RAPORT ȘTIINȚIFIC proiect complex ReTeRom, etapa II - noiembrie 2019

Proiectul 2: TEPROLIN

Activitatea 2.7

Denumire activitate: Transcrierea fonetică a cuvintelor din lexiconul validat

Autori: ICIA

REZUMATUL ETAPEI

Această activitate a proiectului 2, TEPROLIN, are rolul de a îmbogăți lexiconul cu transcrieri fonetice, precum și de a valida și corecta transcrierea fonetică a lexiconului, printr-un proces parțial automatizat, parțial manual, și trecând prin validarea informației de silabificare și accent.

Rezultatele activității: (i) Raport asupra lexiconului îmbogățit cu transcriere fonetică;

(ii) Lexiconul îmbogățit cu transcrieri fonetice devine accesibil public pe site-ul proiectului.

Introducere. Vorbeam într-un raport anterior (Act. 1.8[1]) despre importanța lexiconului pentru recunoașterea automată a vorbirii (ca inventar de cuvinte cunoscute sistemului de transcriere și mijloc de creare de modele acustice) și pentru sinteza vorbirii (furnizând transcrierea fonetică sau pronunția echivalentă formei scrise a unui cuvânt; util în special în cazul abrevierilor și pentru că oferă variație morfologică, incluzând fiecare formă flexionară a unei leme). Pentru toate aceste motive, este esențial ca lexiconul nu doar să existe, ci să aibă și o calitate foarte bună. Calitatea resurselor nu poate fi asigurată decât printr-un proces de validare și corectare; acesta poate fi automatizat parțial, dar o etapă de corectare manuală este obligatorie în contextul unei limbi care nu dispune de corespondență 1:1 între grafie și pronunție, cum este limba română. În raportul 1.8. am exemplificat parțial fenomenele de omografie și omofonie care pun probleme transcrierii fonetice în limba română, dar ele vor fi detaliate în cadrul acestui raport pentru că reprezintă situații de ambiguitate pe care a fost necesar să ne concentrăm în procesul de corectare manuală.

La sfârșitul anului 2018, raportam că în procesul de completare a lexiconului cu informația de silabificare, accent și transcriere fonetică s-au folosit resursele:

- un dicționar de silabificare și accent – RoSyllabiDict (Barbu, 2018) – ce conține 525.534 forme flexionare, corespunzând la aproximativ 65.000 leme

- un dicționar de transcriere fonetică – MaRePhor (Toma et al., 2017) – ce conține 72.375 leme.

Pentru cuvintele regăsite în resursele de mai sus (cele două liste de leme se suprapun doar parțial), informațiile de silabificare și accent pe de o parte și de transcriere fonetică pe de altă parte sunt recuperate printr-o procedură simplă de căutare (*eng. look up*). Pentru cuvintele care sunt în lexiconul nostru dar nu și în resursele menționate, s-a apelat la generarea cu Romanian TTS (Stan et al., 2011), integrat în lanțul de prelucrare TEPROLIN (Ion, 2018), pentru obținerea acestor informații. A rezultat o resursă cu 346.074 intrări cuprinzând informație standardizată asociată unei forme ocurență: lema (forma de dicționar a cuvântului), eticheta

[1] http://www.racai.ro/p/reterom/rapoarte/1.8.pdf

morfo-sintactică în format MSD (Erjavec, 2004), împărțirea în silabe a formei ocurență, marcarea accentului (printr-un apostrof) în fața vocalei accentuate (vocala a, în cazul exemplului de mai jos) și transcrierea fonetică a formei ocurență, între paranteze drepte.

Un exemplu de intrare în lexiconul ReTeRom, care ilustrează formatul generalizat[2] este:

întreagă întreg Afpfsrn în.trea.gă între'agă [1 n t r e_X a g @]

Corectarea erorilor de transcriere fonetică

Metodologia. Pornind de la premisa că resursele utilizate (RoSyllabiDict și MaRePhor) au fost, așa cum susțin autorii lor, validate manual parțial înainte de lansare, ne-am concentrat pe validarea și corectarea acelor intrări din lexiconul ReTeRom care conțin informație generată automat cu Romanian TTS. Au fost implementate[3] reguli de corectură automată a transcrierii fonetice bazate pe silabificare și accent, care presupun că aceste informații sunt corecte (în practică, a existat o etapă anterioară de validare manuală a lor). S-au efectuat corecturi/validări manuale în cazuri de transcriere fonetică pe care le-am identificat ca ambigue/problematice.

Alfabetul fonetic SAMPA. Alfabetul fonetic utilizat în MaRePhor și preluat în lexiconul nostru este SAMPA (vezi tabelul de mai jos pentru exemple). Legat de sunetele ce/ci/ge/gi/che/chi/ghe/ghi, se poate observa ilustrată în tabel următoarea convenție de transcriere: Atunci când în aceeași silabă cu aceste grupuri există o vocală, aceste grupuri trebuie transcrise ca reprezentând un singur fonem (vezi transcrierile pentru "ceas", "ciută", „chiag", „chiar", "geană", „cafegioaică", „lighean", „ghiotură"). Dacă în silabă mai există o semivocală, atunci aceste grupuri se transcriu ca reprezentând două foneme.

litere pentru VOCALE	**exemplu**	**transcriere**	**transcrierea exemplului**
a	cap	a	k a p
ă	hău	@	h @ w
â	câine	1	k 1 j n e
e	eter	e	e t e r
	eu	je	je w
i	vin	i	v i n
î	înapoi	1	1 n a p o j
o	acolo	o	a k o l o
u	sur	u	s u r
litere pentru SEMIVOCALE			
e	neam	e_X	n e_X a m
i	iac	j	j a k
o	soare	o_X	s o_X a r e
u	sau	w	s a w
I șoptit			
i	pomi	i_0	p o m i_0
litere pentru CONSOANE			
b	ban	b	b a n
c	car	k	k a r

[2] *<formă>tab<lemă>tab<etichetă_morfo-sintactică>tab<silabificare>tab<accent>tab<transcriere_fonetică>*

[3] scripturile de implementare a regulilor au fost scrise în C#, sub platforma Microsoft Visual Studio 2019

ce	cerebel	tS	tS e r e b e l
	ceas	tS	tS a s
	cercei	tS	tS e r tS e j
ci	cine	tS	tS i n e
	ciută	tS	tS u t @
che	chema	k_j	k_j e m a
	cheag	k_j	k_j a g
chi	chip	k_j	k_j i p
	chiar	k_j	k_j a r
	rochii	k_j	r o k_j i j
d	dor	d	d o r
f	aftă	f	a f t @
g	ogar	g	o g a r
ge	ager	gZ	a gZ e r
	geană	gZ	gZ a n @
gi	legifera	gZ	l e gZ i f e r a
	cafegioaică	gZ	k a f e gZ o_X a j k @
	lefegiu	gZ	l e f e gZ i w
ghe	ghem	g_j	g_j e m
	lighean	g_j	l i g_j a n
ghi	ghindă	g_j	g_j i n d @
	ghiotura	g_j	g_j o t u r a
h	han	h	h a m
j	joc	Z	Z o c
k	karat	k	k a r a t
	kilogram	k_j	k_j i l o g r a m
l	alt	l	a l t
m	om	m	o m
n	nas	n	n a s
p	pat	p	p a t
q	Qatar	k	k a t a r
qu	Maquis	k_j	m a k_j i s
r	rod	r	r o d
s	sat	s	s a t
ș	ușă	S	u S @
t	ateu	t	a t e w
ț	țar	ts	ts a r
v	vin	v	v i n
w	watt	v	v a t
x	exonera	ks	e ks o n e r a
	examen	gz	e gz a m e n
y	yankeu	j	j a n k_j e w
z	zid	z	z i d

Pentru a ușura procesul de validare manuală, am divizat lexiconul în fișiere cu diferite grade de risc de eroare, după următoarele criterii:

- Informația este sau nu obținută din resursele RoSyllabiDict și MarePhor (am presupus că această informație este de o calitate mai bună decât cea generată cu Romanian TTS)
- conțin sau nu conțin un grup consecutiv de cel puțin două vocale; în acest caz, avem de a face cu situații de hiat, diftong, triftong, pentru care, în funcție de corectitudinea silabificării, se pot implementa reguli de transcriere;
- conțin consoana „x", care se poate transcrie fie drept „ks", fie drept „gz": 7370 intrări, corespunzând la 594 de leme; intrările pentru care forma este identică lemei au fost validate manual, iar informația de transcriere a consoanei a fost copiată automat la toate formele sale;
- conțin grupurile: "ce", „ci", „ge", „gi", „che", „chi", „ghe", „ghi"
- conțin cratimă în interiorul cuvântului (cuvinte compuse)
- conțin prefixul "nemai": în corpus se afla un număr de 6827 de verbe la participiu și gerunziu, silabificate automat cu Romanian TTS în mod eronat (ex.: "ne.maia.vând", în loc de "ne.mai.a.vând");
- conțin mai multe variante de silabificare sau accent pentru aceeași formă ocurentă (nici RoSyllabiDict, nici MarePhor nu dezambiguizează morfo-sintactic formele ocurente):

  Exemplu: iei ie Ncfsoy i.ei 'iei [i e j]

  iei lua Vmip2s iei iei [j e j]

- sunt nume proprii sau abrevieri: acestea sunt generate automat și pun multe probleme, pentru că numerele proprii în limbi străine respectă reguli diferite de pronunție, iar numele proprii în limba română au adeseori pronunții atipice; Romanian TTS nu tratează silabificarea abrevierilor în mod corespunzător; în total, 5440 de nume proprii și 373 de abrevieri au fost corectate manual;

Procesul de corectare manuală și automată s-a desfășurat în ordinea silabificare -> accent -> transcriere fonetică. Această ordine permite automatizarea procesului, deoarece, în multe cazuri, transcrierea fonetică depinde de silabificare și accent. De asemenea, se pot evidenția anumite cazuri de ambiguitate a accentului atunci când două forme identice au silabificări diferite (vezi exemplul de mai sus pentru forma "iei").

Corectarea silabificării s-a făcut integral manual, concentrându-ne pe:

- cuvintele care nu se găsesc în RoSyllabiDict;
- cuvinte care conțin silabe mai lungi de patru litere;
- cuvinte care conțin secvențe de vocale + semivocale: vezi cazul hiat versus diftong, triftong;
- cuvinte la vocativ: 9531 de intrări au fost validate și corectate manual; "o"-ul final al substantivelor la vocativ poate să formeze singur o silabă, sau poate să fie legat de alte litere/sunete în silabă, precum în exemplele de mai jos:

tristețeo tristețe Ncfsvy tris.te.țeo tristețeo [t r i s t e ts e_X o]

informațio informație Ncfsvy I n.for.ma.ți.o inform'ație [i n f o r m a ts j o]

- nume proprii și abrevieri;
- cuvinte cu silabe care conțin mai mult de o vocală, știindu-se că o silabă poate conține o singură vocală și mai multe semivocale; literele "a", „â", „î, "ă" nu pot fi semivocale, iar silabele care conțin două astfel de litere sunt obligatoriu greșite.

În corectarea manuală a accentului, ne-am concentrat pe cuvintele cu două variante de accent (omonimii) și pe numele proprii. În continuare, enumerăm câteva tipuri de omonimii posibile, în funcție de partea de vorbire a formei implicate:

- tipul FOTOGRAFII: substantive care în forma de plural se termină în “ii” și au două leme posibile: FOTOGRAF/FOTOGRAGIE; a) daca are lema terminata în "ie" (FOTOGRAFIE), atunci accentul este pe penultimul i din cuvânt; b) daca are lema terminată în consoană (FOTOGRAF), atunci accentul nu este pe finala "ii".
- tipul DATA/DATĂ: poate fi substantiv sau verb: a) dacă e verb, accentul e pe ultima silabă; b) dacă e substantiv, accentul nu e pe ultima silabă;
- tipul ATRIBUI: forme verbale diferite ale aceleiași leme: a) dacă eticheta MSD este de verb la prezent (Vmip), atunci accentul nu e pe terminație; b) daca eticheta este de verb la trecut (Vmis), indiferent de persoană și număr, sau la infinitiv (Vmnp), sau la viitor, atunci accentul e pe vocala finala “i”;
- tipul ALUNGI: formă verbală identică pentru două leme diferite (ALUNGA, ALUNGI) : a) daca eticheta MSD este de verb la prezent (Vmip), indiferent de persoană și număr, atunci accentul e pe vocala finală; b) daca eticheta MSD este de verb la trecut (Vmis3s), atunci accentul nu e pe vocala finală “i”.

Pentru corectarea erorilor de transcriere fonetică s-au implementat următoarele reguli:

- reguli speciale pentru grupurile ce/ci/ge/gi/che/chi/ghe/ghi, mai precis pentru transcrierea vocalei finale (“e”, “i”):

I. grup la final de cuvânt:

a. litera finală (“e”/“i”) are valoare vocalică: toată silaba e reprezentată de unul dintre grupurile vizată: literele se transcriu ca “e”, respectiv „i”;

exemple: ce (pron.) [tS e], ci (conj.) [tS i], trece [t r e tS e], ghici (vb.), [g_j i tS i], merge [m e r gZ e], amăgi [a m @ gZ i], ureche [u r e k_j e], ochi (vb) [o k_j i], veghe [v e g_j e], zbughi (vb.) [z b u g_j i].

b. litera finală (“e”/“i”) are valoare „zero” (nu se transcrie): atunci când i final asilabic este absorbit în articulația consoanei precedente; deși aflate la finală de cuvânt, aceste grupuri nu formează singure silabă; alături de ele mai există, în aceeași silabă, cel puțin o vocală.

exemple: mici [m i tS], lungi [l u n gZ], ochi (subst.) [o k_j], unghi [u n g_j i], voinici (adj. [v o i n i tS]), oblici [o b l i tS]

II. grup la final de silabă în interiorul cuvântului:

a. litera finală (“e”/“i”) are valoare vocalică, indiferent dacă este urmat de semn vocalic sau de semn consonantic în următoarea silabă; se transcrie “e”/„i”:

exemple: erbacee [e r b a tS e e], salcie [s a l tS i e], geolog [gZ e o l o g] spongios [s p o n gZ i o s], încheia [1 n k_j e j a], închina [1 n k_j i n a], gheonoaie [g_j e o n o_X a j e], ghiocel [g_j i o tS e l]

III. grup în interiorul silabei:

a. litera finală (“e”/“i”) are valoare vocalică atunci când grupul este urmat de consoană: se transcrie ca e sau i;

exemple: certa [tS e r t a], încinge [1 n tS i n gZ e] , îngemăna [1 n gZ e m @ n a], argint [a r gZ i n t], cheltui (vb. prezent) [k_j e l t u j] , chingă [k_j i n g @], chelbe [k_j e l b e], ghindă [g_j i n d @].

b. litera finală ("e"/"i") are valoare vocalică atunci când grupul este urmat de vocală neaccentuată și care nu formează silabă (este semivocală): în cazul acesta e sau i din grup se transcriu cu e, respectiv cu i, iar cealaltă vocală (cea neaccentuată care urmează) se transcrie ca semivocala corespunzătoare; (vezi regulile pentru diftongi pentru a identifica dacă litera care urmează grupul este vocală sau semivocală)

exemple: cercei [tS e r tS e j], cretaceu [k r e t a tS e w], mijlociu [m i Z l o tS i w], funigei [f u n i gZ e j], apogeu [a p o gZ e w], hangiu [h an gZ i w], închei (vb. ind. prez. I sg.) [1 n k_j e j], muchii [m u k_j i j], ochii (verb ind. perf. s. I sg.) [o k_j i j], fistichii [f i s t i k_j i j], pârghii [p 1 r g_j i j].

c. litera finală ("e"/"i") nu are valoare vocalică (sau are valoare „zero") și nu se transcrie (este absorbită în articulația consoanei precedente) atunci când grupul este urmat de o vocală care formează silabă; aici grupul este urmat imediat de o vocală, în aceeași silabă;

exemple: ceaţă [tS a ts @], aicea [a i tS a], ceapraz [tS a p r a z], picior [p i tS o r], ciozvârtă [tS o z v 1 r t @], ciumă [tS u m @], măciucă [m @ tS u c @], ciunti [tS u n t i], geană [gZ a n @], lingea [l i n gZ a], mingea [m i n gZ a], mahalagioaică [m a h a l a gZ o_X a j k @], magiun [m a gZ u n], giuvaer [gZ u v a e r], cheamă [k_j a m @], urechea [u r e k_j a]; chiar [k_j a r], ochios [o k_j o s], chiup [k_j u p], chiuli [k_j u l i], gheaţă [g_j a ts @], ghiozdan [g_j o z d a n], surghiun [s u r g_j u n], unghiul [u n g_j u l].

- **reguli transcriere diftongi (ascendenți vs. descendenți, exc. iu), triftongi, hiat,** pentru a stabili transcrierea vocalică sau semivocalică a vocalelor; cei mai mulți diftongi pot fi clasificați determinist în ascendenți (prima literă este semivocală, a doua este vocală) sau descendenți (prima literă este vocală, a doua este semivocală); excepție face diftongul "iu", care poate fi atât descendent (de exemplu în "fiu") cât și ascendent (de exemplu în "iubit"), și a fost corectat manual; de asemenea, triftongii pot fi centrați (formați din semivocală + vocală + semivocală) sau ascendenți (formați din semivocală + semivocală + vocală). Vocalele în hiat sunt vocale aflate de o parte și de alta a unor silabe, care nu sunt implicate la rândul lor în diftongi sau triftongi.

diftong	Rostire	transcriere
ai	descendent	a j
au	descendent	a w
ei	descendent	e j
eu	descendent	e w
ii	descendent	i j
oi	descendent	o j
ou	descendent	o w
ui	descendent	u j
ăi	descendent	@ j
ău	descendent	@ w

âi	descendent	1 j
âu	descendent	1 w
ea	ascendent	e_X a
eo	ascendent	e_X o
ia	ascendent	j a
ie	ascendent	j e
io	ascendent	j o
oa	ascendent	o_X a
ua	ascendent	w a
uă	ascendent	w @
uo	ascendent	w o

triftong	**rostire**	**transcriere**
eai	centrat	e_X a j
eau	centrat	e_X a w
iai	centrat	j a j
iau	centrat	j a w
iei	centrat	j e j
oai	centrat	o_X a j
ioa	ascendent	j o_X a
eoa	ascendent	e_X o_X a
uea	ascendent	w e_X a
ioi	centrat	j o j

- **regula pentru i final asurzit**: Dacă ultima silabă a unui cuvânt se termină *cu "(orice sau nimic) vocală consoană i"*, atunci „i” final se transcrie ca „i_0” (fac excepție grupurile ce/ci/ge/gi/che/chi/ghe/ghi), care respectă în acest caz regula **I.a.**
  exemplu: apropieri [a p r o p i e r i_0];

Bibliografie

Barbu, Ana-Maria. "Romanian Lexical Data Bases: Inflected and Syllabic Forms Dictionaries." LREC (2008)

Erjavec, Tomaz. "MULTEXT-East Morphosyntactic Specifications: Version 3.0." Supported By EU Projects Multext-East, Concede And TELRI (2004)

Ion, Radu. “TEPROLIN: An Extensible, Online Text Preprocessing Platform for Romanian.” In Proceedings of the International Conference on Linguistic Resources and Tools for Processing Romanian Language (ConsILR 2018), November 22-23, 2018, Iași, Romania. (2018)

Stan, Adriana, Junichi Yamagishi, Simon King, and Matthew Aylett. „The Romanian Speech Synthesis (RSS) corpus: building a high quality HMM-based speech synthesis system using a high sampling rate” In Speech Communication vol.53 442-450. (2011)

Toma, Ştefan-Adrian, et al. "MaRePhoR—An open access machine-readable phonetic dictionary for Romanian." 2017 International Conference on Speech Technology and Human-Computer Dialogue (SpeD). IEEE. (2017)