRANZIȚIE (1830 – 1860)

Cercetător științific **Valentina DEMIDOVA**
Doctor în științe informatice **Ludmila BURȚEVA**
Institutul de Matematică și Informatică al AȘM

DIGITIZATION AND VISUALIZATION OF 1830–1860 TRANSITION ROMANIAN CYRILLIC ALPHABET
Summary. Presented paper concerns problems which arise together with growing of databases of scanned ancient texts. Historical text saved as image is unsuited for modern methods of textual information processing. Automatic digitization resolves this problem, but requires specific methods for representation of unusual today scripting. This paper presents the solution for digitizing of Romanian Cyrillic of printed texts of transitional period 1830–1860.
Keywords: automatic text digitization, Romanian Cyrillic, decyrillization.

Rezumat. Lucrarea prezentată se referă la problemele care apar odată cu creșterea volumului bazelor de texte vechi scanate. Textul istoric păstrat ca imagine nu este potrivit pentru metodele moderne de procesare a informațiilor textuale. Digitizarea automată rezolvă această problemă, dar necesită metode specifice pentru reprezentarea scriptului neobișnuit la ziua de azi. Această lucrare prezintă o soluție pentru digitizarea textelor tipărite în alfabetul chirilic românesc în perioada de tranziție 1830–1860.
Cuvinte-cheie: digitizarea automată a textelor, alfabetul chirilic românesc de tranziție, dechirilizare.

INTRODUCERE

Prin dechirilizare înțelegem procesul de translitera-re a limbii române din scrierea în grafia chirilică din perioada de tranziție (1830–1860) în cea cu grafie latină.

De ce problema dechirilizării a devenit importantă? Cauza principală constă în faptul că cercetările actuale în domeniul conservării patrimoniului cultural s-au concentrat pe particularitățile fiecărei culturi europene [1]. Textele scrise în alfabetul chirilic românesc de tranziție, spre deosebire de cele scrise în grafia chirilică contemporană, sunt răspândite nu doar în Moldova, ci și în România. Citirea acestor texte, care constituie un patrimoniu vast [2], prezintă dificultăți în special pentru specialiștii care nu sunt familiarizați cu grafia chirilică. Odată cu epoca digitală devine oportună și posibilă digitizarea acestor texte vechi, existente doar pe hârtie, astfel ca ele să fie accesibile tuturor.

Drept exemplu, vom apela la un fragment din balada lui Vasile Alecsandri (figura 1), care a fost publicat cu alfabetul chirilic românesc de tranziție, transpus, în paralel, în grafia latină.

1. VIZUALIZAREA SIMBOLURILOR ALFABETULUI CHIRILIC ROMÂNESC DE TRANZIȚIE

Problema vizualizării textelor istorice apare odată cu digitizarea integrală a acestora, în loc de stoca-

Сѹс пе кътпѹл Nistрѹлѹї,	Sus pe cămpul Nistrului,
Сѹв поалеле черѹлѹї,	Sub poalele cerului,
Ла коада Їалпъѹлѹї,	La coada Ialpăului,
Ѹnde фатѣ Zmeoaїчеле	Unde fat' Zmeoaicele.
Ші с'adѹnѣ Зърnoaїчеле	Şi s-adun' zărnoaicele
Ші с'adъпѣ Леѣaїчеле..	Şi s-adăp' Leuaicele..

Figura 1. Fragment din balada *Român Grue-Grozovanu* de V. Alecsandri

re a imaginilor scanate. Prin distribuția suportului standardului de Unicode16 pe o varietate de sisteme de operare au fost alocate coduri pentru orice scriere istorică: cuneiformă, hieroglife, scriere celtică etc. Dispunând de codurile corespunzătoare, problema vizualizării poate fi rezolvată prin intermediul fonturilor specializate.

Pentru alfabetul chirilic românesc de tranziție există deja trei fonturi: Kliment Std, Everson, Roman Cyrillic Std care au diferite imagini, dar care cuprind toate cele 43 de caractere „canonice". Cu toate acestea, în timpul formării versiunii canonice a alfabetului de tranziție au apărut litere care nu sunt incluse în setul final. Astfel, textele electronice obținute sunt incomplete, iar specialiștii în filologie istorică rămân lipsiți de informații foarte necesare.

Problema se poate rezolva pe seama codurilor libere existente în diapazonul Unicode-6 acordat pentru chirilică. În plus, există dificultăți caracteristice pentru alfabetul intermediar, de exemplu, utilizarea caracterelor modificate ale alfabetului latin pentru afișarea caracteristicilor fonetice, în special **„ĭ"**, **„ŭ"** pentru reprezentarea vocalelor scurte. Evident, o astfel de marcă fonetică prezintă mare interes pentru istorici, de aceea ea trebuie să fie stocată în text. Simbolurile relevante sunt disponibile în versiunea extinsă a Unicode latină, astfel problema se rezolvă prin implicarea codurilor suplimentare.

În această lucrare este prezentat un modul al instrumentarului de digitizare a textului istoric din alfabetul chirilic românesc de tranziție. Acest instrumentar [3] a fost elaborat de grupul de lingvistică computațională al Laboratorului „Sisteme de Programare" de la Institutul de Matematică și Informatică al AȘM. Instrumentarul de digitizare oferă utilizatorului acces prin pagina web și permite alegerea perioadei de timp. În funcție de alegerea utilizatorului, se activează modulul respectiv al instrumentarului. Modulul prezentat face conversia alfabetului chirilic românesc de tranziție în alfabetul latin pentru perioada 1830–1860.

Instrumentarul este scris în limbajul Java [4], pentru care suportul multiplatformei și, prin urmare, suportul Unicode, sunt proprietăți de bază. În cazul în care fontul corespunzător este înregistrat în sistemul de operare, instrumentele de programare de interfață ale limbajului Java rezolvă problema de vizualizare printr-o simplă adresare la această înregistrare.

În tabelul 1 este prezentată corespondența literelor specifice ale alfabetului de tranziție cu scrierea lor în grafia latină și codul UTF16.

Tabelul 1

Corespunderea literelor specifice și codurile lor UTF16

Ѣ	Ea	0462
ѣ	ea	0463
Ꙗ	Ia	0465
ꙗ	ia	0464
Ѧ	Â	0466
ѧ	â	0767
Ѧ̀	Î, Îm, În	A64e
ѧ̀	î, îm, in	A65f
Ꙋ	U	A64a
ꙋ	u	A64b
К	C, Ch (înainte de e, и)	041a
к	c, ch (înainte de e, и)	043a
Ӑ̆	Ĭ	012C
ӑ̆	ĭ	012D
Ъ	Ă	042A
ъ	ă	044A
Щ	Șt	0429
щ	șt	0449
Ц	G, Gh (înainte de e, и, ю)	049F
ц	g, gh (înainte de e, и, ю)	044F

2. DESCRIEREA REGULILOR DE TRANSLITERARE (DECHIRILIZARE)

La elaborarea şi descrierea regulilor de transliterare a cuvintelor din limba română, scrise în grafia chirilică a alfabetului de tranziție românesc, într-un echivalent scris în grafia latină, ne-am condus de normele gramaticale în vigoare. Vom identifica două tipuri de reguli: elementare şi complexe.

Alfabetul chirilic de tranziție pentru limba română s-a modificat pe parcursul utilizării sale, înregistrând circa 20 de versiuni. Cercetările noastre vizează un număr maximal de litere care constituie 43 de simboluri. În lucrarea prezentă examinăm 36 de litere întâlnite în textele analizate, dintre care 27 sunt cele contemporane, plus nouă litere vechi: Ѣ ѣ, Ꙗ ꙗ, Ѧ ѧ, Ѧ̀ ѧ̀, Ꙋ ꙋ, Ӑ̆ ӑ̆, Ъ ъ, Щ щ, Ц ц. Menționăm că algoritmul nostru permite adăugarea simplă a literelor care ar putea apărea în procesul viitoarei analize a textelor.

În conformitate cu regulile elementare se efectuează transliterarea a 31 de litere. Pentru simbolurile chirilice, care sunt identice pentru toate variantele alfabetului chirilic, noi folosim regulile descrise în [5]. Regulile elementare sunt cele ce efectuează punerea în corespondență independentă de context a unei litere chirilice cu una sau mai multe litere din grafia latină. Literele prezentate în tabelul 2 sunt traduse în baza regulilor elementare.

Tabelul 2

Regulile elementare de transliterare

Chirilic	а	б	в	д	е	ж	з	и	й	л	м	н	о	п	р	с	т	ф	х	ц	ш	ь	э	ю	ѣ	Ѥ	Ѧ	ѳ	ѵ	ъ
Latin	a	b	v	d	e	j	z	i	i	l	m	n	o	p	r	s	t	f	h	ţ	ş	i ă	iu	ea	ia	â	u	ĭ	ă	

Rămân cinci litere care necesită o procesare suplimentară şi anume: „г”, „к”, „ч”, „ѧ”, „ц”.

Conversiile acestor litere depind de context, ceea ce vom descrie mai jos. Procesul de dechirilizare îl vom prezenta în tabelele 3-6.

Tabelul 3

Reguli pentru literile *„г”* şi *„ц”*

г, ц	**gh**	înainte de: e, и, ю
	g	în restul cazurilor

Exemple: ***цени8*** **– gheniu,** ***богат*** **– bogat**

Tabelul 4

Reguli pentru litera *„к”*

к	**ch**	înainte de: e, и, ю
кс	**x**	
к	**c**	în restul cazurilor

Exemple: ***к8приндe*** **– cuprinde,** ***кѥр*** **– chiar**

Tabelul 5

Reguli pentru litera „ ”

ѧ	**î**	înainte de: m, n
	îm	înainte de: b, p
	în	în restul cazurilor

Exemple: **ѧн – în ,** **ѧнчеп8т– început,** **ѧpune – împune**

Tabelul 6

Reguli pentru litera „ч”

ч	**c**	înainte de: e, и
	ce	înainte de: a
	ci	înainte de: o, y
	ci	la sfârşit de cuvânt
	ci	înainte de consoană

Exemple: причина – pricina, фаче – face, чинститъ – cinstită, ѧнторчї – întorci

3. ANALIZA COMPARATIVĂ A PROCESULUI DE TRANSLITERARE PENTRU DIFERITE PERIOADE DE TIMP

În continuare vom face o analiză comparativă a procesului de dechirilizare pentru literele alfabetului de tranziţie (1830–1860) cu cel descris în [6] pentru alfabetul şi ortografia din perioada 1945–1989. Algoritmul propus în [5] a stat la baza algoritmului prezent. După cum am menţionat anterior, în conformitate cu regulile elementare se efectuează transliterarea a 31 de litere. Pentru unele litere (***„г”, „к”, „ч”, „ц”)*** regulile de transliterare, deşi nu sunt elementare, coincid cu cele pentru alfabetul chirilic contemporan [6]. Aşadar, existenţa unui surplus de litere (43) a adăugat doar o singură regulă adiţională – cea cu litera ѧ, în schimb a rezolvat mai multe probleme care au apărut la procesul de transliterare a chirilicii contemporane, inclusiv:

Dispare problema ce ţine de ortografia nouă, datorită faptului că alfabetul de tranziţie conţine următoarele simboluri pentru sunetul ***„î”*** în poziţii diferite (tabelul 7).

Tabelul 7

Literele care corespund sunetului *„î ”*

Ѧ	**Â**
ѧ	**â**
Ѩ	**Î, Îm, În**
ѧ	**î, îm, in**

Se rezolvă problema ce ţine de transliterarea literei **„я”**.

După cum a fost descris în mod detaliat în [5], ne-am confruntat cu mari dificultăţi la procesul de transliterare a literei ***„я”***. Aceste dificultăţi au fost legate de faptul că transliterarea literei ***„я”*** este ambiguă şi depinde de context. În special, problema dată apare în ocurenţa lui ***„я”*** în interiorul cuvântului. Regula de transliterare poate fi atât „я” – ***„ia”,*** cât şi **„я”** – ***„ea”.*** Menţionăm că stabilirea acestei reguli poate fi efectuată doar prin apelarea la dicţionarele externe, de exemplu la DEX.

Aşadar, alfabetul de tranziţie rezolvă problema ce ţine de litera **„я”**: fără a o conţine, dispune de simboluri specializate pentru diftongi (tabelul 8).

Tabelul 8

Simbolurile pentru prezentarea diftongilor *„ia”* şi *„ea”*

Ѣ	Ea
ѣ	ea
ѧ	Ia
ѥ	ia

Deoarece problemele enumerate mai sus au fost soluţionate, transliterarea alfabetului de tranziţie a fost efectuată cu succes în 99% de cazuri. Trebuie de menţionat faptul că succesul transliterării chirilicii moderne a constituit maximum 91%.

4. EXEMPLE DE TRANSLITERARE

În acest compartiment vom aduce un exemplu de dechirilizare a unui text din ***Gramatica Românească, editată în București, 1835*** (figura 2), care ilustrează complexitatea problemei.

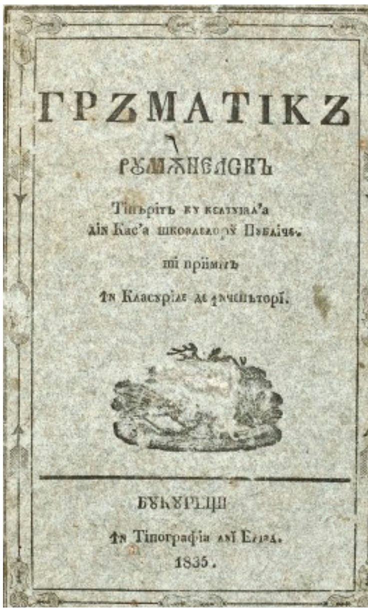

Figura 2. Coperta cărții *Gramatica Românească*

În plus, acest exemplu demonstrează cât de util poate fi instrumentul prezentat pentru specialiști, îndeosebi pentru cei care nu sunt familiarizați în scrisul chirilic:

În figura 3 este prezentat fragmentul textului original. În continuare urmează textul obținut în urma dechirilizării:

ꙟнтродꙋчере.

Грьматік'а есте ꙋнꙋ мешешꙋгꙋ прін каре ꙟнвъцъмꙋ а ворбі ші а скрі о лімбъ фъръ грешалъ.

Лімбъ нꙋмімꙋ ачелꙋ міжлокꙋ прін каре фачемꙋ кꙋноскꙋте ші алтор'а ідеіле саꙋ кꙋноштінцеле ноастре; астꙋ-фелꙋ сѫнт Ворбірea ші Скріерѣ.

Ка съ ворбімꙋ не слꙋжімꙋ кꙋ ніше сꙋнете саꙋ гласꙋрі че ле пронꙋнціъмꙋ ꙟмпреꙋнате, ші пе каре ле нꙋмімꙋ ꙁічері, кѫнд арътꙋ о фіінцъ саꙋ ꙋнꙋ лꙋкрꙋ, кꙋм омꙋ, карте.

Ка съ скрімꙋ не слꙋжімꙋ кꙋ Літеріле саꙋ Словеле, каре сѫнтꙋ ніше семне че арътꙋ сꙋнетеле ꙋнеі лімбі, кꙋм б, а, шчл.

Ноі авемꙋ доъ-зечі ші шапте де слове, ші ле ꙟмпърцімꙋ ꙟн Гласніче ші Консꙋнате.

Гласнічеле сѫнт шапте: а, е, і, о, ꙋ, ъ, ѫ, ші ле зічемꙋ астꙋ-фелꙋ, кьчі сінгꙋре ші фъръ ажꙋторꙋлꙋ алтеі слове факꙋ ꙋнꙋ сꙋнетꙋ, ꙋнꙋ гласꙋ.

Figura 3. Fragmentul textului original

Introducere

Grămatic'a este unu meșteșugu prin care învățămu a vorbi și a scri o limbă fără greșală.

Limbă numimu acelu mijlocu prin care facemu cu noscute și altor-a ideile sau cunoștințele noastre; astu' felu sânt Vorbirea și Scrierea.

Ca să vorbimu ne slujimu cu niște sunete sau glasuri ce le pronunţiămu împreunate, şi pe care le numimu zi'ceri, când arătu o fiinţă sau unu lucru, cum omu, carte.

Ca să scrimu ne slujimu cu Literile sau Slovele, care sântu niște semne ce arătu sunetele unei limbi, cum b, a, şcil.

Noi avemu doă-zecii şi şapte de slove, şi le împărţimu în Glasnice şi Consunate.

Glasnicele sânt şapte: a, e, i, o, u, ă, â, şi le zicemu astu'felu, căcii singure şi fără ajutorulu altei slove facu unu sunetu, unu glasu.

CONCLUZII

Regulile descrise mai sus au servit drept bază la elaborarea algoritmului de dechirilizare a limbii române de tranziție.

Dechirilizarea prezintă un instrument care oferă mijloace de suport pentru specialiști în domeniul filologiei istorice și al conservării patrimoniului cultural.

Transformarea textului în grafia latină asigură aplicabilitatea metodelor de digitizare și a tehnicilor de cercetare a textelor istorice, bazate pe alfabetul latin.

În afară de suportul specialiştilor de cercetări istorice și filologice, instrumentul propus oferă acces liber la texte de importanță patrimonială, deoarece funcționează on-line și prezintă textele istorice într-o formă ușor de înțeles cititorului de astăzi.

Pe parcursul elaborării în continuare, instrumentul propus va fi completat cu mijloace de analiză ale textelor istorice.

Deoarece versiunea curentă a instrumentarului prezentat a fost aplicată cu succes la analiza diacronică manuală [7], realizarea modulului care efectuează această analiza în mod automat este pe primul plan. Implementarea acestui modul include atât aplicarea tehnicilor existente, cât și realizarea tehnologiilor proprii, care țin cont de specificul corpusului.

BIBLIOGRAFIE

Peneva, Juliana and Ivanov, Stanislav and Sotirova, Kalina and Doneva, Rossitza and Dobreva, Milena (2012) Access to Digital Cultural Heritage: Innovative Applications of Automated Metadata Generation Chapter 1: Digitization of Cultural Heritage – Standards, Institutions, Initiatives. In: Access to Digital Cultural Heritage: Innovative

Applications of Automated Metadata Generation. Plovdiv University Publishing House „Paisii Hilendarski", Plovdiv, p. 25-67.

Cojocaru S., Boian E., Ciubotaru C., Colesnicov A., Demidova V., Malahov L. Regeneration of printed cultural heritage: challenges ang technolologies. Chișinău: The Third Conference of Matematical Society of the Republic of Moldova, 19-23 August, 2014, p. 481-489.

S. Cojocaru, L. Burtseva, C. Ciubotaru, A. Colesnicov, V. Demidova, L. Malahov, M. Petic, T. Bumbu, Ș. Ungur. On Technology for Digitization of Romanian Historical Heritage Printed in the Cyrillic Script, Proceedings of MFOI2016. Conference on Mathematical Foundations of Informatics, July 25-31, 2016, Chisinau, Republic of Moldova, p. 160-177.

Richard M. Reese. 2015. Natural Language Processing with Java. Packt Publishing.

Demidova Valentina. Particularitățile dechirilizării limbii române. In: Studia universitatis moldaviae, 2015, nr. 2(82), Seria „Științe exacte și economice". Online 2345-1033, p. 16-20.

Demidova V. Particular Aspects of the Cyrillization Problem. Chișinău: The Third Conference of Matematical Society of the republic of Moldova , 19-23 August, 2014, p. 493-498.

D. Gifu, M. Dascalu, S. Trausan-Matu and L. K. Allen, „Time Evolution of Writing Styles in Romanian Language," 2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI), San Jose, CA, 2016, p. 1048-1054.

Igor Vieru. *Băiețelul din coliba albastră* de Sp. Vangheli. Hârtie, guașă, 1964.
Muzeul Național de Artă al Moldovei