RAPORT ȘTIINȚIFIC proiect complex ReTeRom, etapa I - noiembrie 2018
Proiectul 2: TEPROLIN
Activitatea 1.8
Denumire activitate: **Crearea și validarea (eventual cu corectările manuale necesare) a unui lexicon specific corpusului bimodal și încorporarea sa în lexiconul deja existent**

Autori: ICIA

REZUMATUL ETAPEI

Această activitate a proiectului 2, TEPROLIN, are rolul de a extrage lexicoane din corpusurile orale existente la membrii proiectului, de a le corecta atunci când există erori de diverse feluri și de a le face accesibile publicului larg, pe site-ul proiectului.

Rezultatele activității: (i) Raport asupra modului de lucru și a lexicoanelor extrase din diversele sub-corpusuri; **(ii)** lexicoanele extrase și corectate sunt concatenate și puse la dispoziție public pe site-ul proiectului.

1. Importanța lexiconului în prelucrarea vorbirii

Un lexicon își dovedește utilitatea atât în recunoașterea vorbirii, cât și în sinteza sa. El conține forma scrisă a cuvintelor și pronunția lor. Este important de notat faptul că între grafie și pronunție (între litere și sunete) nu există o corespondență de 1:1. Acest lucru este valabil chiar și pentru limba română, care are o ortografie fonetică: la nivel fonetic, există litera *x* căreia îi corespund două grupuri de câte două sunete (*cs* și *gz*) și grupuri de litere cărora le corespunde un sunet sau două (*ce, ci, che, chi* etc.); la nivel lexical, există cuvinte cu pronunție diferită (deci transcriere fonetică diferită: *haină* - *h a i n ă* sau *h a j n ă*); la nivel lexico-sintactic, există cuvinte (sau părți de cuvinte) care se pronunță într-o singură silabă (*s-au*) și creează uneori chiar fenomenul de omofonie (fără omografie) (*s-au* și *sau*).

Pentru recunoașterea automată a vorbirii (RAV), Adda-Dekker și Lamel (2000) vorbesc despre un rol dublu al lexiconului:

- inventarierea cuvintelor cunoscute de sistemul de transcriere;
- mijlocul de creare a modelelor acustice pentru fiecare intrare.

Pentru sinteza vorbirii (SV), următoarele utilități ale unui lexicon au fost identificate (Liberman și Church, 1992; Klavans și Tzoukermann, 1994):

- oferă modul de citire a abrevierilor;
- suplinește existența unui corpus exhaustiv, prin includerea tuturor formelor flexionare ale unui cuvânt;
- oferă informații morfologice despre un cuvânt.

Reiese, așadar, că ambele aplicații (RAV și SV) au nevoie de lexicoane cu acoperire cât mai mare (pentru minimizarea listei de cuvinte necunoscute - eng. *out-of-vocabulary (OOV) words*), pe de o parte, la nivel lexical, iar pe de altă parte, la nivel lingvistic (tipuri de informație lingvistică reprezentată în lexicon).

În plus, crearea de resurse lingvistice este o preocupare în sine în cadrul ICIA: de aceea, îmbogățirea lexiconului existent (tblwordform), folosit până acum aproape exclusiv pentru prelucrarea limbajului scris, cu informații lingvistice utile pentru prelucrarea limbajului oral este de mare interes.

2. Crearea lexiconului specific corpusurilor bimodale existente la membrii proiectului ReTeRom

Transcrierile (componenta textuală a) corpusurilor bimodale disponibile la toți partenerii proiectului au fost colectate și analizate. Corpusul provenind de la ICIA și UAIC (**CoRoLa - oral**) conține texte provenind din RoWikipedia, știri, interviuri pe teme de actualitate, povești. Transcrierile înregistrărilor sunt tokenizate, lematizate și adnotate morfologic, dar calitatea transcrierii și implicit a adnotării nu este omogenă. S-au observat erori de ortografie și punctuație, precum și convenții diferite de transcriere (de exemplu, cuvintele în limbi străine sunt uneori transcrise așa cum se aude, alteori conform convențiilor de scriere din limba respectivă). Corpusul provenit de la UPB (grupând corpusurile **RSC, SSC-train și SSC-eval**) conține texte ce transcriu înregistrări de știri, talkshow-uri, interviuri, literatură, vorbire spontană. Textele sunt corecte ortografic, dar nu conțin majuscule și nici punctuație. În plus, o mare parte din corpus reprezintă transcrieri automate din emisiuni de știri, bazate pe segmente audio produse tot automat (vezi Cucu et al., 2015) astfel încât să filtreze secvențele non-speech și să conțină rostirea unui singur vorbitor. Transcrierile rezultate sunt arareori propoziții complete. În plus, în altă secțiune a corpusului predomină rostiri ale unor liste de cuvinte, fiecare într-un enunț separat. În consecință, cea mai mare parte a unităților de transcriere din corpusul UPB reprezintă secvențe sub-propoziționale, fără majuscule și punctuație, motiv pentru care nu se pretează unei adnotări automate cu un tagger. Corpusul provenit de la partenerul UTCN conține subcorpusurile **SWARA** (înregistrări în studio cu transcriere ortografică de calitate), **MARA** (din înregistrarea nuvelei Mara de către un vorbitor profesionist de gen feminin, cu transcriere ortografică aproximativă) și **Adevărul.ro** (provenit dintr-un corpus de text compus din articole din ziarul Adevărul - versiunea online, calitate ortografică). Pentru că toate transcrierile sunt ortografice, am putut folosi un instrument de tokenizare, lematizare și adnotare morfologică automată.

Ca rezultat al analizei transcrierilor disponibile, am adoptat două metodologii de lucru, una pentru corpusul UPB care nu dispune de segmentare, lematizare și adnotare automată și alta pentru celelalte corpusuri.

Extragerea lexiconului din transcrierile UPB

Așa cum am menționat, transcrierile din acest corpus nu conțin punctuație sau majuscule. De aceea, sarcina de segmentare se trivializează: am făcut segmentare automată la nivel de spațiu și am dedicat cuvintelor cu cratimă o abordare separată. Acestea au fost segmentate la nivel de cratimă, iar cratima a fost lipită succesiv de primul sau al doilea (cu rare cazuri de 3 cuvinte)

generând diferite variante de combinare posibile. De exemplu, pentru "n-am", s-au generat perechile de cuvinte ("n-", "am") și ("n" si "-am"). Lematizarea și adnotarea morfologică a cuvintelor din acest corpus se face prin apel la un lexicon dezvoltat la ICIA (tbl.wordform.ro, denumit în continuare TBL, pentru a-l distinge de lexiconul specific corpusului bimodal la care lucrăm în această etapă a proiectului), cu peste 1.150.000 de intrări, de forma: *<formă ocurentă>tab<lemă>tab<etichetă morfo-sintactică>*.

Pentru cuvintele care provin din segmentarea cuvintelor cu cratimă, perechea corectă corespunzătoare segmentării poate fi identificată prin apel la TBL (ex. pentru "schimbându-și", perechea (schimbându-, și) nu se regăsește în TBL pentru că acesta nu conține "schimbându-", deci singura variantă de segmentare corectă este (schimbându, -și)) sau poate fi nevoie de validare manuală, atunci când toate cele 4 cuvinte implicate în segmentare (de exemplu "n-", "am", "n" și "-am", pentru secvența "n-am") pot fi regăsite în TBL.

Pentru cuvintele provenind din segmentare la spațiu, recuperarea lemei și a etichetei morfosintactice (MSD, http://nl.ijs.si/ME/Vault/V3/msd/html/) se face tot prin căutare în TBL. Dacă un cuvânt este găsit în TBL, se afișează toate lemele și etichetele asociate; datorită omonimiei în limbă, în anumite situații perechea (lemă, MSD) asociată unei forme ocurență nu poate fi dezambiguizată decât manual, în context, lucru pe care deocamdată nu ni l-am propus. Pentru fiecare pereche (lemă; MSD), întreaga paradigmă morfologică este recuperată ulterior din TBL.

Dacă cuvântul de interes nu este găsit în TBL, este extras separat într-o listă care necesită lematizare și adnotare manuală (vezi secțiunea 3, Validarea lexiconului). Pentru a automatiza (și implicit ușura) o parte din munca de validare, un tratament special între cuvintele negăsite în TBL este acordat celor care încep cu prefixul „nemai", verbe la modul participiu și gerunziu. Prefixul "nemai" a fost separat automat, iar forma rămasă (denumită rădăcină în acest context) a fost căutată la rândul ei in TBL:

- dacă rădăcina a fost găsită, s-au extras toate formele ei de participiu și gerunziu din lexicon și s-a adăugat automat prefixul "nemai"; toate formele au fost validate manual și paradigma a fost completată, când a fost cazul;
- dacă rădăcina nu a fost găsită în lexicon, am observat că toate formele sunt de gerunziu; s-a "ghicit" o lemă, îndepărtând sufixul "nd" iar lemele au fost corectate manual; s-au generat automat formele de participiu și gerunziu pornind de la aceste leme si au fost corectate manual;

Extragerea lexiconului din transcrierile RACAI și UTCN

Pentru aceste corpusuri, metoda de extragere a lexiconului este o variantă simplificată a celei de mai sus. Căutăm în TBL toate tripletele (formă, lemă, MSD) așa cum sunt ele disponibile în corpus și le extragem pentru lexiconul nostru. Dacă nu le găsim în TBL, le căutăm printre intrările în lexiconul TEPROLIN creat anterior din corpusul UPB. Dacă nu se regăsesc nici aici, sunt extrase pentru corectare/validare manuală (vezi secțiunea 3. Validarea lexiconului)

3. Validarea lexiconului

Această etapă de lucru presupune evaluarea fiecărui cuvânt și, dacă acesta este cuvânt corect în limba română (i.e. există sau este o creație ad-hoc posibilă conform regulilor morfologice ale limbii, are toate diacriticele necesare), stabilirea a trei elemente:

- lema sa (forma de dicționar);
- descrierea morfosintactică (PoS tag);
- restul paradigmei flexionare: pentru fiecare formă din paradigmă se notează lema și descrierea morfosintactică.

Erorile frecvent întâlnite și care au dus la eliminarea cuvintelor găsite în corpus din lexiconul creat au fost:

- scrieri greșite: litere inversate, litere lipsă, litere în plus;
- probleme cu diacriticele: lipsa totală a acestora sau existența parțială;
- probleme de segmentare: scrierea împreună a două cuvinte, introducerea unui blanc în interiorul cuvântului.

4. Îmbogățirea lexiconului cu segmentarea în silabe, marcarea accentului și transcrierea ortografică

Lexiconul a fost completat cu informația de segmentare în silabe, de accent și cu transcrierea ortografică apelând la instrumentul descris anterior, Romanian TTS, furnizat de UTCN (mai precis modulul de procesare textuală al acestuia). De exemplu, pentru lista de cuvinte (separate printr-un enter, “copilul”, “plânge”, “neconsolat”), TTS returnează[1]:

SYLLABIFICATION:
['co.pi.lul', [0, 1, 0, 1, 0, 0, 0]]
['plân.ge', [0, 0, 0, 1, 0, 0]]
['ne.con.so.lat', [0, 1, 0, 0, 1, 0, 1, 0, 0, 0]]

ACCENT
["cop'ilul", [0, 0, 0, 1, 0, 0, 0, 0]]
["pl'ânge", [0, 0, 1, 0, 0, 0]]
["neconsol'at", [0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0]]

PHONETIC TRANSCRIPTION
copilul ['k', 'o', 'p', 'i', 'l', 'u', 'l']
plânge ['p', 'l', 'a@', 'n', 'dz', 'e']

[1] Pentru silabificare și marcarea accentului, TTS are două formate de prezentare a informației:1. Separare în silabe prin punct sau marcare prin apostrof (înainte de vocală) a poziției din cuvânt la care apare accentul; 2. Codificare prin vectori binari, unde poziția cifrei 1 in vector marchează poziția de segmentare sau poziția accentului în cuvânt.

neconsolat ['n', 'e', 'k', 'o', 'n', 's', 'o', 'l', 'a', 't']

Aceste informații sunt integrate în lexicon, astfel încât fiecare intrare are ca model pe cea din exemplul următor:

(1) copilul*(tab)*copil*(tab)*Ncmsry *(tab)*['co.pi.lul', [0, 1, 0, 1, 0, 0, 0]]*(tab)*["cop'ilul", [0, 0, 0, 1, 0, 0, 0, 0]]*(tab)*['k', 'o', 'p', 'i', 'l', 'u', 'l']

Formatul generalizat al intrărilor din lexiconul ReTeRom este:

<formăocurentă>tab<lemă>tab<etichetămorfo-sintactică>tab<silabificare>tab<accent>tab<transcriere fonetică>

5. Lexiconul concatenat rezultat

În urma concatenării datelor rezultate direct prin recuperare din TBL, s-au obținut 346.074 intrări de forma (1). În plus, au fost corectate/create și integrate în lexicon 8000 intrări noi (care nu se regăseau în TBL). Ca urmare a procesului de corectare a corpusurilor (care va avea loc într-o etapă ulterioară a proiectului), lexiconul va fi îmbogățit cu noi intrări corectate.

Referințe:

M. Adda-Decker, L. Lamel, The use of lexica in automatic speech recognition. In F. van Eynde & D. Gibbon (Eds.), *Lexicon Development for Speech and Language Processing*, Kluwer Academic Publishers, 2000.

J.L. Klavans, E. Tzoukermann, Machine-readable Dictionaries in Text-to-speech Systems. In *Proceedings of the 15th Conference on Computational Linguistics* - Volume 2 (COLING), 1994, p. 971-975.

M. Y. Liberman, K. W. Church, Text Analysis and Word Pronunciation in Text-to-speech Synthesis. In S. Furui, M.M. Sondhi (eds.), *Advances in Speech Signal Processing*, New York: Marcel Dekker, 1992, p. 791-831.

Horia Cucu, Andi Buzo, Laurent Besacier, Corneliu Burileanu, Enhancing ASR Systems for Under-Resourced Languages through a Novel Unsupervised Acoustic Model Training Technique, in *Advances in Electrical and Computer Engineering*, vol. 15, no. 1, pp. 63-68, Feb 2015, ISSN: 1582-7445, doi:10.4316/AECE.2015.01009