Transliterare automată din engleză în română. Aplicaţii şi rezultate

Tiberiu Boroş[1], Adrian Zafiu[2]

[1]Institutul de Cercetări pentru Inteligenţă Artificială, Academia Română
Strada 13 Septembrie, nr. 13, Bucureşti
E-mail: tibi@racai.ro

[2]Universitatea din Piteşti
Strada Târgul din Vale, nr. 1. Piteşti
E-mail: adrian.zafiu@upit.ro

Rezumat. 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 acest articol, propunem o metodă statistică pentru translatarea din limba engleză în limba română (dar care poate fi adaptată pentru orice pereche de limbi), prezentăm o serie de aplicaţii care utilizează transliterarea în sinteza vorbirii (pentru texte multilingve) şi introducem conceptul de "căutare pe bază de percepţie".

Cuvinte cheie: transliterare, sinteza vorbirii, maximum entropy.

1. Introducere

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, chineză, japoneză, rusă sau bulgară).

Sistemele de sinteză a vorbirii pornind de la text (TTS - text-to-speech) care, prin definiţie, 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 învățate automat pentru limba pe care a fost proiectat sistemul.

O soluție pentru rezolvarea acestei probleme este introducerea unor pachete suplimentare de reguli în vederea obținerii 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 țintă.

În cadrul acestei lucrări, propunem o abordare diferită a transliterării, în care utilizăm un transliterator pentru adaptarea de la o limbă la alta înainte de a aplica transcrierea fonetică. Aceasta din urmă se obține folosind acelaşi pachet de reguli utilizat pentru cuvintele native (ale limbii țintă).

În cazul limbii române, datorită ortografiei preponderent fonetice, urmărim ca acuratețea obținută prin transliterare din engleză în română, urmată de transcrierea fonetică aplicată pentru limba română să fie comparabilă cu acuratețea obținută în situația în care s-ar aplica reguli de transcriere fonetică direct pentru engleză urmând să se facă ulterior o adaptare la nivel fonetic între cele două limbi. Practic, acuratețea globală este limitată în ambele cazuri de performanțele sistemelor de transcriere fonetică pentru limba engleză.

O altă aplicație pentru transliterare este cea numită de noi "căutare pe bază de percepție". Căutarea pe bază de percepție este utilă pentru a găsi o persoană, un loc, o stradă, un oraş sau orice alt cuvânt străin folosind percepția individuală asupra modului în care "sună" cuvântul respectiv.

2. Stadiul actual

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 (1997) au introdus o metodă de transliterare între japoneză şi engleză, utilizând algoritmi de traducere bazați pe maşini cu stări finite, această metodă fiind adaptată de Stalls şi Knight (1998) pentru transliterare bidirecțională între engleză şi arabă. Alte metode de transliterare sunt descrise de Jung et al. (2000), Meng et al. (2001), Virga şi Khudanpur (2003).

În lucrarea lor, Haizhou et al. (2004) clasifică metodele menționate mai sus ca fiind abordări ale transliterării bazate pe nivel fonetic. Ei propun o nouă tehnică, numită de către autori mapare-ortografică-directă (DOM -

direct orthographic mapping) sau model de transliterare pe bază de n-grame (secvențe de *n* litere consecutive care pot apărea în cuvintele unei limbi).

Experimentele noastre sunt focalizate asupra transliterării din engleză în română, fiind parte a studiului efectuat pentru modulul dedicat cuvintelor străine, integrat într-un sistem de sinteză a vorbirii pentru limba română.

3. Corpusul pentru transliterarea din engleză în română

Metoda pentru transliterare necesită un corpus de antrenare format din cuvinte scrise în limba sursă (în acest caz engleză) asociate cu transliterările lor corespunzătoare în limba țintă (limba română).

Faptul că limba română are o ortografie preponderent fonetică ne-a permis să folosim 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 (CMU, 2011). Pe lângă cuvintele uzuale, acest lexicon conține un număr mare de nume proprii, abrevieri şi cuvinte adaptate la limba engleză, care sunt provenite din arabă, germană, franceză, poloneză etc.:

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

Aceste cuvinte încurcă procesul de transliterare din engleză în română deoarece conversia lor în foneme nu se poate obține cu ajutorul regulilor standard. Soluția pentru a învăța doar reguli omogene a constat în filtrarea CMUDict prin alegerea unui set de cuvinte uzuale din limba engleză pe baza cărora am generat lexiconul de transliterare (aproximativ 20.000 de cuvinte). Folosind apoi datele din CMUDict, am generat transcrierea fonetică 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 (vezi tabelul 1).

Pentru limbile fără ortografie fonetică sunt necesari doi paşi suplimentari față de cei prezentați anterior. Primul pas constă în maparea dintre fonemele specifice pentru limba sursă şi fonemele care apar în limba destinație. 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 țintă. Ultimul pas se poate realiza folosind metode automate, dar, pentru rezultate bune, este necesară o iterație suplimentară care constă în validarea manuală a rezultatelor obținute automat.

Tabelul 1. Reguli de conversie din engleză în română

En	***Cuvânt***	***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	**gi**
K+(E/I)	key	**K** IY	**ch**i
K	all	**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	**ău**t
OY	toy	T **OY**	t**oi**
P	pee	**P** IY	**p**i
R	read	**R** IY D	**r**id
S	sea	**S** IY	**s**i
SH	she	**SH** IY	**ş**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 menționat că întregul proces de transliterare se face cu pierdere de informație atât din cauza metodelor statistice folosite cât și pentru că nu toate fonemele au echivalent direct în limba română și, pe baza contextului, unele litere din alfabetul românesc pot avea o pronunție diferită față de cea dorită. Pentru TTS 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 percepție, pentru a diminua efectele nedorite generate de această pierdere de informație, am folosit transliterarea din engleză în română (forward-transliteration) în locul transliterării din română în engleză (backward-transliteration) deși căutarea se face folosind cuvinte scrise în română (ceea ce ar implica backward-transliteration) (Knight și Graehl, 1997).

4. Transliterare din engleză în română

Practic, problema transliterării se poate reformula astfel: găsirea unui set de reguli care pornind de la un şir de simboluri/caractere ce aparţin alfabetului sursă (cuvinte ce trebuie transliterate) obţine un şir de simboluri/caractere ce aparţin alfabetulului destinaţie, astfel încât aplicând pachetul de reguli pentru transcrierea fonetică specific fiecărei limbi pe ambele şiruri (cel de intrare şi repectiv cel de ieşire) similaritatea între sunetele obţinute să fie maximă. O observaţie este că metodele bazate pe manipulare ortografică directă au o acurateţe mai bună decât cele din prima categorie, ceea ce ne-a determinat să alegem şi noi o abordare de tip DOM.

Transcrierea fonetică este o problemă destul de asemănătoare cu cea a transliterării. Diferenţa dintre cele două este că în cazul transcrierii fonetice se caută un pachet de reguli pentru a trece din literele unui cuvânt în simbolurile folosite pentru reprezentarea fonetică a acestuia, faţă de transliterare unde se caută o mapare către ortografia altei limbi.

Într-un articol anterior (Boroş et al., 2012) am introdus o serie de metode statistice pentru transcrierea fonetică automată, implementate într-o unealtă (Bermuda) ce se poate descărca de pe situl de unelte pentru prelucrarea limbajului natural al Institului de Cercetări pentru Inteligenţă Artificială "Mihai Drăgănescu" (http://nlptools.racai.ro). Unealta foloseşte un clasificator de tip Maximum Entropy (MaxEnt), care asociază etichete fiecărei litere dintr-un cuvânt pe baza unei serii de trăsături extrase din contextul lexical al literei respective.

Secvenţa de etichete obţinută pentru o secvenţă de litere a unui cuvânt constituie transcrierea fonetică a acestuia. Notând cu l litera curentă, l_i litera aflată la distanţă i faţă de litera curentă şi p_{-1} eticheta anterioară, trăsăturile folosite de noi sunt:

- l_{-1},l – litera curentă plus litera anterioară;
- l_{-2} ,l_{-1},l - litera curentă plus două litere anterioare;
- l,l_{+1} - litera curentă plus litera următoare ;
- l,l_{+1},l_{+2} - litera curentă plus următoarele două litere ;
- l_{-1},l,l_{+1} - litera curentă plus literele imediat învecinate ;
- p_{-1} - eticheta emisă pentru litera anterioară.

Acuratețea obținută cu ajutorul acestei combinații de trăsături este de 93% pentru cuvinte din afara vocabularului transcrise corect fonetic în limba română și 67% pentru cuvinte din afara vocabularului (extrase din CMUDict) transcrise corect fonetic în limba engleză. Trebuie menționat faptul că acuratețea de 67% pentru CMUDict este una foarte bună în comparație cu majoritatea metodelor propuse de alți autori pentru transcriere fonetică, care au o acuratețe ce variază între valorile 57% și 65% (Black et al., 1998; Bosch și Canisius, 2006; Rama et al., 2009). Performanța maximă, de 71%, pentru acest lexicon a fost obținută folosind Marginal Infused Relaxed Algorithm (MIRA) (Jiampojamarn et al., 2008).

Dat fiind cele enumerate mai sus, am luat decizia de a utiliza aceeași metodă și pentru transliterarea din engleză în română, antrenând același clasificator, dar de data aceasta pe baza lexiconului de transliterare.

Urmărind trăsăturile menționate, clasificatorul atribuie pentru fiecare literă din context o etichetă ce reprezintă un simbol, un grup de simboluri din alfabetul destinație sau, în anumite situații, mulțimea vidă.

Etichetele sunt obținute pe baza lexiconului de antrenare, folosind alinieri automate ale literelor cuvintelor din alfabetul sursă cu literele corespunzătoare din alfabetul destinație. O metodă tipică pentru a obține aceste alinieri este Expectation-Maximization (EM) (Hartley, 1958; Dempster et al., 1977).

Am evaluat transliterarea din engleză în română, folosind metoda 10-fold. Am împărțit setul de date de antrenare în 10 submulțimi egale și am testat acuratețea sistemului pe fiecare submulțime în parte, antrenându-l pe celelalte 9 și calculând rata de acuratețe la nivel de cuvânt (WAR - word-accuracy-rate). Aceasta a fost calculată ca număr de cuvinte transliterate corect raportată la numărul total de cuvinte procesate. Dacă un cuvânt transliterat are chiar și o literă greșită, acesta se contorizează la erori. Acuratețea sistemului nostru, ca medie a celor 10 validări, a fost 78,34%. Trebuie menționat că rezultatul este raportat la cuvinte din afara vocabularului de antrenare. Deoarece limba română are o ortografie preponderent fonetică ne-am aștepta ca rata de acuratețe în cazul transliterării să fie asemănătoare cu rezultatul obținut de transcrierea fonetică. Creșterea cu 10% a acurateții în cazul transliterării, se datorează faptului că am făcut antrenarea pe cuvinte pur englezești (fără cuvinte străine sau abrevieri conținute de CMUDict). În plus, fiind vorba de o mapare ortografică directă între cele două limbi, am redus numărul de

etichete care se puteau atribui unei litere sau grup de litere în momentul clasificării.

5. Aplicații ale transliterării

5.1 Transliterarea din engleză în română pentru sinteza vorbirii

Aşa cum a fost menționat în introducere, există multe situații în care un text ce trebuie sintetizat conține cuvinte provenite din alte limbi, care nu pot fi procesate folosind acelaşi pachet de reguli ca cele specifice limbii țintă pentru care sistemul este destinat (în cazul nostru vorbim despre limba română). În această situație, o abordare evidentă constă în folosirea unor seturi distincte de reguli de transcriere fonetică pentru aceste cuvinte, reguli specifice limbii din care ele fac parte, fiind necesară şi o adaptare ulterioară la nivel fonetic. În cadrul acestui articol, propunem o abordare diferită ce constă în folosirea unui transliterator pentru a genera "pseudo-cuvinte" native. În acest caz, transcrierea fonetică se face folosind reguli specifice limbii române şi nu mai este necesară o adaptare la nivel fonetic a rezultatelor. Există situații în care, dat fiind contextul lexical, anumite litere din limba română generează sunete diferite de cele intenționate. Cu toate acestea, cuvintele sună natural pentru un vorbitor nativ de limba română şi nu prezintă dificutăți în ceea ce priveşte inteligibilitatea.

Diferența între cele două abordări majore este subliniată în figura 1:

1. metodele bazate pe manipulare fonetică se aplică doar în cazul în care sunt cunoscute seturile de reguli pentru conversie din litere în sunete (transcriere fonetică) pentru ambele limbi implicate în proces ;
2. metodele bazate pe manipulare ortografică directă nu necesită cunoaşterea pachetelor de reguli folosite în transcrierea fonetică pentru limba sursă ci doar pentru limba destinație.

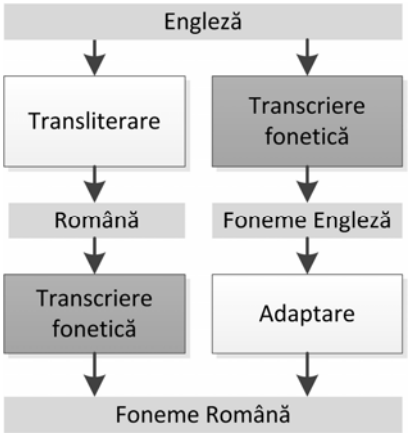

Figura 1 Conversia cuvintelor scrise în engleză pentru sinteza vorbirii

5.2 Detectarea cuvintelor care au nevoie de transliterare

O dificultate cu care se confruntă ambele abordări pentru procesarea cuvintelor străine în sinteza vorbirii, este alegerea corectă a cazurilor în care sistemul TTS trebuie să aplice transliterarea cuvintelor din engleză în română. Principala modalitate de a verifica dacă un cuvânt trebuie transliterat este folosirea lexiconului generat anterior. Este evident faptul că dacă un cuvânt se găseşte în tabela de transliterare el este şi un candidat bun pentru acest proces. Cu toate acestea, trebuie verificat dacă acesta există şi în inventarul de cuvinte al limbii române. De exemplu, nu se poate spune cu exactitate dacă translitererea are sens pentru cuvântul "minus" care are aceeaşi semantică şi ortografie atât în limba română cât şi în limba engleză. Acest tip de cuvinte vor fi, de preferat, stocate într-un fişier separat şi vor rămâne neschimbate.

În situațiile în care un cuvânt nu se găseşte nici în inventarul de cuvinte cunoscute pentru limba română şi nici în cel pentru limba engleză este greu de precizat dacă acesta trebuie transliterat sau nu. O metodă de a rezolva aceste situații o reprezintă folosirea unor indicii lexicali pentru a decide care este limba din care provine acest cuvânt. Anumite grupuri de litere sunt foarte rare sau chiar nu pot exista în limba română. Cuvintele care conțin litera 'y', grupul de litere "ck", ș.a.m.d. sunt candidați buni pentru transliterare. Astfel, am generat o listă de perechi de câte trei litere consecutive posibile în limba română, folosind Dicționarul Explicativ al Limbii Române. Atunci când sistemul întâlneşte un cuvânt necunoscut

(pentru ambele limbi luate în considerare), testează dacă fiecare grup de trei litere consecutive din cuvânt se regăseşte în lista generată anterior. Orice cuvânt care conţine o combinaţie de litere care nu este specifică pentru limba română, este transliterat automat. Celelalte cuvinte sunt lăsate neschimbate (figura 2).

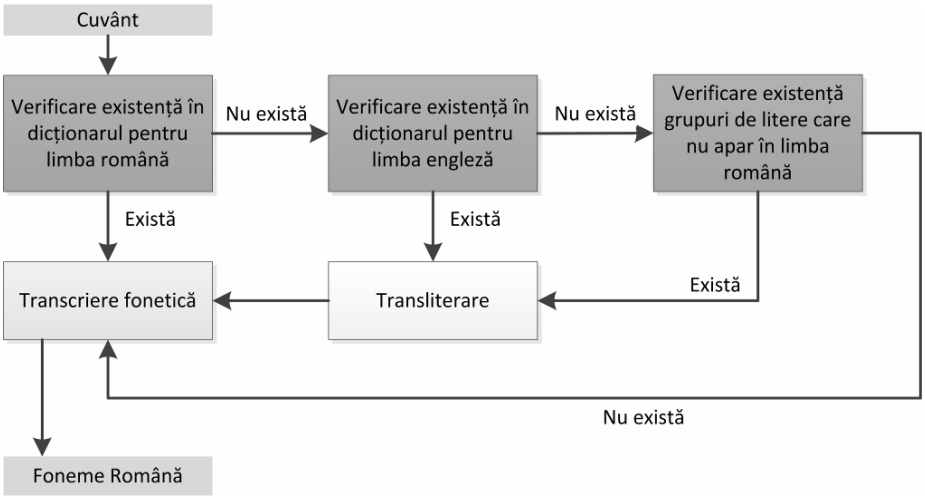

Figura 2. Detectarea cuvintelor care au nevoie de transliterare

În continuare sunt date câteva exemple:

1. Text preluat de la pagina de test a sistemului TTS IVONA (http://www.ivona.com/en/):
   - **Text original:** Scrie orice text folosind corect diacriticele şi apasă pe butonul "**Play**".
   - **Text procesat:** Scrie orice text folosind corect diacriticele şi apasă pe butonul **plei**.
   - **Explicaţii:** Cuvântul “play” a fost identificat direct în lexiconul de transliterare ca fiind fără corespondent în limba română.
2. Text preluat de pe WIKIPEDIA (http://ro.wikipedia.org/wiki/Washington_%28stat%29):
   - **Text original: Washington** este un stat al Statelor Unite ale Americii, situat în aşa numita zonă a Pacificului de Nordvest a Statelor Unite continentale.

- **Text procesat: uaşingtăn** este un stat al Statelor Unite ale Americii, situat în aşa numita zonă a Pacificului de Nordvest a Statelor Unite continentale.
- **Explicații:** Cuvântul "Washington" nu se afla în lexiconul de transliterare însă, combinația de litere "**sh**" a determinat transliterarea automată a acestuia în "uaşingtăn". Conform sursei Wikipedia transliterarea pentru acest cuvânt este "uaşingtăn".

5.3 Căutarea pe bază de percepție

Aşa cum am menționat mai devreme, căutarea pe bază de percepție este un mod de a găsi scrierea corectă a unui cuvânt dintr-o limbă străină (engleză, franceză, germană, rusă etc.) în funcție de modul în care acest cuvânt este perceput (modul în care "sună" cuvântul) pentru un vorbitor nativ (în cazul nostru limba română).

De exemplu, să presupunem că nu am şti nimic altceva despre un oraş cu excepția faptului că sună oarecum ca „ianțiau" şi nu am avea informații cu privire la țara în care se află sau vreo informație despre ortografia pe care ar trebui să o folosim pentru a găsi mai multe date despre locație. Căutarea pe bază de percepție ar putea fi folosită pentru a obține scrierea exactă a denumirii locației prin simpla tastare a cuvântului aşa cum este el perceput în limba nativă a utilizatorului. Un vorbitor nativ de limba română ar introduce doar cuvântul „ianțiau", care este cea mai apropiată formă ortografică din limba sa, iar rezultatul ar fi "燕郊" – localitate aflată în nord estul Chinei, în provincia Hebei.

Această metodologie are câteva neajunsuri:

- Părerea unei persoane nevorbitoare nativ de limba sursă despre cum ar trebui să fie scris un cuvânt în limba sa nativă nu este 100% 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 anumite situații.
- Sunt mai multe cazuri în care scrieri diferite sunt pronunțate la fel (omofone).
- Aşa cum a fost menționat de Knight şi Graehl (1997) transliterarea inversă (backward-transliteration) nu are aceeaşi flexibilitate ca şi transliterarea directă (forward-transliteration) (pierderea de

informație este de două ori mai mare atunci când se transliterează înainte şi înapoi între două limbi).

Prima idee când folosim căutarea bazată pe percepție este de a antrena sistemul să translitereze între limba nativă – sursă - (în care s-a efectuat căutarea) şi toate limbile țintă, alegând cea mai bună variantă pe baza unei funcții de similaritate între şiruri de caractere.

Această metodă este predispusă unei serii de erori, cum ar fi faptul că o reprezentare fonetică "percepută" a unui cuvânt poate corespunde mai multor forme ortografice şi, desigur, pierderea de informație generată de incompatibilitățile pachetelor fonetice ale celor două limbi.

Pentru a compensa asemenea erori am propus o abordare diferită şi anume: toate cuvintele din limba țintă sunt transliterate în limba nativă de căutare. Când se efectuează căutarea, comparăm transliterația curentă (dată de utilizator) cu toate transliterațiile din baza de date folosind distanța Levenshtein (vezi secțiunea următoare pentru rezultate).

5.4 Evaluarea căutării bazate pe percepție

Pentru a valida metoda noastră de căutare am creat alt corpus de test, compus doar din nume de oraşe din Statele Unite ale Americii. Corpusul conține 480 de intrări selectate la întâmplare (nu are nimic în comun cu corpusul de transliterare - aşa cum a fost menționat 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 navigație, 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 înregistrările 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 percepție prezentată anterior. La calcularea acurateții sistemului s-a obținut 99.38% (doar 3 cuvinte nu au fost identificate corect).

6. Concluzii

În articolul de față am prezentat o metodă de transliterare între engleză și română, care, cu câteva adaptări specifice, poate fi aplicată și pe alte perechi de limbi. Am creat un corpus de antrenare pentru transliterare ce poate fi obținut în mod semiautomat (fără efort pentru limbile cu o ortografie fonetică) și toate instrumentele sunt disponibile pentru descărcare pe pagina noastră web.

Acuratețea de 78% a transliterației TTS este raportată la cuvinte din afara vocabularului (out-of-vocabulary – OOV). În practică, nu toate cuvintele străine sunt necunoscute și, 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, intenționam să extindem lexiconul de transliterare la Franceză și Germană.

Rezultatul obținut prin căutarea pe bază de percepție arată că motoarele de căutare și asistenții de călătorie ar beneficia de pe urma unui asemenea instrument. Căutarea după percepție poate îmbunătăți experiența utilizatorului de internet, iar, în același timp, concentrarea pe corectarea greșelilor de ortografie bazată pe similitudini fonetice (care la un anumit nivel poate fi legată de transliterare) între cuvinte poate îmbunătăți procesul de corectare ortografică (Li et al., 2006).

Referințe

Black, A., Lenzo, K. and Pagel, V. (1998) Issues in building general letter to sound rules. ESCA Speech Synthesis Work-shop, Jenolan Caves.

Boroș, T., Ștefănescu, D., Ion, R. (2012) Bermuda, a data-driven tool for phonetic transcription of words. Proceedings of the Natural Language Processing for Improving Textual Accessibility, Language Resources and Evaluation (LREC), Istanbul, Turkey.

Bosch, A., and Canisius, S. (2006) Improved morpho phonological sequence processing with constraint satisfaction inference. Proceedings of the Eighth Meeting of the ACL-SIGPHON at HLT-NAACL, pp. 41–49.

CMU (2011). Carnegie Mellon Pronuncing Dictionary. http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

Dempster, A., Laird, N., and Rubin, D. (1977) Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, 39(1), pp. 1–38.

Hartley, H. (1958) Maximum likelihood estimation from incomplete data: Biometrics, 14, pp. 174–194.

Jiampojamarn, S., Cherry, C. and Kondrak, G. (2008) Joint processing and discriminative training for letter-to-phoneme conversion. Proceedings of ACL-2008: Human Language Technology Conference, Columbus, Ohio, pp. 905–913.

Jung, S. Y., Hong, L. S. şi Paek, E. (2000) An English to Korean Transliteration Model of Extended Markov Window. Proceedings of COLING.

Knight, K. and Graehl, J. (1997) Machine transliteration. Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics, Somerset, New Jersey, pp. 128–135.

Li, H., Zhang, M. şi Su, J. (2004) A joint source-channel model for machine transliteration. Proceedings of the 42nd ACL Annual Meeting, Barcelona, Spain, pp. 159–166.

Li, M., Zhang, Y., Zhu, M. and Zhou, M. (2006) Exploring distributional similarity based models for query spelling correction. Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL, pp. 1025–1032.

Meng, H.M., Lo,W-K., Chen, B. şi Tang, K. (2001) Generate Phonetic Cognates to Handle Name Entities. English-Chinese cross-language spoken document retrieval, ASRU.

Rama, T., Singh, A. K., Kolachina, S. (2009): Modeling Letter-to-Phoneme Conversion as a Phrase Based Statistical Machine Translation Problem with Minimum Error Rate Training. Proceedings of the 2009 Named Entities Workshop, ACL-IJCNLP 2009, Suntec, Singapore, pages 124–127.

Stalls, B.G. şi Knight, K. (1998) Translating Names and Technical Terms in Arabic Text. Proceedings of the COLING/ACL Workshop on Computational Approaches to Semitic Languages.

Virga, P., Khudanpur, S. (2003) Transliteration of Proper Names in Crosslingual Information Retrieval. Proceedings of ACL 2003 workshop MLNER.