RAPORTARE ȘTIINȚIFICĂ

Proiect complex ReTeRom. Proiect component CoBiLiRo

Activitatea A2.4: *Armonizarea colecțiilor existente* (la sfârșitul celui de-al doilea an, colecțiile partenerilor trebuie să se găsească pe platforma nou creată și să fie disponibile acolo)

Faza de predare: noiembrie 2019

Autori: Anca-Diana Bibiri, Daniela Gifu, Mihaela Onofrei, Diana Trandabăț

1. Rezumatul etapei

A doua etapă (2019) a proiectului CoBiLiRO prevede realizarea infrastructurii pentru gestionarea și stocarea corpusului bimodal. Activitățile complementare prevăzute includ descrierea și implementarea unor soluții de armonizare a reprezentărilor colecțiilor existente text/vorbire (metadata și adnotări). Ulterior, aceste soluții urmează să fie utilizate pentru a armoniza formatele colecțiilor existente și a le încărca pe platforma CoBiLiRO. Similar celorlalte etape, este prevăzută o activitate de diseminare atât la evenimente științifice (conferințe, stagii de practică, ateliere de lucru) cât și în mass-media (comunicate de presă). De interes special pentru platforma dezvoltată este respectarea drepturilor de autor și a anonimizării solicitate pentru contributorii de resurse pe platformă.

2. Rezumatul activității

În această etapă, este prevăzută încărcarea pe platforma CoBiLiRo[1] a resurselor partenerilor din Consorțiul ReTeRom și realizarea concordanței/standardizării colecțiilor existente de corpusuri bimodale (text/vorbire) în funcție de soluțiile de armonizare a reprezentărilor acestora (metadate și adnotări) stabilite de comun acord în cadrul consorțiului.

3. Descrierea științifică și tehnică

Resursele existente până în prezent pe platforma CoBiLIRo (a se vedea Fig. 1) sunt: CoRoLa-IIT, CoRoLa-IIT_2, CoRoLa_RASC, SWARA, SoRoEs, MARA, RoDigits, Rador_2.

Resursele încărcate pe platforma CoBiLiRo respectă formatele agreate de comun acord de aliniere text-vorbire: tipul *file* pentru fișierele audio, însoțite de textele corespunzătoare, cel de-al doilea tip, *start-stop*, care comportă un singur fișier, și tipul *file-start-stop*.

[1] http://85.122.23.18:81/

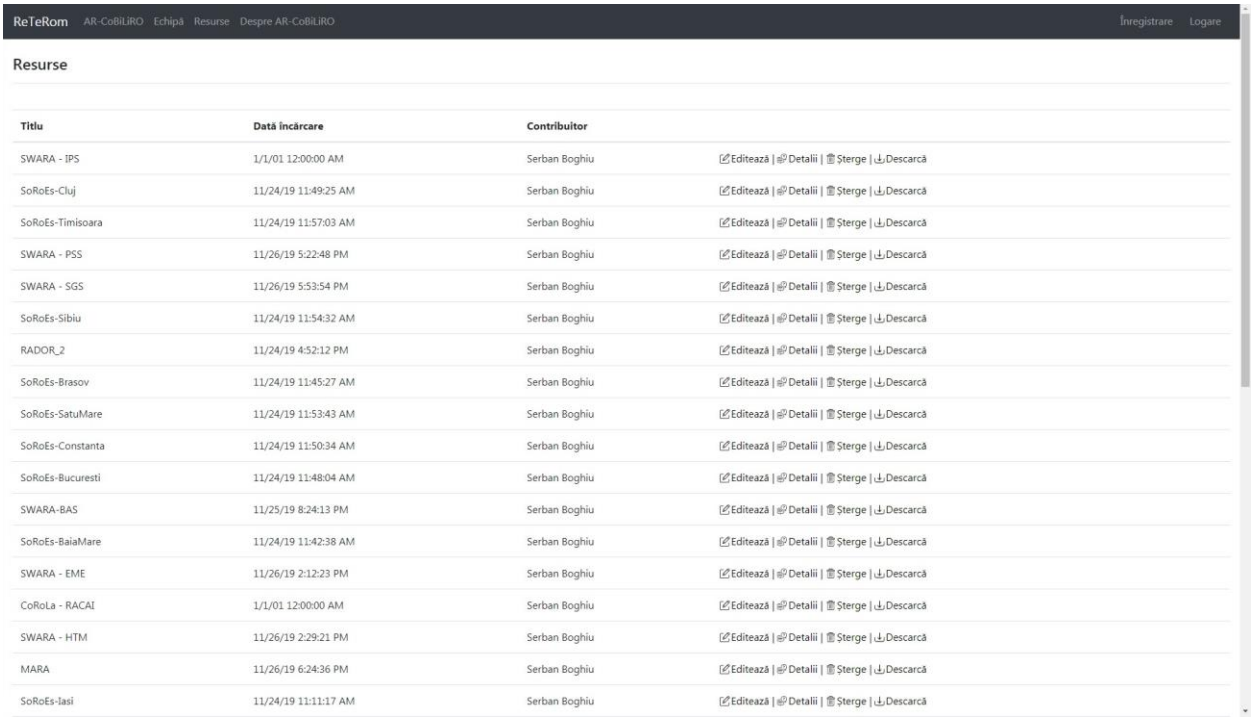

Figura 1: Interfața platformei CoBiLiRo la încărcarea unei resurse

Cele trei formate identificate în resursele partenerilor sunt:

a. FORMATUL PHS/LAB care conține patru fișiere cu extensiile:
    1. *.wav* (înregistrarea secvenței audio),
    2. *.txt* (transcrierea textului, rostit în cadrul înregistrării audio),
    3. *.phs* (aliniere la nivel de fonem a textului transcris),
    4. *.lab* (varianta procesată a textului din fișierul *.txt*, din care s-au eliminat semnele de punctuație).

b. FORMARTUL MULTEXT/TEI (dezvoltat în cadrul proiectului de cercetare științifică *Multilingual Text Tools and Corpora*) are în componență:
    1. un fișier *.xml* (care conține o serie de metadate și varianta text asociată undei sonore) și
    2. un set de fișiere cu înregistrări audio cu extensia *.wav;*

c. FORMATUL TEXTGRID - cele trei fișiere asociate acestui format sunt:
    4. *.wav* (reprezintă înregistrarea audio),
    5. *TextGrid* (alinierea elementelor vocalice segmentate în PRAAT, cu timpul de început și timpul de sfârșit exprimate prin *xmin* și *xmax*),
    6. *.txt* (parametrii acustici ai fiecărei vocale: durata (ms), energia (dB) și frecvența fundamentală (extrasă în câte trei puncte) (Hz).

Aceste formate au fost descrise exhaustiv în rapoartele anterioare, în special în Activitatea A2.2: *Raport asupra soluțiilor de armonizare a reprezentărilor colecțiilor existente text /vorbire (metadate și adnotări)* și în Activitatea A2.3: *Realizarea de convertoare de format pentru armonizarea diferitelor reprezentări ale partenerilor la reprezentarea standard agreată în consorțiu* (proiectate și implementate de membrii proiectului CoBiLiRo).

În ceea ce privește respectarea Legii drepturilor de autor, pentru realizarea înregistrărilor audio-video s-au încheiat acorduri de colaborare și au fost semnate fișele de consimțământ ale subiecților intervievați pentru respectarea confidențialității ca datele înregistrate să fie utilizate strict în scopul cercetării științifice.

De exemplu, în cazul înregistrărilor SoRoEs, fiecărui subiect i s-a atribuit un cod de identificare (pentru respectarea legii dreptului de proprietate), iar etichetarea unui enunț cuprinde 4 simboluri, la care am adăugat și simbolul pentru domeniul limbii române folosit în alte proiecte (AMPER și AMPRom), reprezentat de cifra 9: pentru punctul de anchetă respectiv, codul subiectului în funcție de gen, vârstă și nivelul de studii, codul enunțului – cu un simbol alcătuit din cifre și litere și numărul enunțului înregistrat (1, 2, 3, ...). De exemplu, 9I5c_86a reprezintă enunțul 86, rostit de subiectul de gen feminin, cu vârsta peste 50 de ani, având studii superioare, din localitatea Iași, domeniul limbii române. Aceste informații, astfel codificate pentru necesitățile proiectelor de origine (AMPER și AMPRom), au fost însă explicitate și în metadatele asociate resurselor.

7. Concluzii

Toate resursele încărcate până în acest moment în Platforma CoBiLiRo au fost trecute prin programele de conversie de format, ceea ce le face compatibile cu standardul CoBiLiRo agreat în consorțiu.

Toate obiectivele incluse în plan la această activitate au fost realizate.

Bibliografie

[1] Pistol. I., Pădurariu C., Boghiu Ș., Scutelnicu A., Raport Activitatea A2.2: *Raport asupra soluțiilor de armonizare a reprezentărilor colecțiilor existente text /vorbire (metadate și adnotări),* proiectul ReTeRom.

[2] Ionuț Pistol, Andrei Scutelnicu, Cristian Pădurariu, Șerban Boghiu**,** Activitatea A2.3: *Realizarea de convertoare de format pentru armonizarea diferitelor reprezentări ale partenerilor la reprezentarea standard agreată în consorțiu* (noiembrie 2019) proiectul ReTeRom.

[3] Paul Boersma și David Weenink (2018), *Praat: doing phonetics by computer [Computer program]***.** Version 6.0.37, retrieved on 15 November 2019 from http://www.praat.org/.