RAPORT ȘTIINȚIFIC etapa I – 2018

A1.2. Inventarierea colecțiilor de date lingvistice românești disponibile la parteneri sau în terțe coaliții

REZUMATUL ETAPEI

Prima etapă a proiectului 1 COBILIRO prevede activități premergătoare realizării platformei de resurse audio și textuale, care au ca principal obiectiv identificarea convențiilor de adnotare optime, prin inventarierea corpusurilor multimodale existente la parteneri și la nivel internațional, precum și prin armonizarea formatelor de reprezentare, adnotare și metadate.

1. Introducere

Denumire activitate: 1.2. Inventarierea colecțiilor de date lingvistice românești disponibile la parteneri sau în terțe coaliții și a formatelor de stocare a acestora.

Rezultate Etapa:

În această etapă au fost inventariate resursele lingvistice existente la partenerii din cadrul proiectului complex, precum și la terțe coaliții. Fiecare resursa a fost descrisă conform unui set de trăsături unitar, cu scopul de a facilita identificarea unei convenții de adnotare optime pentru desfășurarea proiectului.

Au fost inventariate 11 resurse, în marea majoritate multimodale, cu următoarea distribuție per partener:

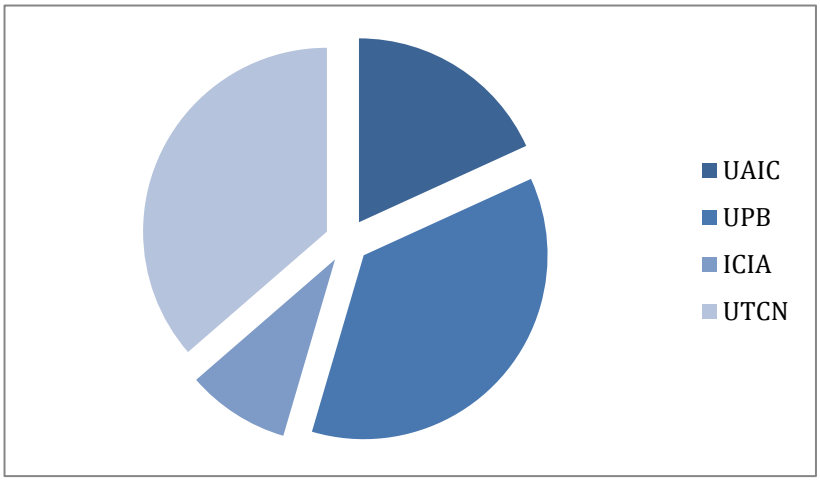

Fig. 1. Distribuția corpusurilor inventariate per partener

2. Descrierea colecțiilor de date lingvistice

Resursele lingvistice sunt colecții de date de limbaj, scris sau vorbit, însoțite de o descriere, într-un format care poate fi citit de o mașină, folosit pentru construirea, îmbunătățirea sau evaluarea

algoritmilor sau sistemelor de limbaj natural și de vorbire. Exemple de resurse lingvistice sunt corpurile scrise și vorbite, lexicoane computaționale, ontologii, baze de date terminologice, colecții de vorbire etc. Instrumentele de bază de prelucrarea vorbirii sunt de asemenea esențiale pentru achiziționarea, pregătirea, colectarea, gestionarea, personalizarea și utilizarea acestor resurse lingvistice.

Cercetarea lingvistică fundamentală și aplicată are o lungă istorie de generare și utilizare a resurselor de limbaj textuale și, mai recent, a resurselor multimodale. Aceste resurse sunt utilizate pentru o varietate de scopuri: lingviștii le folosesc pentru a crea și a testa noi ipoteze lingvistice; informaticienii le folosesc pentru a testa programele de prelucrarea limbajului scris și vorbit și stabilirea parametrilor pentru algoritmii de învățare. În ultima perioadă au fost investite sume mari de bani pentru crearea de noi resurse lingvistice sau extinderea resurselor existente astfel încât să includă o varietate de intrări și adnotări multimodale (text, sunet, video, urmărirea gesturilor sau a mișcării ochilor etc.).

Aceste resurse lingvistice au fost descrise de obicei în temeni care exprimau caracteristicile lor, precum „resursa include înregistrări ale unui subiect de 10 ani, sex, născut în mediu urban” sau „această resursă include propoziții însoțite de gesturi înregistrate când oamenii au fost rugați să dea informații despre cum se poate ajunge la gară”. Aceste descrieri sunt numite metadate, și sunt folosite pentru a caracteriza succint conținutul resursei. Majoritatea resurselor lingvistice includ aceste metadate fie ca parte a resurselor în sine, fie în fișiere separate, într-un format specific fiecărui corpus. Astfel, fiecare corpus și-a definit propria structură de metadate, adecvate obiectivelor sale.

În era web-ului semantic, se dorește armonizarea modului de descriere a metadatelor, cu scopul de a facilita identificarea conținutului resurselor și descoperirea resurselor relevante fiecărui task prin urmărirea unui standard pentru structura și semantica acestor meta-descrieri. Standardul Dublin-Core [1], cel mai cunoscut standard de descriere a resurselor, a fost definit de comunitatea de bibliotecari pentru a descrie resursele din librării. Ulterior, a devenit formatul de referință pentru postarea resurselor multimodale pe platforma Europeana[1].

3. Trăsături ale colecțiilor de date

Inventarierea colecțiilor de date lingvistice existente la parteneri s-a realizat având în vedere un set de trăsături, în acord cu standardul internațional de descriere a resurselor Dublin Core. Setul de metadate Dublin Core conține cincisprezece trăsături generice, utile pentru descrierea unei mari varietăți de resurse. Numele “Dublin” se datorează originii sale, ideea acestui standard luând naștere la un atelier din Dublin, Ohio, în anul 1995. Cele cincisprezece elemente din Dublin Core fac parte dintr-un set mai larg de metadate și specificații tehnice stabilite de Dublin Core Metadata Initiative (DCMI). Setul complet, DCMI-TERMS, include de asemenea seturi de clase de resurse, scheme de codificare a metadatelor și a sintaxei. Aproape toate trăsăturile din Dublin Core au fost preluate în descrierea resurselor partenerilor (mai puțin trei trăsături), și sunt:

- Tipul colecției de date (vorbire/text/bimodal)
- Titlul colecției, numele dat resursei de către creator sau de cel care a publicat-o

[1] Europeana este o platformă online care oferă acces la peste 50 de milioane de articole digitizate: cărți, muzică, lucrări de artă și multe altele, utilizând instrumente de căutare și filtrare avansate, pentru a împărtăși moștenirea culturală în scopuri educative, de cercetare și divertisment.

- Nume contribuitor: partener care pune această resursă la dispoziția consorțiului
- Creator: entitatea principală (persoană sau organizație) care a fost responsabilă de crearea resursei
- Descriere: o descriere generală a resursei, a aplicabilității sau a modului recomandat de folosire
- Data: Data creării resursei sau a distribuirii ei
- Topic/Domeniu: cuvinte cheie care descriu conținutul resursei
- Limba: în ce limba este resursa
- Sursa: de unde provine textul/vorbirea care compun resursa.
- Format: modul de reprezentare folosit în resursa lingvistică (este detaliat în ultima secțiune)
- Aliniere: nivel aliniere text/voce, sau aliniere cu alte resurse, dacă este cazul.
- Drepturi: drepturile de utilizare cu care vine resursa

În plus, au fost introduse trei trăsături diferite, specifice resurselor lingvistice adnotate, astfel:

- Adnotări specifice/niveluri de adnotare, unde poate fi inclusă o descriere a etichetelor folosite pentru adnotare.
- Dimensiune: statistici privind numărul de cuvinte/fraze/ore de înregistrare/etc.
- Localizarea corpusurilor: unde pot fi găsite, online sau offline.

Fiecare partener a identificat resursele lingvistice proprii, precum și cele la care poate avea acces prin terțe coaliții, și a completat tabele de descriere pentru fiecare resursă în secțiunea de mai jos.

4. Inventarul de resurse

Tipul colecției	Corpus bimodal
Titlul colecției	CoRoLa - oral
Nume contribuitor	Institutul de Cercetări pentru Inteligență Artificială “Mihai Drăgănescu”
Creator	Institutul de Cercetări pentru Inteligență Artificială “Mihai Drăgănescu”
Descriere	Colecție de înregistrări de voce reprezentând știri, povești, interviuri, însoțite de transcrieri.
Data	2017
Topic/Domeniu	RoWikipedia, știri, interviuri pe teme de actualitate, povești
Limba	Română
Sursa	Înregistrări radio, înregistrări de către amatori voluntari, înregistrări în studio cu amatori și profesioniști
Format	Fișierele audio au extensia wav, transcrierile au extensia .txt și .lab, iar adnotările au extensia .phn.
Adnotări	Transcrierile înregistrărilor sunt tokenizate, lematizate, adnotate morfologic, descompuse în silabe.
Aliniere	Transcrierile sunt aliniate cu semnalul vocal la nivel de propoziție, cuvânt și fonem.
Drepturi	Institutul de Cercetări pentru Inteligență Artificială “Mihai Drăgănescu”

Dimensiune	Fișierele audio prelucrate totalizează 151h 57’ 21’’. Transcrierile conțin ~800.000 cuvinte.
Localizare	Pentru interogare, resursa poate fi consultată la adresa: http://89.38.230.23/corola_sound_search/index.php

Tipul colecției	Corpus de vorbire adnotat la nivel de propoziție
Titlul colecției	Read Speech Corpus (RSC)
Nume contribuitor	Universitatea Politehnica din București
Creator	Universitatea Politehnica din București
Descriere	Corpus de vorbire continuă, citită, în limba română. **Achiziție.** Corpusul a fost achiziționat prin înregistrarea de fișiere audio utilizând o aplicație disponibilă online [2]. Vorbitorii au fost în general studenți ai Universității Politehnice din București. Aplicația afișează textul ce trebuie citit și apoi preia semnalul audio de la microfon. **Conținut.** Textele citite de vorbitori reprezintă: (i) cuvinte izolate (WORDS din [3]), (ii) propoziții din articole de știri sau interviuri text online (CS_01, CS_04 și CS_05 din [3]) și (iii) propoziții din texte literare (Paolo Coelho - Jurnalul unui mag și Marin Preda - Viața ca o prada; CS_06 din [3]). Mai multe informații despre corpus pot fi găsite în [3] și [4].
Data	2014
Topic/Domeniu	știri, interviuri, literatură
Limba	Română
Sursa	Înregistrări efectuate de către amatori voluntari folosind propriul sistem de achiziție audio
Format	Fișierele audio sunt în format wav, 16kHz, 16bps. Fișierele text sunt în format txt, UTF-8
Adnotări	Transcrierea text nu conține niciun fel de adnotări
Aliniere	Alinierea transcrierii cu semnalul vocal la nivel de propoziție
Drepturi	Universitatea Politehnica din București
Dimensiune	Audio: ~147 mii de fișiere audio, 105.8 ore de la 165 vorbitori;
Localizare	Server de fișiere Universitatea Politehnica din București

Tipul colecției	Corpus de vorbire adnotat la nivel de rostire
Titlul colecției	Spontaneous Speech Corpus (SSC-train)

Nume contribuitor	Universitatea Politehnica din București
Creator	Transcrieri: Universitatea Politehnica din București Audio: televiziuni și posturi de radio din România
Descriere	Corpus de vorbire spontană, în limba română. **Achiziție.** Fișierele audio și transcrierile aproximative au fost descărcate de pe website-urile principalelor televiziuni din România. Materialul audio și transcrierile aproximative au fost aliniate automat folosind metodologia descrisă în [5]. **Conținut.** Emisiuni de știri și talkshow-uri transcrise în mod automat. Mai multe informații despre corpus pot fi găsite în [3] și [4].
Data	2014
Topic/Domeniu	știri, talkshow-uri
Limba	Română
Sursa	Internet
Format	Fișierele audio sunt în format wav, 16kHz, 16bps. Fișierele text sunt în format txt, UTF-8
Adnotări	Transcrierea text nu conține niciun fel de adnotări
Aliniere	Alinierea transcrierii cu semnalul vocal la nivel de secvență de cuvinte
Drepturi	Transcrieri: Universitatea Politehnica din București Audio: -
Dimensiune	Audio: ~54 mii de fișiere audio, 27.5 ore de vorbire;
Localizare	Server de fișiere Universitatea Politehnica din București

Tipul colecției	Corpus de vorbire adnotat la nivel de rostire
Titlul colecției	Spontaneous Speech Corpus (SSC-eval)
Nume contribuitor	Universitatea Politehnica din București
Creator	Transcrieri: Universitatea Politehnica din București Audio: televiziuni și posturi de radio din România
Descriere	Corpus de vorbire spontană, în limba română. **Achiziție.** Fișierele audio au fost descărcate de pe website-urile principalelor televiziuni din România, iar etichetarea lor s-a facut manual. **Conținut.** Emisiuni de știri și talkshow-uri transcrise în mod automat. Mai multe informații despre corpus pot fi găsite în [3] și [5].

Data	2014
Topic/Domeniu	știri, talkshow-uri
Limba	Română
Sursa	Internet
Format	Fișierele audio sunt în format wav, 16kHz, 16bps. Fișierele text sunt în format txt, UTF-8
Adnotări	Transcrierea text nu conține niciun fel de adnotări
Aliniere	Alinierea transcrierii cu semnalul vocal la nivel de secvență de cuvinte
Drepturi	Transcrieri: Universitatea Politehnica din București Audio: -
Dimensiune	Audio: 3035 fișiere audio, 3.5 ore de vorbire;
Localizare	Server de fișiere Universitatea Politehnica din București

Tipul colecției	Corpus de vorbire adnotata la nivel de rostire
Titlul colecției	Spontaneous Speech Corpus 2 (SSC-train2)
Nume contribuitor	Universitatea Politehnica din București
Creator	Transcrieri: Universitatea Politehnica din București Audio: televiziuni și posturi de radio din România
Descriere	Corpus de vorbire spontană, în limba română. **Achiziție.** Fișierele audio au fost descărcate de pe website-urile principalelor televiziuni din România. O parte din materialul audio a fost transcris automat folosind metodologia descrisă în [6]. **Conținut.** Emisiuni de știri transcrise în mod automat. Mai multe informații despre corpus pot fi găsite în [6] și [7].
Data	2017
Topic/Domeniu	știri, talkshow-uri
Limba	Română
Sursa	Internet
Format	Fișierele audio sunt în format wav, 16kHz, 16bps. Fișierele text sunt în format txt, UTF-8
Adnotări	Transcrierea text nu conține niciun fel de adnotări
Aliniere	Alinierea transcrierii cu semnalul vocal la nivel de secvență de cuvinte

Drepturi	Transcrieri: Universitatea Politehnica din București Audio: -
Dimensiune	Audio: 170.292 fisiere audio, 103 ore de vorbire;
Localizare	Server de fișiere Universitatea Politehnica din București

Tipul colecției	Corpus audio
Titlul colecției	SWARA Speech Corpus
Nume contribuitor	Universitatea Tehnică din Cluj-Napoca
Creator	Proiect SWARA
Descriere	Corpus audio de aproximativ 21 de ore, compus din înregistrări în studio a 17 vorbitori.
Data	2016
Topic/Domeniu	Sinteză text-vorbire/ Recunoașterea vorbirii
Limba	Română
Sursa	Înregistrări în studio Universitatea Tehnică din Cluj-Napoca
Format	wav, 48kHz, 16bps, segmentate la nivel de propoziție
Adnotări	Transcriere ortografică și adnotare semi-automată la nivel de fonem
Aliniere	Aliniere semi-automată la nivel de fonem
Drepturi	https://creativecommons.org/licenses/by-sa/4.0/
Dimensiune	7.5GB
Localizare	https://speech.utcluj.ro/swarasc/

Tipul colecției	Corpus text
Titlul colecției	Corpus de text Adevărul.ro
Nume contribuitor	Universitatea Tehnică din Cluj-Napoca
Creator	Universitatea Tehnică din Cluj-Napoca
Descriere	Corpus de text compus din articole din ziarul Adevărul - versiunea online
Data	2011
Topic/Domeniu	Procesare de text
Limba	Română
Sursa	Adevărul.ro

Format	Text
Adnotări	-
Aliniere	-
Drepturi	https://creativecommons.org/licenses/by-nc/3.0/
Dimensiune	1871 articole - 4MB
Localizare	server Universitatea Tehnică din Cluj-Napoca

Tipul colecției	Corpus audio
Titlul colecției	Cartea Sonoră - Mara - Audiobook
Nume contribuitor	Universitatea Tehnică din Cluj-Napoca
Creator	Cartea sonoră
Descriere	Corpus audio de aproximativ 11 de ore, compus din înregistrarea nuvelei Mara de către un vorbitor profesionist de gen feminin
Data	2014
Topic/Domeniu	Sinteză text-vorbire/ Analiză prozodică
Limba	Română
Sursa	Înregistrări în studio Cartea Sonoră
Format	mp3, 44.1kHz, 16bps, segmentate la nivel de capitol
Adnotări	Transcriere ortografică aproximativă
Aliniere	Fără aliniere
Drepturi	https://creativecommons.org/licenses/by-nc/3.0/
Dimensiune	500MB
Localizare	https://speech.utcluj.ro/corpora/mara.html

Tipul colecției	Corpus audio
Titlul colecției	RO-GRID
Nume contribuitor	Universitatea Tehnică din Cluj-Napoca
Creator	Universitatea Tehnică din Cluj-Napoca
Descriere	Corpus audio de aproximativ 9 de ore, compus din înregistrări scurte cu format prestabilit: <command: 4> <color: 4> <preposition: 4 words> <letter: 25> <digit: 10> <adverb: 4> Numerele indică numărul de variante pentru fiecare tip. De exemplu, command 4 indică patru posibile comenzi: vezi, muta, pune, sari. Mai multe detalii în articolul de la adresa

	https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6256337
Data	2012
Topic/Domeniu	Recunoașterea vorbirii, domeniu limitat
Limba	Română
Sursa	Înregistrări în incintă acustică semi-izolată fonic
Format	wav, 16kHz, 8bps, segmentate la nivel de propoziție
Adnotări	Transcriere ortografică și adnotare la nivel de fonem
Aliniere	Aliniere manuală la nivel de fonem
Drepturi	https://creativecommons.org/licenses/by-nc/3.0/
Dimensiune	11 vorbitori, aproximativ 50 min/vorbitor, 9 ore de voce în total
Localizare	server Universitatea Tehnică din Cluj-Napoca

Tipul colecției	Corpus bimodal
Titlul colecției	Corpus IIT
Nume contribuitor	Universitatea Alexandru Ioan Cuza din Iași
Creator	Universitatea Alexandru Ioan Cuza din Iași și Institutul de Informatică Teoretică a Academiei Române, Filiala Iași.
Descriere	Corpus de vorbire spontană, în limba română. Fișierele audio au fost obținute în urma unui acord cu Radio Iași și Radio Universitas. Corpusul constă în general din dezbateri/interviuri și a fost transcris și aliniat manual (în general folosind Praat).
Data	2017
Topic/Domeniu	Sinteză text-vorbire, aliniere text/vorbire, recunoașterea vorbirii
Limba	Română
Sursa	Radio Iași și Radio Universitas
Format	Fișierele audio sunt în format wav, 16kHz, 16bps. Alinierea (și transcrierea) în fișiere textgrid (create în Praat).
Adnotări	Textul nu conține alte adnotări în afara transcrierii fonetice și a alinierii.
Aliniere	Alinierea la nivel de rostire (utterance).
Drepturi	Institutul de Informatică Teoretică, Academia Română, Filiala Iași
Dimensiune	Aproximativ 30 ore aliniate, plus 20 doar transcrise.

Localizare	pe serverul 89.38.230.23la home/corola/corpusIasi/

Tipul colecției	Corpus bimodal
Titlul colecției	Corpusul SoRoEs
Nume contribuitor	Universitatea ‘Alexandru Ioan Cuza’ din Iași
Creator	Departamentul de Cercetare Interdisciplinar - Domeniul Socio-Uman Universitatea ‘Alexandru Ioan Cuza’ din Iași
Descriere	Corpus de vorbire spontană și semi-spontană, în limba română. Corpusul a fost obținut prin înregistrările realizate în 10 centre culturale ale țării și prin aplicarea a 2 chestionare. Chestionarul fix a fost aliniat manual (folosind utilitarul PRAAT), fiind adnotate elementele vocalice ale enunțurilor pentru 100 de persoane intervievate. Corpusul liber, spontan conține conversații libere și /sau dirijate, nealiniate.
Data	2015-2017
Topic/Domeniu	Analiză prozodică
Limba	Română
Sursa	Înregistrări de teren, realizate în perioada 2015-2017 în 10 centre culturale ale țării: Baia Mare, Brașov, București, Cluj-Napoca, Constanța, Craiova, Iași, Sibiu, Suceava, Timișoara.
Format	Fișierele audio sunt în format wav, 48kHz. Chestionarul fix aliniat în PRAAT în fișiere TextGrid, iar trăsăturile frecvenței fundamentale, duratei și intensității în fișiere .txt
Adnotări	Pentru chestionarul fix au fost adnotate elementele vocalice (vocale, diftongi triftongi).
Aliniere	Alinierea la nivel de rostire (utterance).
Drepturi	Departamentul de Cercetare Interdisciplinar - Domeniul Socio-Uman din cadrul Universității “Alexandru Ioan Cuza” din Iași
Dimensiune	Înregistrări în cadrul a 10 anchete, cu răspunsuri la 92 de întrebări dintr-un chestionar.
Localizare	http://soroes.ro/

5. Observații privind colecțiile de date lingvistice existente la parteneri

5.1 Statistici

În urma inventarierii corpusurilor se observă că majoritatea corpusurilor sunt bimodale, incluzând fișiere audio și transcrieri sau alinieri ale acestora. Un singur corpus raportat este doar corpus de texte, așa cum tot un singur corpus este doar corpus de voce. Așa cum a fost detaliat în secțiunea 11, au fost identificate 11 corpusuri. În ceea ce privește dimensiunea, corpusurile raportate însumează peste 450 de ore de înregistrare, conform distribuției din figura 2 de mai jos, la care se adaugă 1871 de articole de ziare în format text.

Fig. 2. Distribuția orelor de înregistrare per partener

5.2. Formate

Formatul de adnotare este diferit de la partener la partener. Deoarece se are în vedere propunerea unui format standard pentru înregistrările care vor fi incluse pe platforma ReTeRom, format ce va fi detaliat în raportul 1.3, am considerat utilă prezentarea mai detaliată a formatelor fișierelor de aliniere.

Corpusurile inventariate de UPB conține, pe lângă fișierele audio, un fișier cu transcrierea înregistrării, care conține textul transcrierii. Pe baza acestui fișier, poate fi făcută alinierea la nivel de cuvânt în mod automat, folosindu-se sistemul propriu de recunoaștere a vorbirii.

Corpusul inventariat de ICIA conține, pentru fiecare fișier .wav cu înregistrarea audio, câte 3 alte fișiere, cu extensia .txt, .lab și respectiv .phn. Fișierul .txt include transcriere, de exemplu:

Purtătorul de cuvânt al Biroului Electoral Central, Marian Muhuleț,…

Fișierul cu extensia .lab conține normalizarea textului, normalizare care elimină semnele de punctuație, ca în exemplul:

Purtătorul de cuvânt al Biroului Electoral Central Marian Muhuleț…

Al treilea fișier, cel cu extensia .phs conține alinierile la nivel de fonem. Pentru exemplul de mai sus, începutul fișierului de aliniere este:

3700000	4400000	p	purtătorul
4400000	4900000	u	
4900000	5200000	r	
5200000	5600000	t	
5600000	6200000	@	

6200000	6700000	t	
6700000	7200000	o	
7200000	7500000	r	
7500000	8100000	u	
8100000	8400000	l	
8400000	8800000	d	de
8800000	9200000	e	

Primele două coloane reprezintă intervalul de timp (începutul, respectiv sfârșitul pronunțării fonemului), a treia coloană este transcrierea fonemului, iar a patra coloană apare doar la începutul unui nou cuvânt și marchează întregul cuvânt.

Corpusul SWARA inventariat de UTCN este descris în rapoartele proiectului SWARA, disponibile pe site-ul https://speech.utcluj.ro/swarasc/. Fișierele audio sunt complementate și în cazul acestui corpus de alte 3 fișiere, cu aceleași extensii ca în cazul corpusului ICIA. Fișierele .txt și .phn conțin același tip de informați ca cea descrisă anterior. Fișierul .lab conține o structură mai complexă, care include informații despre fonemului curent, împreună cu contextul lui fonematic (două foneme înainte și două după), silaba curentă împreună cu contextul ei, cuvântul și respectiv propoziția curentă, accentuarea din silaba/cuvânt/grup, număr de silabe accentuate sau nu, în cuvânt/grup și în context, dar include și informații despre partea de vorbire estimată a cuvântului curent și a celui anterior. Un exemplu de adnotare pentru sintagma “până atunci” este prezentată mai jos. Detalierea etichetelor este prezentată în anexa 1.

1050000 1950000 #~p-a@+n=@:2_1/A/0_0_0/B/1-1-2:1-2&1-12#0-4$0-4>0-0<0-0|a@/C/0+0+2/D/content_0/E/content+2:1+8&0+0#0+0/F/content_0/G/0_0/H/12=8:1=1&L-L%/I/0_0/J/12+8-1

1950000 2250000 p~a@-n+@=a:1_2/A/0_0_0/B/0-0-2:2-1&2-11#1-3$1-3>1-2<1-2|@/C/0+0+1/D/content_0/E/content+2:1+8&0+0#0+0/F/content_0/G/0_0/H/12=8:1=1&L-L%/I/0_0/J/12+8-1

2250000 2700000 a@~n-@+a=t:2_1/A/0_0_0/B/0-0-2:2-1&2-11#1-3$1-3>1-2<1-2|@/C/0+0+1/D/content_0/E/content+2:1+8&0+0#0+0/F/content_0/G/0_0/H/12=8:1=1&L-L%/I/0_0/J/12+8-1

2700000 3650000 n~@-a+t=u:1_1/A/0_0_2/B/0-0-1:1-2&3-10#1-3$1-3>2-1<2-1|a/C/1+1+5/D/content_2/E/content+2:2+7&0+0#0+0/F/content_0/G/0_0/H/12=8:1=1&L-L%/I/0_0/J/12+8-1

3650000 4500000 @~a-t+u=n:1_5/A/0_0_1/B/1-1-5:2-1&4-9#1-3$1-3>3-0<3-0|u/C/0+0+2/D/content_2/E/content+2:2+7&0+0#0+0/F/content_0/G/0_0/H/12=8:1=1&L-L%/I/0_0/J/12+8-1

4500000 5400000 a~t-u+n=ch:2_4/A/0_0_1/B/1-1-5:2-1&4-9#1-3$1-3>3-0<3-0|u/C/0+0+2/D/content_2/E/content+2:2+7&0+0#0+0/F/content_0/G/0_0/H/12=8:1=1&L-L%/I/0_0/J/12+8-1

Corpusurile inventariate de UAIC conțin fișiere audio în format .wav, un fișier TextGrid și un fișier .txt. Fișierul TextGrid conține segmentarea vocalelor din fișierul audio și identificarea vocalelor. Un extras dintr-un fișier .TextGrid este prezentat mai jos, unde se observă intervalele de început, respectiv sfârșit pentru fiecare vocală.

```
File type = "ooTextFile"
Object class = "TextGrid"
xmin = 0
xmax = 1.5149375
tiers? <exists>
size = 1
item []:
    item [1]:
        class = "IntervalTier"
        name = "voyelles"
        xmin = 0
        xmax = 1.5149375
        intervals: size = 17
```

```
        intervals [1]:
            xmin = 0
            xmax = 0.17960261055804663
            text = ""
        intervals [2]:
            xmin = 0.17960261055804663
            xmax = 0.23306660377413285
            text = "\'1u\~^"
…
```

Fișierul .txt conține, după header, câte o linie pentru fiecare vocală, unde sunt marcați parametrii acustici ai fiecărei vocale: durata (ms), energia (dB) și frecvența (extrasă în câte trei puncte). De asemenea, în partea de jos a fișierului, sunt marcate momentele în timp unde au fost citite valorile. Un exemplu este detaliat mai jos:

	duration [ms]	energy [dB]	fo1	fo2	fo3 [Hz]
1	57	84	147	152	153
2	71	81	154	150	145
3	76	84	134	123	108
4	70	81	119	120	116
5	91	79	112	111	109
6	48	66	108	107	107
7	88	73	103	103	102
8	161	70	99	83	76

values at:

2669 3125 3581 4597 5167 5737 6690 7302 7913 8867 9427 9986 10836 11562 12287 14194 14578 14961 20143 20848 21553 22320 23605 24890

După cum se observă, formatele partenerilor sunt variate, astfel fiind necesară o standardizate pentru formatul care urmează a fi folosit pe platforma ReTeRom, și realizarea de convertoare pentru aderarea la formatul comun.

Bibliografie

[1] Weibel, S., Kunze, J., Lagoze, C., & Wolf, M. (1998). Dublin core metadata for resource discovery (No. RFC 2413).

[2] Aplicația de înregistrare a vorbirii: https://speech-recorder.speed.pub.ro

[3] Horia Cucu, Andi Buzo, Lucian Petrică, Dragoș Burileanu and Corneliu Burileanu, "*Recent Improvements of the SpeeD Romanian LVCSR System*", in the Proceedings of the 10th International Conference on Communications (COMM), Bucharest, 2014, pp. 111-114.

[4] Horia Cucu, "*Towards a speaker-independent, large-vocabulary continuous speech recognition system for Romanian*", PhD Thesis, University "Politehnica" of Bucharest, Oct 2011 (scientific coordinator: prof. Corneliu Burileanu).

[5] Andi Buzo, Horia Cucu, Corneliu Burileanu, "*Text Spotting In Large Speech Databases For Under-Resourced Languages*", in the Proceedings of the 7th Conference on Speech Technology and Human-Computer Dialogue (SpeD), Cluj-Napoca, 2013, pp. 77-82, ISBN: 978-1-4799-1065-6.

[6] 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.

[7] Alexandru-Lucian Georgescu, Horia Cucu, Corneliu Burileanu, "*SpeeD's DNN Approach to Romanian Speech Recognition*," in the Proceedings of the 9th Conference on Speech Technology and Human-Computer Dialogue (SpeD), Bucharest, 2017, 8p, ISBN 978-1-5090-6496-0.

Anexa 1

Descrierea formatului .lab pentru corpusul SWARA

*p*1^*p*2- *p*3+*p*4=*p*5 @*p*6 *p*7
/A:*a*1 *a*2 *a*3 /B:*b*1-*b*2-*b*3 @*b*4-*b*5 &*b*6-*b*7 #*b*8-*b*9 $*b*10-*b*11 !*b*12-*b*13 ;*b*14-*b*15 |*b*16 /C:*c*1+*c*2+*c*3
/D:*d*1 *d*2 /E:*e*1+*e*2 @*e*3+*e*4 &*e*5+*e*6 #*e*7+*e*8 /F: *f*1 *f*2
/G:*g*1 *g*2 /H:*h*1=*h*2 @*h*3=*h*4|*h*5 /I:*i*1 *i*2
/J: *j*1+ *j*2- *j*3
*p*1 the phoneme identity before the previous phoneme
*p*2 the previous phoneme identity
*p*3 the current phoneme identity
*p*4 the next phoneme identity
*p*5 the phoneme after the next phoneme identity
*p*6 position of the current phoneme identity in the current syllable (forward)
*p*7 position of the current phoneme identity in the current syllable (backward)
*a*1 whether the previous syllable stressed or not (0: not stressed, 1: stressed)
*a*2 whether the previous syllable accented or not (0: not accented, 1: accented)
*a*3 the number of phonemes in the previous syllable
*b*1 whether the current syllable stressed or not (0: not stressed, 1: stressed)
*b*2 whether the current syllable accented or not (0: not accented, 1: accented)
*b*3 the number of phonemes in the current syllable
*b*4 position of the current syllable in the current word (forward)
*b*5 position of the current syllable in the current word (backward)
*b*6 position of the current syllable in the current phrase (forward)
*b*7 position of the current syllable in the current phrase (backward)
*b*8 the number of stressed syllables before the current syllable in the current phrase
*b*9 the number of stressed syllables after the current syllable in the current phrase
*b*10 the number of accented syllables before the current syllable in the current phrase
*b*11 the number of accented syllables after the current syllable in the current phrase
*b*12 the number of syllables from the previous stressed syllable to the current syllable
*b*13 the number of syllables from the current syllable to the next stressed syllable
*b*14 the number of syllables from the previous accented syllable to the current syllable
*b*15 the number of syllables from the current syllable to the next accented syllable
*b*16 name of the vowel of the current syllable
*c*1 whether the next syllable stressed or not (0: not stressed, 1: stressed)
*c*2 whether the next syllable accented or not (0: not accented, 1: accented)
*c*3 the number of phonemes in the next syllable
*d*1 gpos (guess part-of-speech) of the previous word
*d*2 the number of syllables in the previous word
*e*1 gpos (guess part-of-speech) of the current word
*e*2 the number of syllables in the current word
*e*3 position of the current word in the current phrase (forward)
*e*4 position of the current word in the current phrase (backward)
*e*5 the number of content words before the current word in the current phrase
*e*6 the number of content words after the current word in the current phrase
*e*7 the number of words from the previous content word to the current word
*e*8 the number of words from the current word to the next content word
*f*1 gpos (guess part-of-speech) of the next word
*f*2 the number of syllables in the next word
*g*1 the number of syllables in the previous phrase
*g*2 the number of words in the previous phrase
*h*1 the number of syllables in the current phrase
*h*2 the number of words in the current phrase
*h*3 position of the current phrase in utterence (forward)

*h*4 position of the current phrase in utterence (backward)
*h*5 TOBI endtone of the current phrase
*i*1 the number of syllables in the next phrase
*i*2 the number of words in the next phrase
*j*1 the number of syllables in this utterence
*j*2 the number of words in this utternce
*j*3 the number of phrases in this utterence