//8.4.16 v0.6

Algoritmizarea procesului de transliterare pentru actualizarea textelor chirilice românești

Introducere

În această lucrare de licență voi prezenta detaliat demersul realizării unei aplicații de transliterare (convertare a textelor din grafia chirilico-slavonă în cea latină), și realizarea unui algoritm eficient pe baza căruia va fi realizată aplicația dată. De asemeni voi prezenta detaliat despre ceea ce trebuie cunoscut și ceea ce trebuie de luat în considerare atunci cînd se realizează un astfel de convertor *(aplicație de convertare)*. Voi trece în revistă principiile de bază a alfabetelor și ortografiei textelor chirilice din 3 perioade istorice: chirilica sovietică (sec. XX), chirilica română slavonă (sec. XVIII) și chirilica tranzițională (sec. XIX). În această lucrare voi evidenția principiile de bază realizării unui algoritm pentru automatizarea tranzliterării*(conversiei)* dintr-o grafie în alta.

Pe parcursul lucrării voi pune accent atît pe importanța realizării unui astfel de algoritm cît și pe crearea unei aplicații finale cu interfață grafică accesibilă și un funcțional clar, și totodată performant. Voi defini caracteristicele importante la nivel vizual și tehnic pentru diferite tipuri de aplicații de convertare(*convertori*) clasificate după scop. Comportamentul și așteptările unui utilizator care se folosește de astfel de aplicații sunt corelate cu tipul și scopul acelor aplicații. Voi vorbi despre cum ar trebui să fie redactate textele inițiale (*chirilice*) și ce criterii ar trebui să satisfacă pentru ca aplicația de convertare să producă cel mai eficient rezultat posibil. Deoarece aplicația convertează textele dintr-o grafie în alta, ea nu poate corecta greșelile de ortografie sau de conținut(*sens*), așadar textul final în grafie latină va conține toate acele grașeli odată ce ele au fost prezente în textul chirilic. De asemeni dicționarul de cuvinte recunoaște doar cuvintele chirilice și latine scrise în forma lor exactă(*corectă*), așadar el nu va avea nici un efect asupra cuvintelor cu greșeli ortografice. Astfel aplicația de convertare(*transliterare*) nu se rezultă doar convertarea nemijlocită a textelor propuse, ci și la extragerea potențialului(*rezultatului*) maxim, utilizînd tot pachetul de instrumente și funcționalități propuse (dicționarul de cuvinte, lista de prefixe, posibilitatea alegerii standardului necesar etc.). Atît cum aplicația necesită să aibă o interfață grafică accesibilă și ușor de înțeles, iar pachetul de intrumente să fie la fel ușor accesibil, o aplicație de tipul dat nu se resumă doar la garfică sau design. Orice utilizator pune în primul rînd calitatea conținutului, iar calitatea unei aplicații se poate determina atît pe baza aspectelor vizuale cît și pe cele tehnice. Voi face un studiu comparativ între mai multe aplicații de translare(*convertare*) atît de realizare străină cît și a celor românești. Analiza succintă se va evalua atît la nivel tehnic cît și vizual, cu scopul de a extrage concluzii în vederea realizării proiectului de licență: aplicația de transliterare(*pe 3 nivele*) ABConv.

Pentru realizarea unei aplicații de acest gen, dar mai ales pentru realizarea unui algoritm de convertare automată dintr-o grafie în alta sunt necesare cunoștințe din diferite ramuri desprinse din filologie, istorie lingvistică etc. Prin aprofundare domeniilor precum tipografia, istoria alfabetelor românești și a celor chirilice, studiul ortografiei slavone, realizarea unui astfel de de algoritm corect și eficient este considerabil facilitată, și rezultatele sunt net superioare.

Pentru a realiza conversia(*transliterarea*) eficientă și rapidă a unui text din grafia chirilică(*slavonă*) în cea latină aplicația realizată va trebui să înglobeze elemente tehnice și ortografice specifice acestora: regulile ortografice, corectitudinea textului, păstrarea proporțiilor(*spațiilor*) textului inițial, definirea unor cazuri speciale pentru caracterele ce au semnificație multiplă etc. Așadar un text chirilic nu poate fi convertat mot-a-mot, acest proces trebuie realizat ținînd cont de numeroase reguli lingvistice pentru a asigura păstrarea mesajului inițial al textului. Programatorul este nevoit să aibă în prealabil un grad de cunoaștere destul de înalt atît în domeniul programării cît și al unor domenii precum filologia, lingvistica, istoria și evoluția alfabetică(*alfabetelor*), ortografia grafiei slavone etc. Toate aceste tehnologii vor fi folosite pentru a asigura corectitudinea transliterării.

Orice aplicație, atît desktop cît și web, la început nu este altceva decît un algoritm de procesare a datelor de intrare și obținerea datelor de ieșire necesare, iar fiecare algoritm la început se eloaborează, testează, cizelează, pentru ca în final să fie implementat în aplicația dată. Interactivitaea aplicației se bazează pe pe aceste tehnologii, iar limitele impuse la nivel de implementare trebuie luate în considerare în conceperea unei aplicații. Mai mult decît atît, este important de luat în considerare și grupul țintă de utilizatori căruia se adresează aplicația. Chiar și dacă aplicația nu este îndreptată spe un public larg de utilizatori, oricum trebuie de ținut cont de necesitățile și posibilitățile lor (vîrsta, profesia etc.), cu scopul de a realiza interfața grafică și pachetul de instrumente cît mai eficient. În acest mod se poate determina mai ușor cum trebuie să arate designul, ce funcționalități trebuie să implementeze și ce grad de accesibilitate este necesar. Totodată, prin efectuarea studiilor de marketing (studierea aplicațiilor asemănătoare, definirea grupului țintă de utilizatori etc.) gui-ul și unele funcționalități ale aplicației pot fi considerabil schimbate, astfel încît să corespundă necesităților utilizatorilor.

Principii generale

În general, fiecare aplicație indiferent de tipul ei trebuie să fie unică, deci să aibă elemente distincte atît la nivel de design cît și la nivel tehnic, ceea ce ține de funcționalitățile oferite utilizatorului. Programatorul are datoria de a realiza o aplicație unică prin designul interfeței grafice și funcționalitățile(*instrumentele*) utilizate. O aplicație bine realizată va fi ușor accesibilă atît la nivel vizual cît și tehnic, iar conținutul prezentat și specificul acestuia trebuie să fie conceput astfel încît utilizatorului să-i fie cît mai ușor să-l acceseze.

Realizarea oricărui proiect începe cu planificarea, și în etapa de planificare trebuie stabilite tipul de aplicație(instiuțională, educativă, de prezentare a unui produs etc.), ce intrumente vor fi oferite utilizatorului, preferințele vizuale, structura aplicației la lansare, ce informații vor fi la prima vedere, unde și cum vor fi amplasate butoanele, meniurile, tehnologiile necesare pentru realizarea aplicației, etc. Toate aceste informații sunt foarte importante în realizarea interfeței, deoarece influențează foarte mult și în mod direct forma finală a aplicației proiectate. Interfața grafică este prima ce utilizatorul vede și unica lui sursă de a comunica cu algoritmul aplicației, respectiv aspectul ei vizual și comoditatea utilizării este unul din pilonii principali pe care se bazează dezvoltarea unuei aplicații. O aplicație bogată, cu structură de meniuri mare și multe informații nu poate fi în principiu foarte intensă la nivel vizual, deoarece în acest mod va fi subminat scopul aplicației prin distragerea atenției de la conținut, și respectiv reducerea importanței sale.

Astfel, o interfață trebuie să fie mai simplă, mai concisă, pentru a pune în valoare conținutul și instrumentele necesare, astfel a le expune utilizatorului într-un mod simplu și accesibil.

/\*

\* 7.4.16

\* Art. din cartea fotografiată/scanată.

\*

\*/

Digitizarea, recunoașterea și conservarea patrimoniului cultural istoric

**Rezumat.** În lucrare se abordează probleme ce apar în procesul de digitizare și recunoaștere lexicografică a textelor vechi românești, se argumentează necesitatea creării resurselor electronice specifice care caracterizează evoluția limbii române moderne. Se prezintă rezultatele statistice obținute la recunoașterea unui text românesc din secolul al XIX-lea, utilizîndu-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. Aceată tehnologie se bazează pe transliterare și pe aliniere paralelă a textelor.

**Cuvinte-cheie:** digitizare, resurse lingistice românești, recunoașterea textului, tehnologia limbajului, alfabet chirilic, transliterare, alinierea textelor.

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

Dezideratele principale ale politicii culturale pentru zonele unde se vorbește limba română țin de studierea, valorificarea și digitizarea patrimoniului cultural-istoric. Procesul de digitizare a patrimoniului necesită soluționarea unui șir de probleme legate de recunoașterea, editarea, traducerea, interpretarea, circularea și recepționarea textelor tipărite atît în limba română cît și în alte limbi moderne. Soluționarea acestor probleme pentru patrimoniul istorico-lingvistic 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, în particular cîteva „alfabete de tranziție” chirilico-latine. Dificultățile în digitizarea și conservarea acestui tezaur țin de recunoașterea corectă a literelor chirilico-latine, dar și de inexistența unui lexicon adecvat perioadei de tipărire a resursei. O soluție pentru problema lexiconului ar fi alinierea la normele lingvistice contemporane ale textelor vechi (**1**).

Istoric, limba română a parcurs o cale lungă și bogată de dezvoltare. Există studii care explică apariția foneticii și ortografierii caracteristice etapelor concrete de evoluție a limbii, care sunt necesare atît pentru determinarea alfabetului, cît și a literelor specifice (**2,3**). Cunoașterea acestor legități ne permite să construim resurse lingvistice utilizînd un instrument special elaborat pentru o perioadă istorică concretă.

Prima carte tipărită pe teritoriul românesc a fost *Liturghierul slavon*, îngrijit de ăctre ieromonoahul Macarie în anul 1508, iar prima carte tipărită în limba română a fost *Catehismul Românesc* al diaconului Coresi, apărut la Brașov în anul 1535(**3**).

Biblioteca Națională a Republicii Moldova deține o colecție de aproximativ 21.000 cărți vechi și rare. Circa 20 de cărți din această colecție sunt tipărite în limba română, în Basarabia (Chișinău și Dubăsari), utilizînd alfabetele chirilic și tranzițional(**4,5**). Bibliotecile publice din Sankt Petersburg dețin importante mostre de carte româneacă veche (secolele XVI-XIX). Dintre cele 66 de titluri incluse, spe exemplu, în *Catalogul edițiilor chirilice ale slavilor de sud și ale românilor*, 45 de volume revin slavilor de sud, iar 21 de volume - țărilor românești(**6**).

Studiile existente explică aspectele legate de dezvoltarea componentelor principale ale limbii: alfabet, lexicon, ortografie cu referire la etapele specifice din evoluția limbii. Această informație este utilă pentru a crea resurse și instrumente lingvistice racordate la anumite perioade din istoria limbii. Ținînd cont de particularitățile 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 Socialită 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 adîncul secolelor”.

**Perioadele de evoluție a limbii române**

Istoria limbii române cunoaște două epoci în dezvoltarea sa. Prima se referă la formarea dialectului dacoromân, începînd cu căderea Sarmisegetuzei (106 A.D.) pînă în secolul al XV-lea(**2**). Se utiliza alfabetul chirilic grație influenței masive a Bisericii Ortodoxe.

Epoca a doua de dezvoltare a limbii române literare (secolul XVI-XX) începe cu apariția primelor texte scrise în limba română și constitue rezultatul unei îndelungate și complexe evoluții(**3**). Procesul de unificare lingvistică este marcat de apariția Bibliei de la București (1688), care a condus ulterior la stabilirea a două mari etape în evoluția lingvistică(**7**).

Etapa întîi începe cu apariția primelor texte literare românești și se încheie la începutul secolului al XVIII-lea. În cadrul acestei etape pot fi distinse 3 perioade:

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

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

* Anii 1715-1780, momentul primei unificări, aproximativ în 1750;
* Anii 1780-1836, diversificarea lingvistică;
* Anii 1836-1881, constituirea principalelor norme ale limbii literare de atăzi;
* Ani 1881-1960, definitivarea formării normelor limbii române literare contemporane.

Ultima perioadă ne descrie consolidarea stilurilor limbii române literare. În 1904, prin modificările aduse ortografiei, se stabilesc definitiv bazele scrierii fonetice. păstrate, cu unele retușări ulterioare, pînă în prezent. Voi arăta în fig. 1-8 exemple de texte tipărite în diverse perioade din evoluție a limbii române.

/\*

\* Figurile 1-8, de modelat în ps.

\*/

**Perioadele 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 cîteva litere grecești, în mare parte pentru redarea adecvată a numelor proprii, dar și litere originale românești. De exemplu litera /\*-în-\*/ utilizată pentru a reda prefixul (prepoziția) **în**, **îm**, sau litera **î** modernă la începutul cuvîntului. Acest alfabet a fost utilizat la tipărirea *Cazaniei* lui Varlaam la Iași în anul 1643 (fig. 1). Primul abecedar românesc a fost tipărit în 1699 la Belgrad (Alba-Iulia), iar prima gramatică românească a fost tipărită în 1757 de Dimitrie Eustatievici.

Începînd cu anul 1830 și pînă la adoptarea oficială a alfabetului român în 1862, nu exista un alfabet stabil, astfel în această perioadă au fost utilizate cel puțin șapte modificări ale așa-numitului „alfabet de tranziție”, chirilico-latin, care conținea atît litere latine, cît și litere chirilice (fig 4, 7). De exemplu, **e** - **e** (1830) - **з** (1846); **к** - k; **щ** - **шt**; s - **дз** - **dz** - /\*-d-\*/ (1846).

Utilizarea grafiei latine în România nu a influențat activitatea tipografică din Basarabia. După **alipirea** Basarabiei la Imperiul Rus în 1812, limba oficială la Chișinău devine 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 Chișinău tipărit cu caractere chirilice. Tipografia bisericească din Chișinău a fost sistată în perioada 1883-1890, procesul fiind reluat la începutul secolului XX.

Spre deosebire de alfabetul chirilic utilizat pentru scrierea limbii române din secolele XIV-XV pînă în anul 1862, alfabetul chirilic folosit în Republica Autonomă Sovietică Socialistă Moldovenească (RASSM) începînd cu anii 1930 și, ulterior în Republica Sovietică Socialistă Moldovenească (RSSM) și 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 Molfova alfabetul chirilic a fost utilizat pînă în 1989.

Voi prezenta mai jos (tab. 1) perioadele de evoluție a alfabetului român începînd cu *Cazania* de Varlaam. Pe lîngă alfabet, există și alți factori care caracterizează evoluția limbii, precum ortografia și lexiconul.

/\*

\* Tabelul 1 - Evoluția alfabetului român începînd cu anul 1642 (pag. 64)

\*/

**Recunoașterea textelor tipărite**

Procesul de digitizare și de recunoaștere 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 recunoaștere și lexicoane specializate. Procesul de digitizare și recunoaștere e constituit din următoarele etape (fig. 9):

* Digitizarea (scanarea) textului pentru obținerea copiei electronice grafice;
* Recunoașterea 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, așa-numitul proces de conversie. Transliterarea textului se va efectua ținînd cont de literele specifice utilizate în textul inițial.
* Verificarea textului recunoscut se produce utilizînd resursele lingvistice reutilizabile specializate pentru perioada respectivă.

/\*

\* Figura 9 - Etapele tehnologice de recunoaștere a textelor tipărite (pag. 65)

\*/

Digitizarea textelor constă în scanarea lor și obținerea variantei electronice în formă de imagine. Pentru recunoașterea textelor din imagine se aplică OCR. Sistemele standard OCR utilizează diferite metode de recunoaștere a textelor. Am cercetat posibilitățile a două sisteme: IRIS și ABBY FineReader. Rezultatele experiențelor de recunoaștere a textului tipărit în secolul al XIX-lea sunt expuse în continuare. Am determinat că sistemul IRIS, în procesul de instruire, nu poate selecta orice fragment din imagine textului și de aceaste acest sistem nu satisface scopul de recunoaștere a textului tipărit cu scrisul vechi român. Așadar, sistemul IRIS nu a mai fost utilizat în scop practic, ci doar ca un model comparativ (**de referință**).

În continuare se vor folosi tehnici de recunoaștere a formelor pentru identificarea individuală a caracterelor unei pagini de text, inclusiv și semnele de punctuație, pațiile și sfîrșitul de linie. Textul recunoscut se va prezenta în final ca un fișier editabil.

Transliterarea este un proces strict individual ce depinde de perioada examinată. În funcție de textul inițial, se vor utiliza programe care conțin informație despre caracterele specifice întîlnite în text. Transliterarea presupune stabilirea une relații bidirecționale univoce între două sisteme de scriere astfel, încît un cunoscător să poată reconstitui textul original din varianta transliterată.

Verificarea textului se efectuează cu aplicații special elaborate (**9**), care utilizează resursele reutilizabile specifice pentru perioada istorică a textului tipărit. Totodată, cuvintele noi obținute se vor introduce în lexiconul corespunzător.

**Procesarea textelor tipărite cu alfabet chirilic în RASSM și RSSM**

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

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

Mă voi referi în continuare le 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 recunoaște litera adițională **ӂ** și pentru a crea lexiconul respectiv. Existența unui lexicon caracteristic acestei perioade ar permite automatizarea procesului de verificare și validare a cuvintelor recunoscute prin procedeele expuse mai sus. Acest lexicon poate fi creat:

* manual,
* prin transliterearea cuvintelor românești scrise cu caractere latine în varianta corectă scrisă cu caractere chirilice;
* prin alinierea variantelor de text tipărite în paralel cu caractere latine și caractere chirilice.

Prin transliterare vom înțelege transcrierea unui cuvînt din limba română în forma echivalentă scrisă cu caractere chirilice și conform normelor lingvistice acceptate în perioada 1967-1989 în RSSM. **/\*-8.4.16-\*/** Metoda transliterării s-ar potrivi ideal în cazul dacă se reușește formalizarea tuturor regulilor de transcriere. Un studiu prealabil arată că acest proces este anevoios și nu poate fi automatizat în totalitate din cauza iregularităților legate de discordanța dintre fonologia, morfologia și sintaxa limbii române și normele lingvistice acceptate în RSSM. Parțial acest proces poate fi automatizat implicînd elemente de formalizare a regulilor de transcriere, de intervenție manuală și de aliniere.

Dificultăți evidente apar la transliterarea cuvintelor de proveniență străină. Dacă în limba română aceste cuvinte se scriu, de regulă, ca și în limba originală, atunci transcrierea lor cu caractere chirilice se face conform pronunțării. 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 parțial automatizat. În acest scop s-au stabilit reguli de transcriere a literelor și îmbinărilor de litere. Exemplu de aceste reguli sunt prezentate mai jos:

* Reguli de transcriere „literă → literă”. De exemplu, ***a****→****а****,* ***ă****→****э****,* ***b****→****б****,* ***d****→****д****,* ***f****→****ф****,* ***l****→****л****,* ***m****→****м****,* ***n****→****н****,* ***r****→****р****,* ***ș****→****ш****,* ***t****→****т****,* ***ț****→****ц****,* ***v****→****в****,* ***z****→****з*** *(****bardă - бардэ****,* ***zarvă - зарвэ****,* ***măr - мэр****).*
* Reguli de transcriere pentru literele ***î*** și ***â***. Îmbinările ***âi*** sau ***îi*** sw vor transcrie în ***ы*** pentru cuvintele ***mâine, pâine, câine*** și derivatele lor (***mâine - мыне, pâine - пыне, câine - кыне, mîine - мыне, pîine - пыне, cîine - кыне***). În alte situații se va aplica regula „literă → literă”: ***â****→****ы, î****→****ы*** (***român*** *→* ***ромын, întâi*** *→* ***ынтый***).