

# RAPORTARE ȘTIINȚIFICĂ

## Proiect complex ReTeRom. Proiect component CoBiLiRo

### *Activitatea 3.3 - Realizarea de statistici privind corpusul bimodal voce/text*

Faza de predare: noiembrie 2020

Autori: Pistol Ionuț, Scutelnicu Andrei, Serban Boghiu, Felix Cristian Pericică

#### 1. Rezumatul etapei

În această etapă a proiectului complex ReTeRom consorțiul și-a propus să consolideze și apoi să exploateze rezultatele acumulate în primii doi ani, cu obiectivele: existența unui Portal pregătit a primi și prelucra resurse bilingve românești, dezvoltarea în continuare a unei colecții de resurse care să corespundă formatului agreat de consorțiu, perfecționarea lanțurilor de prelucrări lingvistice și sonore, atât asupra componentelor textuale cât și vocale ale resurselor bimodale, care să permită alinieri între componentele vocale și textuale, recunoașterea cu minimum de erori a vocii, generarea expresivă a vocii și antamarea de aplicații bazate pe aceste tehnologii.

A treia etapă (2020) a proiectului CoBiLiRO prevede completarea inventarului de resurse disponibile pe portal cât și valorificarea lor atât în cadrul platformei (statistici și instrumente integrate) precum și în afara platformei, propunând o serie de proiecte ce utilizează tehnologiile dezvoltate în cadrul proiectelor partenere. Similar celorlalte etape, este prevăzută și o activitate de diseminare, atât la evenimente științifice cât și în mass-media. De interes special pentru platforma dezvoltată este respectarea drepturilor de autor și a anonimizării solicitate pentru contribuitorii de resurse pe platformă.

#### 2. Rezumatul activității

Activitatea 3.3 are ca obiectiv dezvoltarea platformei CoBiLiRo prin oferirea unor statistici privind resursele disponibile, atât la nivel individual cât și la nivelul întregului corpus. Statisticile prezintă, la cererea utilizatorului, parametri calculați din exploatarea informațiilor conținute în fișierele text și audio, în alinierea dintre cele două, cât și în metadatele aferente.

#### 3. Descrierea științifică și tehnică

##### 3.1 Scurtă descriere a tehnologiilor relevante

Componenta de procesare lingvistică necesară obținerii de valori statistice, este realizată sub forma unui script Python, încadrat într-un web API, care poate fi accesat folosind principiile arhitecturii Restful, apelat doar la nivel intern (pe server); aici s-a folosit tehnologia **Flask Restful**.

Pentru realizarea prelucrării și procesării la nivel de limbaj natural a textelor a fost utilizată librăria open-source **Spacy**, mai exact modelul **ro\_core\_news\_sm**; acest model atribuie vectori de tokeni specifici contextului și etichete ale părților de vorbire având o acuratețe de 95.62%. Ca alternativă poate fi folosită componenta de analiză a textului produsă de TEPROLIN; motivul pentru care nu a fost inclusă în versiunea inițială este că nu toate fișierele de pe platformă au fost procesate de acea componentă. Pentru finalul proiectului avem în plan tranziția spre tehnologia TEPROLIN.

### 3.2 Statistici la nivelul întregii colecții de resurse

La nivelul întregii colecții de resurse se pot aplica mai multe tipuri de filtre, ele întorcând valori numerice rezultate din numărarea resurselor pentru care filtrele întorc valori *true*. Utilizatorul poate afla numele fișierelor pentru care filtrele dau valoarea *true*:

- Pe tipul de segmentare a fișierelor audio:
  - **“File”** - un fișier audio corespunde unei zone din text. Alinierea se face prin marcajul *file* ce indică fișierul audio corespunzător unei zone din text.
  - **“Start-Stop”** - un singur fișier audio corespunde întregului text. Alinierea se face prin marcaje *start* și *stop* ce indică momentul de început și de sfârșit din înregistrarea conținută în fișierul audio ce corespunde unei zone din text.
  - **“File-Start-Stop”** - mai multe fișiere audio corespund întregului text. Alinierea se face prin marcajul *file* ce indică fișierul audio corespunzător și prin marcajele *start* și *stop* ce indică momentul de început și de sfârșit din înregistrarea conținută în fișierul audio ce corespunde unei zone din text.
- Pe tipul de fișier audio disponibil:
  - **“wav”** (Waveform Audio File Format) - calitate foarte bună dar dimensiune mare.
  - **“mp3”** (MPEG Audio Layer 3) - calitate medie, dimensiune scăzută.
- Pe nivelul de adnotare:
  - **“Propoziție”** - alinierea dintre audio și text se face la nivel de propoziție.
  - **“Cuvânt”** - alinierea dintre audio și text se face la nivel de cuvânt.
  - **“Fonem”** - alinierea dintre audio și text se face la nivel de fonem.
  - **“Prozodie”** - alinierea dintre audio și text marchează pe text adnotările de prozodie din semnalul sonor.

Pe lângă filtrele de mai sus platforma include și statistici ce indică numărul de resurse disponibile pentru fiecare tip menționat mai sus. De asemenea este calculată și durata totală a fișierelor audio de fiecare tip, folosind *ffprobe*<sup>1</sup> - un multimedia stream analyzer.



Fig.1: Statistici generale

### 3.3 Statistici la nivelul fișierelor text

Înainte de a afișa statisticile pe interfață, fișierele text asociate resurselor sunt prelucrate și procesate pentru a obține rând pe rând următoarele componente:

- **Densitatea lexicală** a textului transcris asociat resursei - după procesul de tokenizare, se aplică etichetarea părților de vorbire având grijă ca tokenii să fie clasificați în două mari grupe (cuvinte lexicale și cuvinte funcționale); rezultatul final este dat de factorii din grupa tokenilor lexicali, prin formula:  $nr. \text{ de cuvinte lexicale} / nr. \text{ total de cuvinte} * 100$ ;
- **Raportul type/token** - după ce se calculează numărul total de cuvinte, se grupează cele care apar o singură dată, în forma flexionară; rezultatul final este dat de formula:  $nr. \text{ cuvintelor care apar o dată} / nr. \text{ total de cuvinte} * 100$ ;
- **Numărul de tokeni**;
- **Numărul de lemme** - se aplică procesul de lematizare și o lemă se ia în considerare doar o singură dată;
- **Numărul formelor flexionare distincte** - se aplică o procedură similară ca și în cazul numărului de lemme;
- **Media formelor flexionare distincte** - este dată de formula  $nr. \text{ total de forme flexionare distincte} / nr. \text{ total de tokeni}$ ;
- **Media de lemme din text** - componentă ce returnează rezultatul calculului:  $nr. \text{ de lemme distincte} / nr. \text{ total de tokeni}$  ; se iau în considerare toate lemmele din text, numărate o singură dată;

<sup>1</sup> <https://ffmpeg.org/ffprobe.html>

- **Numărul de lemme care apar de mai multe ori** - după ce se aplică procesul de lematizare, se grupează lemmele în două categorii: cele care apar o singură dată (*hapax legomenon*) și cele care apar de mai multe ori; este luată în considerare ultima grupă;
- **Raportul de lemme care apar de mai multe ori** - redat prin formula  $nr. \text{ de lemme care apar de mai multe ori} / nr. \text{ total de lemme luate în considerare o singură dată} * 100$ . În abordarea prelucrărilor lingvistice ulterioare se va avea în vedere și inventarierea lemmelor unicate, urmând ca pe interfață să fie afișată o nouă componentă - *hapax legomenon*.

Toate fișierele sunt curățate în prealabil de semnele de punctuație și eventualele spații și/sau simboluri nejustificate.

### 3.4 Statistici la nivelul fișierelor audio

Din cauza numărului mare de fișiere disponibile în platformă și a dimensiunii ridicate a acestora, taskul care calculează durata fișierelor a trebuit paralelizat (folosind librăria *TPL*<sup>2</sup>) și totodată aplicată o *tehnică de caching*, deoarece calculul total al duratelor este un proces foarte costisitor. Prin urmare, am decis utilizarea unui *Background Job Worker* (folosind librăria *HangFire*<sup>3</sup>), care recalculează zilnic durata totală a fișierelor încărcate pe platformă.

Pentru a putea accesa metadatele interne ale înregistrărilor și a afla durata acestora (mp3, mpeg, mp4, wav), a fost folosit analizatorul de stream-uri *ffprobe*. Cu ajutorul lui, am reușit să extragem durata fiecărui fișiere sunet.

Una dintre provocările întâlnite a fost configurarea lui pe serverul CentOS și integrarea lui în mediul de dezvoltare .NET Core. A fost necesară modelarea și implementarea unor optimizări pentru a face față problemelor generate de memoria aflată la dispoziție și de numărul maxim de fișiere ce puteau fi deschise simultan.

### 3.5 Statistici la nivelul alinierilor

Pentru alinierea la nivel de cuvânt, cele mai relevante în majoritatea aplicațiilor care valorifică alinieri (antrenarea *text-to-speech* și *speech-to-text*, aliniatoare automate), am pregătit un set de statistici ce pot fi generate automat. Scopul principal al acestora este oferirea unei priviri de ansamblu asupra calității alinierilor disponibile. Trebuie remarcat faptul ca această calitate depinde nu numai de procesul de aliniere (fie automată sau manuală) cât și de calitatea versiunilor paralele text și voce. Pentru o discuție mai detaliată despre posibilele probleme, a se vedea raportul *A3.4 Proiectare de aplicații de exploatare a corpului bimodal și a tehnologiilor de prelucrare textuale și voce, create în proiectele P2, P3, P4*.

---

<sup>2</sup> <https://docs.microsoft.com/en-us/dotnet/standard/parallel-programming/task-parallel-library-tpl>

<sup>3</sup> <https://www.hangfire.io>

File	AutoText	AlignedText	LongestUnmatch	StartsAt	SingleUnmatched
Alma-Mater-Iasiensi-24-Facultatea-de-Drept	5379	3761	44	30.51	193
Alma-Mater-Iasiensis-1-Aniversarile-Universitatii	7078	4400	121	2944.26	248
Alma-Mater-Iasiensis-11-Prof-Univ-Dr-Andrei-Margha	7072	4947	65	2688.87	266
Alma-Mater-Iasiensis-12-Facultatea-de-Fizica	5232	3312	52	3282.42	166
Alma-Mater-Iasiensis-13-Universitatea-din-Iasi-prezent-si-perspective	7006	5220	61	2109.15	281
Alma-Mater-Iasiensis-14-Universitatea-din-Iasi-Dimensiunea-culturala	6301	4558	41	2090.97	265

Fig.2 : Statistici pe alinieri la nivel de cuvânt

În scopul obținerii acestor statistici a fost utilizată alinierea produsă de TADARAV ce aliniază textul extras și aliniat automat de tehnologia TADARAV cu transcrierea manuală disponibilă pentru resursa respectivă [1].

În figura 2 semnificația coloanelor este:

- *AutoText* - dimensiunea (în cuvinte) a textului extras automat din fișierul audio;
- *AlignedText* - dimensiunea (în cuvinte) a textului aliniat cu transcrierea originală;
- *LongestUnmatch* - cea mai lungă secvență de cuvinte din transcrierea originală pentru care nu există nici o aliniere;
- *StartsAt* - indexul cuvântului de la care începe secvența de mai sus;
- *SingleUnmatched* - numărul de cuvinte din transcrierea originală care nu au fost aliniate, deși cuvântul imediat precedent și cel imediat ulterior au fost aliniate.

O serie de observații pot fi făcute pe aceste statistici, de exemplu se observă că procentul cuvintelor alinate corect (pe cele 6 fișiere din figura 2) este de aproximativ 70%. Se observă că în fișierul al doilea există o zonă lungă de cuvinte nealinate (121 cuvinte) ce poate indica prezența în transcrierea manuală a unui fragment care nu se regăsește și în înregistrarea audio. O altă observație ce ar putea fi făcută pe aceste statistici ar fi că în fișierul 5 există un număr relativ mare de cuvinte izolate nealinate (281) ceea ce ar putea indica o calitate scăzută a semnalului audio sau un accent/prozodie nestandard a vorbitorului.

Aceste statistici sunt implementate și în curs de integrare cu platforma online, versiunea finală a platformei, ce va fi predată în aprilie 2021, va include și aceste statistici.

### 3.6 Concluzii

Oferirea unor statistici relevante pentru o colecție de resurse contribuie atât la mărirea încrederii unor potențiali noi contributory cât și la potențialul resurselor din colecție de a-și găsi utilitatea în diverse contexte. Statisticile deja incluse pe platformă precum și cele care vor fi incluse până în aprilie 2021 promit să asigure acest avantaj pentru colecția de resurse bimodale, principalul rezultat al proiectului CoBiLiRO.

Până la terminarea proiectului avem în vedere ca procesele de determinare a lungimii fișierelor să fie lansate doar asupra fișierelor nou introduse sau updatate. Serverul e capabil să urmărească fișierele asupra cărora s-a umblat și să calculeze/recalculeze lungimile doar pentru acestea. Odată aflată o lungime, ea se va înscrie în metadate, ceea ce înseamnă că nu va trebui recalculată la fiecare cerere de lungimi din partea utilizatorului.

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

#### **Bibliografie**

[1] C. Burileanu, D. Burileanu, H. Cucu (2019). Raportul aferent activității A2.11 *Proiectarea și implementarea unei soluții de bază de filtrare și aliniere a transcrierilor aproximative cu semnalul de vorbire*, proiectul RETEROM