//22.4.16 16:57 v1.7

Universitatea Academiei de Științe din Republica Moldova

Facultatea de Științe Exacte

Specializarea: Informatica

LUCRARE DE LICENȚĂ

Coordonator științific, Absolvent,

Gladei Anatol Ungur Ștefan

2016

Universitatea Academiei de Științe din Republica Moldova

Facultatea de Științe Exacte

Specializarea: Informatica

Algoritmizarea procesului de transliterare pentru actualizarea textelor chirilice românești

Coordonator științific, Absolvent,

Gladei Anatol Ungur Ștefan

2016

**Cuprins**

1. Introducere

2. Principii generale, introducere în procesul de transliterare

2.1. Scurt istoric

2.2. Principii generale ale transliterării

2.3. Exemplu practic de transliterare lingvistică (din engleză în română)

2.4. Aplicații practice ale transliterării

2.5. Evaluarea metodei practice

2.6. Digitizarea, recunoașterea și conservarea patrimoniului cultural istoric

2.7. Perioadele de evoluție a limbii române

2.8. Perioadele de evoluție al alfabetului român

2.9. Recunoașterea textelor tipărite

2.10. Procesarea textelor tipărite cu alfabet chirilic

2.11. Procesarea textelor tipărite cu alfabet latini și litere adiționale

2.12. Procesarea textelor tipărite cu alfabete tranziționale

2.13. **Procesarea textelor tipărite cu alfabetul slavon (sec. XVIII - XIX)**

3. Aplicația practică

3.1. Caracteristici generale ale aplicației

3.2. Interfața grafică

3.3. Tehnologii utilizate

3.4. Descrierea algoritmului (algoritmelor), regulilor și excepțiilor

3.5. Exemplu practic de text transliterat

3.6. Dezvoltarea aplicației pe viitor

3.8. Împachetarea și distribuția

4. Concluzie

**1. 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 când 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 transliterării*(conversiei)* dintr-o grafie în alta.

Pe parcursul lucrării voi pune accent atît pe importanța realizării unui astfel de algoritm cât 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 greșeli 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, utilizând tot pachetul de instrumente și funcționalități propuse (dicționarul de cuvinte, lista de prefixe, posibilitatea alegerii standardului necesar etc.). Atât cum aplicația necesită să aibă o interfață grafică accesibilă și ușor de înțeles, iar pachetul de instrumente să fie la fel ușor accesibil, o aplicație de tipul dat nu se rezumă doar la grafică sau design. Orice utilizator pune în primul rând calitatea conținutului, iar calitatea unei aplicații se poate determina atât pe baza aspectelor vizuale cât și pe cele tehnice. Voi face un studiu comparativ între mai multe aplicații de translare(*convertare*) atât de realizare străină cât și a celor românești. Analiza succintă se va evalua atât la nivel tehnic cât 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 ținând 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 atât 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, atât desktop cât și web, la început nu este altceva decât un algoritm de procesare a datelor de intrare și obținerea datelor de ieșire necesare, iar fiecare algoritm la început se elaborează, testează, cizelează, pentru ca în final să fie implementat în aplicația dată. Interactivitatea aplicației se bazează pe pe aceste tehnologii, iar limitele impuse la nivel de implementare trebuie luate în considerare în conceperea unei aplicații. Mai mult decât atât, este important de luat în considerare și grupul țintă de utilizatori căruia se adresează aplicația. Chiar și dacă aplicația nu este îndreptată spre un public larg de utilizatori, oricum trebuie de ținut cont de necesitățile și posibilitățile lor (vârsta, profesia etc.), cu scopul de a realiza interfața grafică și pachetul de instrumente cât 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 încât să corespundă necesităților utilizatorilor.

Principii generale ale gui-ului

În general, fiecare aplicație indiferent de tipul ei trebuie să fie unică, deci să aibă elemente distincte atât 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(instituțională, educativă, de prezentare a unui produs etc.), ce instrumente 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 unei 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.

/\*

\* handwritten,

\* to be continued

\*

\*/

**2. Principii generale, introducere în procesul de transliterare**

**2.1. Scurt istoric**

De-a lungul anilor au fost propuse câteva tehnici de transliterare între două limbi, fiind orientate, în principal, pe transliterarea ortografică a numelor proprii englezeşti în chineză, japoneză, coreeană sau arabă.

Knight şi Graehl, în 1997, au introdus o metodă de transliterare între japoneză şi engleză, utilizând algoritmi de traducere bazaţi pe maşini cu stări finite, această metodă fiind adaptată de Stalls şi Knight în anul succesor pentru transliterare bidirecţională între engleză şi arabă. Alte metode de transliterare au mai fost descrise de Jung et al. (2000), Meng et al. (2001), Virga şi Khudanpur (2003). În lucrarea lor, Haizhou et al. (2004) clasifică metodele menţionate mai sus ca fiind abordări ale transliterării bazate pe nivel fonetic. Ei propun o nouă tehnică, numită de către autori mapare-ortografică-directă (direct orthographic mapping sau DOM), cu alte cuvinte un model de transliterare pe bază de n-grame (secvenţe de n litere consecutive care pot apărea în cuvintele unei limbi).

Experimentele realizate au fost focalizate asupra transliterării textelor românești tipărite cu grafie chirilică în cea română, fiind parte a studiului efectuat pentru modulul dedicat transliterării automatizate, integrat într-un sistem de sinteză a vorbirii pentru limba română.

Transliterarea nu este altceva decât reprezentarea caracterelor unei scrieri alfabetice sau silabice în caracterele alfabetului în care se face conversia. Transliterarea se face în principiu caracter cu caracter și redă grafia (nu pronunțarea) cuvintelor. Se utilizează în actele notariale, de stare civilă, în comunicațiile poștale, precum și în lucrări de istorie, geografie, cartografie, bibliografie, lingvistică etc.

Transliterarea se face după sistemele oficiale ale limbilor supuse transliterării sau după norme naționale sau internaționale de conversie. Ea nu depinde de normele ortografice ale limbii în care se transliterează. Standardele internaționale pentru conversia sistemelor de scriere urmăresc asigurarea unui sistem de convertire riguros, univoc și complet reversibil, fără ambiguități. Ele trebuie să permită schimbul internațional de informații prin comunicarea cu mijloace manuale, mecanice sau electronice a mesajelor scrise între oameni sau mașini, care să le poată transmite și reconstitui automat. Din acest motiv se face abstracție de considerente fonetice sau estetice și de uzanțele naționale. Reprezentările ce rezultată exclusiv în baza transliterării nu pot fi considerate mereu exacte conform uzanțelor fonetice ale limbii supuse conversiei. Totuși, cititorul care cunoaște limba convertită poate reconstitui în mod neechivoc grafia și, în anumite limite, pronunția originală.

Transliterarea respectă obiectul tradiției și convențiilor locale. De exemplu, în tradiția limbii române litera chirilică х se transliterează prin litera latină h, dar în tradiția țărilor anglofone aceeași literă se transliterează prin digrama kh. Astfel, în română toponimul Сахалин se transliterează Sahalin, iar în engleză ca Sakhalin. De asemenea, limba română redă ش arab, שׁ ebraic sau ш chirilic prin "ș", astfel că denumiri precum precum שבועות ,مراكش sau Шумeн, se transcriu respectiv în litere latine prin Șauia, Șavuot sau Șumen, pe când în țările anglofone sau francofone apar sub forme cu digrame: Shawia, Shavuot sau Shumen în engleză, respectiv Chaouia, Chavouot sau Choumen în franceză.

Desigur, în limba română, este preferabilă folosirea transliterării românești. Când este vorba, în schimb, de nume proprii străine deja scrise în litere latine în limbile țărilor respective, nume care așadar nu necesită să fie transliterate, tradiția este preluarea lor ca atare: Ouagadougou, Saskatchewan, Szczecin, cu indicarea pronunțării transcrisă între paranteze (Oagadugu, Sascatceoan, Șcețin). Aceste tradiții au fost în general respectate în cărțile și publicațiile apărute în România până la începutul deceniului 1980.

Cazuri particulare sunt:

* cele ale limbilor asiatice precum chineza, japoneza, coreeana sau limbile peninsulei indiene, pentru care se folosesc, atât în română cât și în majoritatea limbilor folosind scrierea romanică, sisteme de transliterare specifice (de exemplu sistemul Pinyin pentru chineză);
* cel simplificat al limbii bulgare;
* cele două sisteme de transliterare folosite de-o parte și de alta a Prutului pentru denumirile, toponimele și patronimele din limbile rusă sau ucraineană: К, Ы, și Я se transliterează în România prin K, Â și IA (în dependență de conținut), iar în Republica Moldova prin C (CH înainte de E sau I), Î și EA sau IA (la fel în dependență de conținut).

Transliterarea și transcrierea se aseamănă, dar trebuie confundate. Prin transcriere se reprezintă sunetele vorbirii într-un anumit sistem de scriere, fără a ține seama de modul de notare a sunetelor în sistemul de scriere propriu al limbii respective, deci transcrierea este mai degrabă o metodă de notare fonetică (este metoda cea mai folosită în limbile folosind alfabetul chirilic). Transliterarea în schimb face conversia unui sistem de scriere în altul, nefiind condiționată de aspectul fonetic.

**2.2. Principii generale ale transliterării**

Transliterarea în prelucrarea limbajului natural a fost introdusă pentru a translata numele proprii dintr-o limbă în alta în situaţiile în care cele două limbi folosesc un inventar fonetic incompatibil sau o ortografie total diferita. În această lucrare, propun o metodă pentru translatarea din grafia chirilică și slavonă în cea latină (dar care, ulterior, poate fi adaptată pentru orice altă pereche de limbi sau grafii), și prezint o aplicaţie care utilizează transliterarea în sinteza vorbirii, astfel automatizînd procesul nemijlocit de conversie.

În traducerea automată (MT – machine translation) sunt multe situaţii în care întâlnim nume proprii ai căror echivalenţi de traducere nu sunt cunoscuţi. În cazul în care ortografia celor două limbi este asemănătoare, o practică frecventă este ca aceste cuvinte să rămână neschimbate. Acest lucru nu este posibil însă, dacă una dintre limbi foloseşte o ortografie total diferită de cealaltă (de exemplu, traducerea din engleză într-una din limbile arabă, japoneză, rusă, chineză sau bulgară).

Sistemele de sinteză a vorbirii pornind de la text au sarcina de a sintetiza vocea pornind de la un text arbitrar, se confruntă cu următoarea problemă: pentru textele care conţin cuvinte sau nume proprii provenind din alte limbi nu se poate aplica direct transcrierea fonetică folosind aceleaşi reguli de transcriere specificate manual sau învăţate automat pentru limba pe care a fost proiectat sistemul. O soluţie pentru rezolvarea acestei probleme este introducerea unor pachete suplimentare de reguli în vederea obţinerii transcrierii fonetice din diferite limbi sursă. Însă nu toate limbile au acelaşi pachet fonetic, iar lexicoanele străine necesită adaptări pentru a fi corelate cu limba sau grafia ţintă.

În cadrul acestei lucrări, propun o abordare diferită a transliterării, în care se utilizează un transliterator pentru adaptarea de la o limbă la alta înainte de a aplica transcrierea fonetică. Aceasta din urmă se obţine folosind acelaşi pachet de reguli utilizat pentru cuvintele native (ale grafiei ţintă). În cazul limbii române, datorită ortografiei preponderent fonetice, este necesar ca acurateţea obţinută prin transliterare, urmată de transcrierea fonetică aplicată pentru grafia latină să fie comparabilă cu acurateţea obţinută în situaţia în care s-ar aplica reguli de transcriere fonetică direct pentru chirilică, urmând să se facă ulterior o adaptare la nivel fonetic între cele două grafii. Practic, acurateţea globală este limitată în ambele cazuri de performanţele sistemelor de transcriere fonetică.

**2.3. Exemplu practic de transliterare lingvistică (din engleză în română)**

În cadrul lucrării mi-am propus că realizez în mod practic, transliterarea atât dintr-o grafie în alta, cât și dintr-o limbă în alta. Spre a ușura condiția experimentului am considerat limba-sursă ca fiind cea engleză, deoarece face parte din aceiași familie lingvistică latină ca și limba română, și deoarece materiale legate de conversia din limba engleză în cea română sunt foarte multe și ușor de accesat. Așadar, algoritmul de transliterare din engleză în română va necesita un corpus de antrenare format din cuvinte scrise în limba sursă (în cazul dat engleză) asociate cu transliterările lor corespunzătoare în limba ţintă (în cazul dat limba română).

Faptul că limba română are o ortografie preponderent fonetică a permis să folosesc transcrierea fonetică a cuvintelor englezeşti ca pivot pentru tehnica semi-automată pe care am folosit-o în crearea corpusului de antrenare. Am ales ca lexicon de plecare pentru transcriere fonetică în limba engleză CMUDict. Pe lângă cuvintele uzuale, acest lexicon conţine un număr mare de nume proprii, abrevieri şi cuvinte adaptate la limba engleză, care sunt provenite din arabă, germană, franceză, poloneză, precum:

* Italiană: braggiotti, castelli, castelluccio
* Germană: aachen, abbenhaus, schlender, schlenker
* Poloneză: zawistowski.

Aceste cuvinte ar încurca procesul de transliterare deoarece conversia lor în foneme nu se poate obţine cu ajutorul regulilor standard. Soluţia pentru a învăţa doar reguli omogene a constat în filtrarea CMUDict prin alegerea unui set de cuvinte uzuale din limba engleză pe baza cărora a fost generat lexiconul de transliterare. Folosind apoi datele din CMUDict, a avut loc generarea transcrierei fonetice pentru aceste cuvinte şi am folosit un set de reguli pentru a trece de la fonemele limbii engleze la litere şi/sau grupurile de litere din alfabetul românesc (mai detaliat descris în Tabelul 1).

Pentru limbile fără ortografie fonetică sunt necesari doi paşi suplimentari faţă de cei prezentaţi anterior. Primul pas constă în maparea dintre fonemele specifice pentru limba sursă şi fonemele care apar în limba destinaţie. Al doilea pas implică trecerea din forma fonetică a cuvintelor înapoi la forma ortografică a acestora, de data aceasta folosind un pachet de reguli specific pentru limba ţintă. Ultimul pas se poate realiza folosind metode automate, dar, pentru rezultate bune, este necesară o iteraţie suplimentară care constă în validarea manuală a rezultatelor obţinute automat.

Tabelul 1. Reguli de conversie

|  |  |  |  |
| --- | --- | --- | --- |
| ***Engleză*** | ***Cuvînt*** | ***Transcriere fonetică*** | ***Transliterație*** |
| AA | odd | **AA** D | **a**d |
| AE | at | **AE** T | **e**t |
| AH | hut | HH **AH** T | h**a**t |
| AO | ought | **AO** T | **o**t |
| AW | cow | K **AW** | c**au** |
| AY | hide | HH **AY** D | h**ai**d |
| B | be | **B** IY | **b**i |
| CH | cheese | **CH** IY Z | **ci**z |
| D | dee | **D** IY | **d**i |
| DH | thee | **DH** IY | **z**i |
| EH | Ed | **EH** D | **e**d |
| ER | hurt | HH **ER** T | h**ăr**t |
| EY | ate | **EY** T | **ei**t |
| F | fee | **F** IY | **f**i |
| G | green | **G** R IY N | **g**rin |
| HH | he | **HH** IY | **h**i |
| IH | it | **IH** T | **i**t |
| IY | eat | **IY** T | **i**t |
| JH | gee | **JH** IY | **g**i |
| K+(E/I) | key | **K** IY | **ch**i |
| K | call | **K** AO L | **c**ol |
| L | lee | **L** IY | **l**i |
| M | me | **M** IY | **m**i |
| N | knee | **N** IY | **n**i |
| NG | ping | P IH **NG** | pi**ng** |
| OW | oat | **OW** T | **ă**ut |
| OY | toy | T **OY** | t**o**i |
| P | pee | **P** IY | **p**i |
| R | read | **R** IY D | **r**id |
| S | sea | **S** IY | **s**i |
| SH | she | **SH** IY | **ș**i |
| T | tea | **T** IY | **t**i |
| TH | theta | **TH** EY T AH | **t**eta |
| UH | hood | HH **UH** D | h**u**d |
| UW | two | T **UW** | t**u** |
| V | vee | **V** IY | **v**i |
| W | we | **W** IY | **u**i |
| Y | yield | **Y** IY L D | **i**ild |
| Z | zee | **Z** IY | **z**i |
| ZH | seizure | S IY **ZH** ER | si**j**ăr |

Trebuie menţionat că întregul proces de transliterare se face cu pierdere de informaţie atât din cauza metodelor statistice folosite cât şi pentru că nu toate fonemele au echivalent direct în limba română şi, pe baza contextului, unele litere din alfabetul român pot avea o pronunţie diferită faţă de cea dorită. Pentru cazul dat acest efect nu constituie o problemă deoarece cuvintele sună „natural” pentru un vorbitor nativ de limba română în momentul în care sunt sintetizate. În cazul căutării pe bază de percepţie, pentru a diminua efectele nedorite generate de această pierdere de informaţie, a fost folosită transliterarea „forward-transliteration” în locul „backward-transliteration” Diferența dintre aceste două metode se rezumă la faptul că prima implică inițial convertarea din engleză în română, iar a doua invers, și anume din română în engleză. Deşi căutarea se face folosind cuvinte scrise în română (ceea ce ar implica „backward-transliteration”), probleme la pasul dat nu au fost întâlnite.

Problema transliterării poate fi reformulată în modul următor: realizarea unui set de reguli, care pornind de la un şir de simboluri sau caractere ce aparţin alfabetului sursă (cuvintele ce trebuie transliterate nemijlocit), rezultă în obţinerea un şir de simboluri (caractere) ce aparţin alfabetulului destinaţie, astfel încât aplicând setul de reguli pentru transcrierea fonetică specifică fiecărei limbi pe ambele şiruri, atât cel de intrare cât şi cel de ieşire, asemănarea între sunetele obţinute să fie maximă. Metodele bazate pe manipulare ortografică directă au o acurateţe mai bună decât cele din prima categorie, ceea ce m-a determinat să aleg o abordare de tip DOM.

Transcrierea fonetică este o problemă asemănătoare cu cea a transliterării propriu zise. Diferenţa dintre aceste două se formulează în modul următor: în cazul transcrierii fonetice se caută și se formulează un pachet de reguli pentru a trece din literele unui cuvânt anumit în simbolurile folosite pentru reprezentarea fonetică a acestuia, pe când transliterarea implică căutarea unei mapări către ortografia limbei țintă. Așadar metoda dată foloseşte un clasificator de tip-ul Maximum Entropy, care parcurge și asociază etichete fiecărui caracter dintr-un cuvânt pe baza unei serii de trăsături extrase din contextul lexical al literei respective.

Secvenţa de etichete, sau cu alte cuvite pachetul de reguli obţinut pentru o secvenţă de caractere a unui cuvânt constituie transcrierea fonetică a acestui cuvânt. Dacă notăm cu l litera curentă, cu li litera aflată la distanţa respectivă de lungime i faţă de litera curentă, şi notăm cu p-1 eticheta anterioară, atunci obținem o legitate a trăsăturilor, fomulate în modul următor:

* l-1, l – litera curentă și litera anterioară;
* l-2, l-1, l – litera curentă și două litere anterioare;
* l, l+1 – litera curentă și litera succesoare;
* l, l+1, l+2 – litera curentă și două litere succesoare;
* l-1, l, l+1 – litera curentă împreună cu literele imediat învecinate;
* p-1 – eticheta emisă pentru litera predecesoare.

Acurateţea medie obţinută cu ajutorul unei astfel de combinaţii de trăsături este în jur de 90% pentru cuvintele din afara vocabularului ce sunt transcrise corect fonetic în limba română şi în jurul 60% pentru cuvintele luate din CMUDict (deci din afara vocabularului), ce sunt transcrise corect fonetic în limba sursă (în cazul dat engleza). Trebuie neapărat menţionat faptul că acurateţea de 60% pentru CMUDict este destul de bună în comparaţie cu alte metode propuse de alţi autori pentru transcrierea fonetică. Acele metode au o acurateţe ce poate varia între valorile 57% şi 62%. Performanţa maximală obținută pentru acest lexicon este de 71%, și a fost obţinută folosind algoritmul Marginal Infused Relaxed Algorithm.

Așadar, din cele enumerate mai sus, am luat decizia de a utiliza aceeaşi metodă de transliterare şi pentru convertarea textelor din limba engleză în română, antrenând astfel acelaşi clasificator, dar de data aceasta pe baza unui lexicon de transliterare. Ținând cont de trăsăturile menţionate, clasificatorul atribuie fiecărei litere din context un etichet ce reprezintă un simbol nemijlocit, o mulțiem de simboluri din alfabetul destinaţie, sau mulţimea vidă.

Etichetele sunt obţinute în acest mod, utilizând lexiconul de antrenare și folosind alinieri automate ale unor litere din cuvintele alfabetului sursă cu literele corespunzătoare din alfabetul destinaţie, au determinat un nivel înalt de precizie al algoritmului obținut.

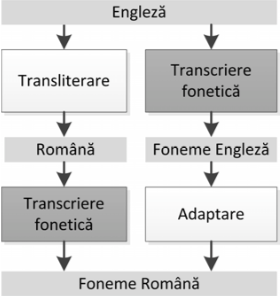
Astfel am cercetat și evaluat transliterarea din limba engleză în română, folosind o metodă numită ten-fold. Setul de date utilizat la antrenare a fost împărțit în 10 submulţimi de mărime egală, şi a fost testată acurateţea sistemului obținut pe toate submulţimeile în parte, astfel antrenând sistemul pe celelalte 9 şi calculând procentul de acurateţe la nivelul de cuvânt. Rata acurateții a fost calculată în modul următor: numărul de cuvinte transliterate în mod corect fiind raportat la numărul total de cuvinte transliterate. Chiar și dacă un cuvânt transliterat are o literă greşită, atunci acesta se atribue la numărul de erori. Acurateţea sistemului dat, ca o medie a acelor 10 validări, a fost de 78%. Trebuie neapărat de menţionat că rezultatul a fost raportat la cuvintele din afara vocabularului utilizar la antrenare. Din motiv că limba română are o ortografie preponderent fonetică s-ar fi putut aştepta ca rata acurateţii în cazul dat să fie nemijlocit asemănătoare cu rezultatul obţinut la transcrierea fonetică. Creşterea cu 9% al acurateţii în cazul dat al transliterării se datorează faptului că antrenarea a fost făcută pe cuvintele doar în limba engleză (fără oricare cuvinte străine sau alte abrevieri ce se conțin în CMUDict). Mai mult ca atât, fiind vorba de o oarecare mapare ortografică dintre cele două limbi, a fost redus numărul etichetelor care se atribuiau unei litere sau unui grup de litere direct în momentul clasificării.

**2.4. Aplicații practice ale transliterării**

Aşa cum a mai fost menţionat mai sus, se întâlnesc mai multe situaţii în care textul ce trebuie sintetizat conţine mai multe cuvinte provenite din alte limbi, care evident nu pot fi procesate corect dacă vom utiliza acelaşi pachet de reguli ca și la cuvintele specifice limbii ţintă pentru care este destinat sistemul (în cazul dat limba română). În situaţia dată, o abordare evidentă va fi folosirea unor seturi diferite de reguli de transcriere fonetică pentru aceste cuvinte. Aceste reguli vor fi specifice limbii din care ele fac parte, totodată fiind necesară şi o adaptare ulterioară la nivelul fonetic. În cadrul acestei lucrări, propun o abordare puțin diferită, ce constă în folosirea unui transliterator pentru generarea “pseudo-cuvintelor” native. Așadar, transcrierea fonetică se va face folosind reguli specifice limbii române şi nu va mai fi necesară adaptarea rezultatelor la nivel fonetic. Totuși există unele situaţii în care, fiind dat un context lexical, unele litere din limba română vor generea sunete diferite de cele care au fost intenţionate. Cu toate acestea, cuvintele vor suna cât se poate de clar pentru orice vorbitor nativ de limba română şi nu vor prezinta dificutăţi în ceea ce priveşte inteligibilitatea. Diferenţa între aceste două abordări majore este subliniată în Figura 1 de ma jos:

* metoda bazată pe manipularea fonetică se aplică atunci când sunt cunoscute seturile de reguli pentru transcrierea fonetică (conversia din litere în sunete) pentru ambele limbi implicate în proces;
* metoda bazată pe manipularea ortografică nu necesită cunoaşterea setului de reguli folosite pentru transcrierea fonetică a limbei sursă, ci doar pentru limba destinaţie.

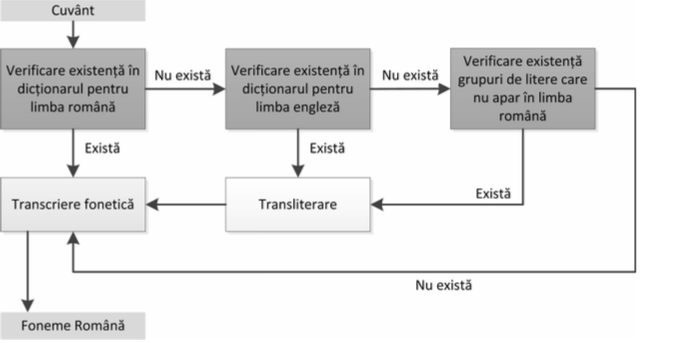
Figura 1. Conversia cuvintelor scrise în engleză pentru sinteza vorbirii



Dificultatea cu care se confruntă ambele metode pentru procesarea cuvintelor străine este alegerea corectă a cazurilor în care sistemul trebuie să aplice transliterarea cuvintelor din limba sursă în limba țintă. Principala modalitate de a verifica dacă un cuvânt trebuie sau nu transliterat este folosirea lexiconului care a fost generat anterior. Este evident că dacă un cuvânt se regăseşte în tabela de transliterare atunci el este şi un candidat bun pentru procesul dat. Totuși este necesar de verificat dacă acesta există şi în inventarul de cuvinte al limbii țintă. Spre exemplu, nu putem spune cu exactitate dacă transliterearea este necesară pentru un cuvânt de tipul “minus” sau “bat”, deoarece ele au aceeaşi semantică şi ortografie atât în limba română cât şi în cea engleză. Acest tip specific de cuvinte vor face parte din lista excepțiilor, și preferabil stocate într-un oarecare fişier separat, şi ca rezultat vor rămâne neschimbate. În cazurile în care un cuvânt nu se va regăsi nici în inventarul de cuvinte cunoscute pentru limba sursă şi nici în cel pentru limba țintă, va fi greu de precizat dacă acesta trebuie neapărat transliterat sau nu.

O metodă clasică de a rezolva aceste situaţii ar fi folosirea unor indicii lexicali pentru a decide care este limba din care provine cuvântul dat. Spre exemplu, dacă luăm limba sursă ca rea română atunci există unele anumite grupuri de caractere ce sunt foarte rare sau chiar nu pot exista în limba română. Deci toate cuvintele care conţin litera ‘y’ sau grupul de litere “ck” vor candidaţi buni pentru transliterare. Astfel, putem genera o listă de perechi de câte trei sau mai multe litere consecutive posibile în limba română, folosind DEX-ul. Atunci când sistemul va întâlni un cuvânt necunoscut (pentru ambele limbi procesate), va avea loc testarea dacă fiecare grup de litere consecutive din cuvânt se regăseşte în lista generată anterior. Orice cuvânt va care conţine o combinaţie de litere care nu este specifică pentru limba română va fi transliterat automat. Celelalte cuvinte vor fi lăsate neschimbate. Schemaeste descrisă în Figura 2.

Figura 2. Detectarea cuvintelor trebuie transliterate



/\*

\*

\* Exepmlu din Lucr\_Translit\_Engleza\_Romana (pag.10)

\*

\*/

Aşa cum a fost menţionat mai sus, căutarea pe bază de percepţie reprezintă o modalitate de a găsi scrierea corectă a unui cuvânt dintr-o limbă sursă (de regulă limbă străină, precum cea rusă engleză, germană, franceză etc.) în funcţie de modul în care acest cuvânt este perceput (modul în care “sună” cuvântul) pentru un vorbitor nativ, în cazul dat de limba română. Spre exemplu, presupunem că ştim nimic altceva despre un oarecare oraş cu excepţia faptului că denumirea lui sună în modul următor: „ianţiau”. Așadar, nu avem nici o informaţie cu privire la ţara în care se află orașul dat sau vreo altă informaţie despre ortografia pe care ar trebui să o folosim pentru a găsi mai multe date despre locaţie. În acest caz putem folosi căutarea pe bază de percepţie pentru a obţine scrierea exactă a denumirii locaţiei pur și simplu prin tastarea cuvântului aşa cum este el perceput în limba nativă. Deci un vorbitor nativ de limba română ar introduce cuvântul respectiv „ianţiau”, care reprezintă cea mai apropiată formă ortografică din limba sa, iar rezultatul ar fi “燕郊” – o localitate ce se află la nordul Chinei.

Această metodologia are și câteva neajunsuri:

* există mai multe cazuri în care unele scrieri diferite se pronunţă la fel (omofone);
* părerea unei anumite persoane care nu este vorbitoare nativă de limba sursă despre cum ar trebui să fie scris cuvântul în limba sa nativă nu poate fi absolut exactă, deoarece nu toate limbile au acelaşi inventar fonetic iar regulile de conversie de la forma ortografică la cea fonetică sunt destul de complexe în unele situaţii.
* Aşa cum a fost menţionat de Knight şi Graehl (1997) transliterarea inversă (backward-transliteration) nu are aceeaşi flexibilitate ca şi transliterarea directă (forward-transliteration), deoarece pierderea de informaţie este de două ori mai mare atunci când se transliterează înainte şi înapoi între acele două limbi.

Când folosim căutarea bazată pe percepţie af fi bine de a antrena sistemul să translitereze între limba nativă (limba sursă în care nemijlocit s-a efectuat căutarea) şi toate limbile ţintă, alegând astfel cea mai bună variantă pe baza unei funcţii de similaritate între şiruri de caractere. Totodată, această metodă este predispusă și unei serii de erori, precum ar fi faptul că o reprezentare fonetică "percepută" a unui cuvânt poate corespunde mai multor forme ortografice şi, desigur, pierderea informaţiei din cauza incompatibilităţii pachetelor fonetice ale celor două limbi.

Pentru a compensa astfel de erori am propus o abordare puțin diferită, şi anume: toate cuvintele din limba ţintă să fie transliterate în limba nativă de căutare. Când se efectuează căutarea, comparăm transliteraţia curentă (dată de utilizator) cu toate transliteraţiile din baza de date folosind distanţa Levenshtein.

**/\*-22.4.16-stopped-here-\*/**

**2.5. Evaluarea metodei practice de transliterare**

Pentru a valida metoda propusă de căutare am creat alt corpus de test, compus doar din nume de oraşe din Statele Unite ale Americii. Corpusul conţine 480 de intrări selectate la întâmplare (nu are nimic în comun cu corpusul de transliterare - aşa cum a fost menţionat anterior – niciun nume propriu nu a fost păstrat în CMUDict). Alegerea noastră s-a bazat pe faptul că o astfel de metodă de căutare şi-ar găsi foarte uşor locul într-un sistem de navigaţie, sau asistent de călătorie.

După ce am selectat aceste nume, am folosit Google Speech API pentru a sintetiza fiecare cuvânt şi am rugat un număr de 5 persoane să asculte înregistările şi să scrie cuvintele în română aşa cum le aud. Fiecare persoană a putut să asculte acelaşi cuvânt de cel mult 3 ori.

Cuvintele din corpusul de test nou creat au fost procesate corespunzător metodologiei de căutare pe bază de percepţie prezentată anterior. La calcularea acurateţii sistemului s-a obţinut 99.38% (doar 3 cuvinte nu au fost identificate corect).

În articolul de față am prezentat o metodă de transliterare între engleză şi română, care, cu câteva adaptări specifice, poate fi aplicată şi pe alte perechi de limbi. Am creat un corpus de antrenare pentru transliterare ce poate fi obţinut în mod semiautomat (fără efort pentru limbile cu o ortografie fonetică) şi toate instrumentele sunt disponibile pentru descărcare pe pagina noastră web.Acurateţea de 78% a transliteraţiei TTS este raportată la cuvinte din afara vocabularului (out-of-vocabulary – OOV). În practică, nu toate cuvintele străine sunt necunoscute şi, chiar dacă apar erori de transliterare pentru unele cuvinte OOV, ele sunt de preferat în sinteza vorbirii în defavoarea formei lor directe. Ca parte din dezvoltarea sistemului nostru TTS românesc, intenţionam să extindem lexiconul de transliterare la Franceză şi Germană.Rezultatul obţinut prin căutarea pe bază de percepţie arată că motoarele de căutare şi asistenţii de călătorie ar beneficia de pe urma unui asemenea instrument. Căutarea după percepţie poate îmbunătăţi experienţa utilizatorului de internet, iar, în acelaşi timp, concentrarea pe corectarea greşelilor de ortografie bazată pe similitudini fonetice (care la un anumit nivel poate fi legată de transliterare) între cuvinte poate îmbunătăţi procesul de corectare ortografică (Li et al., 2006).

**2.6. Digitizarea, recunoașterea și conservarea patrimoniului cultural istoric**

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 atât în limba română cât 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 câteva „alfabete de tranziție” chirilic-latine. Dificultățile în digitizarea și conservarea acestui tezaur țin de recunoașterea corectă a literelor chirilic-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 atât pentru determinarea alfabetului, cât și a literelor specifice (**2,3**). Cunoașterea acestor legități ne permite să construim resurse lingvistice utilizând un instrument 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 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), utilizând alfabetele chirilic și tranzițional(**4,5**). Bibliotecile publice din Sankt Petersburg dețin importante mostre de carte românească veche (secolele XVI-XIX). Dintre cele 66 de titluri incluse, spre 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. Ținând 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 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”.

**2.7. Perioadele de evoluție a limbii române**

Istoria limbii române cunoaște 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 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 constituie 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 întâi 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-ardeleană);
* 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, până 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.

\*

\*/

**2.8. 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 câteva litere grecești, în mare parte pentru redarea adecvată a numelor proprii, dar și litere originale românești. De exemplu litera ꙟ 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 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.

Începând cu anul 1830 și până 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”, chirilic-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 | |  |
| **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 și cel slavon bisericesc; ocazional alfabet tranzițional 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) | |
| 1960 – 1993 (alfabet latin modificat) | 1944 – 1989 (alfabet chirilic bazat pe alfabetul rus; apare litera **ӂ**) | |
| 1993 – prezent (alfabet modern român bazat pe alfabetul latin) | 1989 – prezent (alfabet modern român bazat pe alfabetul latin) | |

**2.9. 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 ținând cont de literele specifice utilizate în textul inițial.
* Verificarea textului recunoscut se produce utilizând resursele lingvistice reutilizabile specializate pentru perioada respectivă.

|  |
| --- |
| Figura 9. Etapele tehnologice de recunoaștere a textelor tipărite |

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 ABBYY 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 aceasta 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, spațiile și sfârșitul 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 întâlnite în text. Transliterarea presupune stabilirea unei relații bidirecționale univoce între două sisteme de scriere astfel, încât 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.

**2.10. Procesarea textelor tipărite cu alfabet chirilic**

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 local (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 exemplu **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 transliterarea 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 cuvânt din limba română în forma echivalentă scrisă cu caractere chirilice și conform normelor lingvistice acceptate în perioada 1967-1989 în RSSM.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 implicând 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***).
* Reguli pentru **ea** și **ia**. Se transcriu în **я**, cu o singură excepție: pronumele **ea** se transcrie **еа**; în același timp, verbul **ia** se transcrie ca **я**;
* Transcrierea lui **i** are loc prin trei litere diferite: **и**, **й**, **ь**. Concomitent, menționez existența cazurilor cînd litera **i** este omisă (***iepure*** *→* ***епуре***), sau trecută în litera ***ы*** (***introducere*** *→* ***ынтродучере***).

Reguli de transcriere pentru litera ***c***:

* ***c→к***, dacă după ***c*** urmează una din vocalele ***a***, ***â***, ***î***, ***o***, ***u*** sau o consoană diferită de ***h*** (***încrețit → ынкрецит, clocot → клокот, casă → касэ, cucoș → кукош, câmp → кымп***).
* combinațiile ***che***, ***chi*** se vor transcrie în ***ке*** și, respectiv ***ки*** (***cheltuială → келтуялэ, chihlimbar → кихлимбар, chibzui → кибзуи***).
* Dacă după îmbinarea ***ce*** nu urmează litera ***a***, atunci se aplică regula ***ce→че*** (***cercel → черчел, cep → чеп***).
* ***cea→ча*** (***ceară → чарэ, ceas → час, ceață → чацэ, ceașcă → чашкэ***). Excepție pentru articolul demonstrativ ***cea*** (***acea***) ***→*** ***чя*** (***ачя***).
* Dacă după îmbinarea ***ci*** nu urmează una din vocalele a, o, u, atunci se aplică regula ***ci→чи*** (***ciment → чимент, ciclu → чиклу, cimbrișor → чимбришор***). Dacă cuvîntul se termină în ***ci***, atunci poate fi aplicată una din regulile: ca excepție ***ci→ч*** (a**rici *→* арич, beci *→*беч, prichici *→* прикич**); ***ci - чь*** pentru plural (***saci - сачь, maci - мачь***); ***ci→чи*** pentru alte situații (***aci → ачи, răci → рэчи, înveșnici → ынвешничи***).
* ***cio→чо*** (***ciorbă → чорбэ, ciocârlie → чокырлие, cioban → чобан, cocioabă → кочоабэ***).
* ***ciu→чу*** (***ciuperci → чуперчь, ciubotă → чуботэ, bucium → бучум***).

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

|  |
| --- |
| Figura 10: Text digitizat 1894 (Densușianu, 1984, p. 130) |

**2.11. Procesarea textelor tipărite cu alfabet latini și litere adiționale**

Pentru ilustrarea tehnologiei descrise vom cerceta procesul de recunoaștere și verificare a unui text digitizat din cartea(**10**), tipărită în anul 1894 (fig. 10). Textul din figura 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 obține **tnsălbătăcitu** în loc se **însélbătăcitu**.

Acest rezultat nu poate fi îmbunătățit, deoarece IRIS nu posedă capacitatea de a selecta fragmente arbitrare din imagine. Utilizarea unui lexicon modern permite să se recunoască **avutú** ca **avută**, varianta corectă pentru acest context fiind **avut**. Cuvintele specifice lexiconului secolului al XIX-lea nu pot fi recunoscute corect, deoarece pentru aceasta sunt necesare dicționare corespunzătoare perioadei dare care, în cazul nostru, ar conține cuvintele **rămasú, viéța, împératú** etc.

/\*

\* Textul din figura 10 recunoscut, (cu greșeli - și mai jos cu corectare)

\*

\*/

|  |
| --- |
| Româniî, deşi au avută o miie de an! se sufere in ­vasiunele barbare, care au distrusă tote operele măreţe ale architectureT romane, în cătu acesta faptă a rămasă până adt în dicerea populară "n 'a rămasă petră pe petră", totuşt nici moravuri1e niel sufletuln loru nu s'a tnsălbătăcitu. Ei au păstrate o adâncă intimitate şi doioşie in vieţa familiară. Căsătoria este încungiurată de-o mulţime de ceremonil când grave, când vesele. Miresa este "o, fată de tmpăratn", mirele "ficioru de împărată", ceea ce indică respectă şi fericire. Căsătoria este "pe vieţă şi morte", pentru aceea şi jelirea la m6rtea u­nuia dintre sot} este adâncă şi lungă. In ceealaltă lume tnsă er' se tntelnescn pentru a trăi împreună. Cultulă moşiloră (sufletele răposatilorn) este în forte mare 0- n6re până adt. Anumite sărbătort peste anu suntu con­sacrate acestui cultă. Figura. 11. Textul recunoscut cu ajutorul sistemului OCR IRIS |

Dacă în textul recunoscut se vor restabili literele specifice și textul obținut se va verifica cu ajutorul 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, ortografia cărora a rămas intactă față de perioada secolului al XIX-lea, de exemplu **sufere, aceasta, fericire**. Cuvintele „suspicioase” sunt cele afectate de modificări otografice, de exemplu **ceealaltă** (***cealaltă***), **doioșie** (***duioșie***), **miie** (***mie***), **avutú** (***avut***), **aɖĭ** (***azi***).

Pentru recunoașterea corectă a textului trebuie de instruit sistemul OCR ca să recunoască literele și să completeze lexiconul cu cuvinte noi, specifice secolului al XIX-lea. De exemplu: **avutú, o miie, invasiunele, pétră, nicĭ, sufletulú, lorú, însălbătăcitú, doioșie, viéța, ficiorú, împăratú, miresa** etc. Ținînd cont că sistemul OCR ABBYY FineReader este înzestrat cu facilități de instruire, am mai efectuat un experiment. Sistemul a fost instruit în mod special ca să poată recunoaște literele specifice secolului al XIX-lea. Mai jos prezint câteva din literele care au fost utilizate în procesul de instruire:

* **ü** (literă finală, mută sau citită),
* **é** (é se pronunță ca diftongul ea),
* **ó** (ó se pronunță ca diftongul oa),
* **ɖ** (ɖ se citea ca z sau dz, în dependență de conținut),
* **ê** (ê se folosea ca litera 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 recunoaște caracterele specifice;
* să fie elaborat un lexicon cu cuvinte și fraze uzuale specifice perioadei;
* să fie extinse facilitățile corectorului ortografic (spellchecker) pentru a utiliza și lexiconul elaborat.

**2.12. Procesarea textelor tipărite cu alfabete tranziționale**

Există cel puțin șapte versiuni ale alfabetului tranzițional (mixt chirilic-latin). Majoritatea literelor acestor alfabete pot fi recunoscute de ABBYY FineReader prin evidențierea codurilor respective din setul Unicode. O singură literă specifică pentru aceste alfabete lipsește în unicode - ꙟ. În acest caz urmează să fie inclusă o variantă de literă echivalentă (de exemplu o săgeată ↑ sau slavonica „yus” Ѧ ѧ) și instruit sistemul pentru recunoașterea acestei variante grafice.

|  |  |  |  |
| --- | --- | --- | --- |
| **Tabelul 2:** Rezultatele experimentelor OCR cu texte din secolul XIX-lea | | |  |
| **Modul de recunoaștere** | **Cuvinte corecte** | **Cuvinte suspecte** | |
| IRIS | 57% | 43% | |
| ABBYY FineReader, fără nici o instruire | 63% | 37% | |
| ABBYY FineReader, cu instruire și dicționar pentru o pagină | 98% | 2% | |
| ABBYY FineReader, cu instruire, mai multe pagini, aceiași carte | 95% | 5% | |
| ABBYY FineReader, cu instruire, pagini din altă carte | 95.4% | 4.6% | |

**2.13. Procesarea textelor tipărite cu alfabetul slavon (sec. XVIII - XIX)**

/\*

\* algoritmul transliterării textelor vechi slavone

\*

\*

\*

\*/

**Concluzii preliminare**

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 recunoașterea literelor specifice perioadei; **2**. Elaborarea instrumentarului și interfețelor necesare pentru crearea resurselor lingvistice (lexiconului) corespunzător perioadei în scopul eficientizării procesului de recunoaștere a cuvintelor și de aliniere a normelor lingvistice contemporane.

La trecerea de la o perioadă la alta, în limitele posibilităților, se vor utiliza intrumentarul și resursele deja elaborate, materializînd astfel principiul „din prezent în adîncul secolelor”. Resursele electronice create pot fi amplasate pe internet pentru acces public, contribuind la dezvoltarea mediului de comunicare informațională pentru limba română. În plus, aceste resurse ar constitui un suport esențial pentru cercetători, iar convertite în text literar ar putea fi utilizate ca materiale didactice în procesul de instruire.

**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, 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 și pe aliniere paralelă a textelor.

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