Raport științific și tehnic PROIECTUL NR 73PCCDI/2018
Resurse și tehnologii pentru dezvoltarea interfețelor om-mașină în limba română (RETEROM)
Etapa I-a, an 2018

Termen: 30 **Noiembrie 2018**

Aceste rezultate au fost obținute prin finanțare în cadrul Programului PN-III Proiecte complexe realizate în consorții CDI, derulat cu sprijinul MEN – UEFISCDI,
Cod: PN-III-P1-1.2-PCCDI-2017-0818, Contract Nr. 73 PCCDI/2018

Denumirea organizației participante în proiect	Acronim organizație	Tip organizație	Rolul organizației în proiect (Coordonator/partener)
Institutul de Cercetări Pentru Inteligență Artificială "Mihai Drăgănescu"	ICIA	UNI	CO
Universitatea Tehnică din Cluj-Napoca	UTCN	UNI	P1
Universitatea Politehnica din București	UPB	UNI	P2
Universitatea "Alexandru Ioan Cuza" din Iași	UAIC	UNI	P3

Rezumatul etapei

Proiectul complex RETEROM este format din 4 proiecte componente: COBILIRO, TEPROLIN, TADARAV și SINTERO, fiecare cu obiective și realizări specifice. În continuare vor fi prezentate pe scurt, pentru fiecare dintre ele, informațiile tehnico-științifice așa cum au fost elaborate de responsabilii proiectelor componente (descrierile detaliate sunt cuprinse în rapoartele de activitate, indicate prin adresa web de unde pot fi descărcate), cu punerea in evidenta a rezultatelor etapei și gradul de realizare a obiectivelor, incluzând realizarea indicatorilor de rezultat atinsi; de asemenea, fiecare descriere de proiect component al proiectului complex prezintă oferta de servicii de cercetare si tehnologice cu indicarea link-ului din platforma Erris;

Comunicarea între parteneri a fost constantă, prin schimbarea a zeci de mesaje e-mail, prin convorbiri telefonice, prin 3 conferinte Skype și doua workshopuri în iunie la București și în noiembrie la Iași.

Locurile de muncă susținute prin program sunt reprezentate de membrii echipelor de cercetare (23 cercetători, din care 5 nou angajați deja la UPB (3), ICIA (1) și UAIC (1)- doi urmând a fi angajați la UTCN luna vitoare).

Partenerii consorțiului nu au folosit cecurile pe anul 2018 din varii motive, cel mai important fiind stadiul incipient al proiectului. O parte din fondurile alocate cecurilor au fost realocate la cheltuielile de management.

MINISTERUL EDUCAȚIEI NAȚIONALE

UEFISCDI

PROIECTUL COBILIRO

Anca-Diana Bibiri, Dan Cristea, Daniela Gîfu, Mihaela Plămadă-Onofrei, Ionuț Cristian Pistol, Andrei Scutelnicu, Diana Trandabăț

Responsabil: Universitatea A. I. Cuza, coordonator Prof. Dan Cristea, McAR

1. Rezumatul etapei

Etapa I a proiectului component COBILIRO s-a intitulat **Acțiuni preparatorii realizării unui corpus bimodal (vorbire/text) pentru limba română**. Conform planului, rezultatele așteptate au fost: 1. Elaborarea unui studiu atent asupra nivelului atins în străinătate în elaborarea de corpusuri bimodale voce/text. 2. Inventarierea colecțiilor existente. Indicatorii de realizare planificați au fost: Predarea studiului și a inventarului (setul de livrabile D1.1.1, D1.1.2, D1.1.3 și D1.1.4). Bugetul alocat de la stat pentru realizarea acestei etape a fost de 289863 lei.

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. Alte activități, urmăresc realizarea comunicării în cadrul consorțiului, diseminarea rezultatelor, coordonare și management și raportare.

2. Descrierea științifică și tehnică,

Descriem în această secțiune cele 4 activități planificate, punând în evidență rezultatele etapei, gradul de realizare a obiectivelor și modul de diseminare a rezultatelor. Menționăm că toți indicatorii de rezultat planificați au fost atinși.

2.1. Activitatea *1-1-1 Studiu state-of-the-art asupra realizării corpusurilor bimodale*

Această activitate și-a propus realizarea unui studiu cuprinzător asupra nivelului cercetărilor în lume în realizarea de corpusuri bimodale și s-a materializat prin raportul livrabil: *D1.1.1: Raport asupra corpusurilor bimodale similare din străinătate* (http://www.racai.ro/p/reterom/rapoarte/1.1.pdf). Instituțiile implicate în realizarea acestei etape au fost toți cei 4 parteneri din proiectul complex: UAIC, UPB, UTCN, ICIA. Activitatea inclusă în această etapă, pe care o vom rezuma mai jos, se include în categoria de activitate cercetare fundamentală.

În această etapă au fost identificate, analizate și evaluate studiile centrate pe realizarea corpusurilor bimodale în vederea selectării acelora care pot constitui modele necesare producerii unei tehnologii integrate pentru procesarea limbajului natural în limba română și adnotarea pe diferite niveluri lingvistice a corpusului bimodal generat în cadrul Proiectului 1 COBILIRO.

Limbajul natural este modalitatea principală, având proprietăți invariabile în mediul auditiv - ca limbă vorbită, în mediul vizual - ca limbă scrisă, în mediul tactil - ca Braille și în mediul cinetic - ca limbaj al semnelor. Din acest punct de vedere, termenul de bi-modal (care înseamnă utilizarea combinată a două modalități) are un înlocuitor preferat, care este bi-medial (care înseamnă, utilizarea combinată a două medii). Cele două medii de comunicare ale limbajului, în cazul proiectului nostru, sunt textual și auditiv.

În accepțiunea proiectului ReTeRom, prin corpus bimodal[1] (care este un caz particular de corpus multimodal, la rândul lui reprezentând un tip particular de corpus) vom înțelege o colecție de înregistrări orale însoțite de transcrierile lor și de metadatele corespunzătoare. Un corpus bimodal este găzduit pe o platformă specializată, împreună cu serviciile și aplicațiile web de acces, dezvoltare și întreținere ale lui, unde

[1] V. și http://www.aclweb.org/anthology/R09-1044

sunt specificați algoritmii de utilizare ai corpusului și, în unele cazuri, de unde pot fi descărcate exemple de aplicații care utilizează corpusul. Corpusurile pot conține texte scrise, înregistrări orale sau ambele modalități de redare a unei limbi naturale. În proiectul de față ne interesează ultimul caz.

Corpusurile orale sunt clasificate în două tipuri: voce-în-citire (read speech) (incluzând lecturi din cărți, știri, liste de cuvinte, secvențe de numere) și vorbire spontană (spontaneous speech) (incluzând: dialoguri între două sau mai multe persoane, narative, relatări despre trasee pe hartă, stabilirea de întâlniri, simulări "Vrăjitorul din Oz").

În variantă bimodală, corpusurile orale sunt transcrise în text. Tipurile de transcrieri posibile într-un corpus oral sunt: transcrierea ortografică fără aliniere în timp cu semnalul sonor, transcrierea ortografică aliniată în timp, transcriere fonetică realizată cu ajutorul simbolurilor.

În livrabilul D1.1 au fost trecute în revistă 11 corpusuri bimodale construite în diverse laboratoare de cercetare de pe mapamond. Informațiile au fost grupate în 10 categorii, care au fost identificate în majoritatea descrierilor: limba, tipul de corpus, dimensiunea (în număr de cuvinte), vorbitori, informații private relative la vorbitori, metadate, conținut și mod de realizare, formatul înregistrărilor, suport pentru realizare, distribuție și copyright.

Printr-o analiză comparativă a mai multor propuneri de standardizare, s-a propus o structură de metadate și transcrieri aliniate, care include următoarele tipuri de informații:

- specificații relative la vorbitori: vârstă, gen, educație, calitate a vocii, dialect, accent, dar și un ID ales de vorbitor sau generat automat;
- specificații relative la contextul producerii înregistrării: dacă înregistrarea reprezintă vorbire spontană sau voce-în-citire; în cazul vorbirii spontane - dacă ea este solicitată (răspuns la întrebări etc.), în cazul vocilor-în-citire - metadatele textului reprodus, dacă înregistrarea reprezintă un dialog sau un monolog, dacă vorbirea e expresivă;
- specificații relative la înregistrare: nivelul tehnic al echipamentului cu care s-a produs înregistrarea (microfon, platforma soft-hard de înregistrare), contextul înregistrării (în laborator, în casă, pe stradă, în mașină etc.), nivelul zgomotului de fond (în db), detalii asupra digitalizării semnalului, data, momentul și locul înregistrării, un ID al înregistrării (ales sau generat) pus în legătură cu numele fișierului[2];
- specificații de adnotare: formatul convențiilor de adnotare ale transcrierii sunet-text (aliniere semnal-timp și text-timp sau direct semnal-offset caracter), adnotări ortografice, fonetice, prozodice, sintactice etc., instrumentele de segmentare și aliniere utilizate;
- criterii de validare: dacă au fost utilizate anumite criterii de validare a corpusului, la nivel de înregistrare dar și global, la nivelul corpusului ca întreg;
- specificații de distribuție și stocare: planul de distribuție, mediul de stocare, condiții backup etc.

Aceste elemente sunt detaliate în livrabilul D1.1.3 (http://www.racai.ro/p/reterom/rapoarte/1.3.pdf), rezumat mai jos în secțiunea 2.3.

Problemele etice în înregistrarea și publicarea în spațiul public a ceea ce a fost produs într-un cadru privat și destinat unei audiențe limitate sunt întâlnite mai frecvent în tratarea textelor vorbite decât în cele scrise (TEI[3]). Ca urmare, aspectele legale la constituirea unui corpus bimodal trebuie să aibă în vedere următoarele:

- drepturile de autor ale proprietarilor textelor pronunțate (în cazul corpusurilor voce-în-citire),
- acceptul vorbitorilor de a utiliza înregistrările făcute cu vocea lor, pentru scopuri didactice, de cercetare sau comerciale (depinzând de caz),
- acceptul vorbitorilor de a include în metadate informații de natură personală,
- convenții legale semnate între producătorii, distribuitorii și utilizatorii corpusului.

[2] Lungimea numelui fișierului trebuie să respecte specificațiile atributelor de fișiere descrise în ISO-9960.

[3] https://quod.lib.umich.edu/cgi/t/tei/tei-idx?type=pointer&value=TSOV

Convenția de colaborare trebuie semnată între producător și vorbitor înainte de începerea înregistrărilor.

Colecția de resurse ce urmează a fi organizată (dezvoltată, standardizată) în cadrul proiectului ReTeRom nu se adresează strict membrilor consorțiului proiectului. Această activitate are un orizont științific de durată, cu un pronunțat caracter de generalitate. Ea deschide calea pentru formarea unei viziuni științifice care să ghideze colectarea, adnotarea și distribuția de corpusuri de acest gen în România.

Activitatea A1-1-1 din proiectul complex ReTeRom lămurește câteva chestiuni de terminologie a domeniului, sintetizează aspectele relevante în crearea corpusurilor bimodale, incluzând modalități de achiziție, prezintă un număr de exemple notorii de astfel de corpusuri, cu sintetizarea a 10 trăsături care au putut fi revelate din literatură, și prezintă preocupări de standardizare în realizarea corpusurilor bimodale, precum și aspecte legale. **Toate obiectivele incluse în plan au fost realizate.**

2.2 Activitatea *1-1-2 Inventarul colecțiilor de date lingvistice românești disponibile la parteneri sau în terțe coaliții și a formatelor de stocare ale acestora*

Această activitatea și-a propus realizarea unui inventar al colecțiilor lingvistice românești aflate la partenerii din proiectul ReTeRom și s-a materializat prin raportul livrabil: *D1.1.2 Raport privind inventarul colecțiilor existente și a formatelor de reprezentare și adnotare* (http://www.racai.ro/p/reterom/rapoarte/1.2.pdf). Instituțiile implicate în realizarea acestei etape au fost toți cei 4 parteneri din proiectul complex: UAIC, UPB, UTCN, ICIA.

În această etapă au fost inventariate resursele lingvistice existente la partenerii din cadrul proiectului complex, precum și la terți proprietari. Fiecare resursă 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, majoritatea multimodale de la parteneri.

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 corpusuri 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.

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. 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.

Inventarierea colecțiilor de date lingvistice existente la parteneri s-a realizat având în vedere un set de trăsături, în acord cu standardul internațional Dublin Core de descriere a resurselor. Setul de metadate Dublin Core conține cincisprezece trăsături generice, utile pentru descrierea unei mari varietăți de resurse. 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
- 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 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 amintit mai sus, au fost identificate 11 corpusuri. În ceea ce privește dimensiunea, corpusurile raportate însumează peste 450 de ore de înregistrare, la care se adaugă 1871 de articole de ziare în format text.

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 a formatelor fișierelor de aliniere.

Corpusurile inventariate de UPB conțin, pe lângă fișierele audio, un fișier cu transcrierea înregistrării, care includ 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 transcrierea ortografică. Fișierul cu extensia .lab conține normalizarea textului, normalizare care elimină semnele de punctuație. Al treilea fișier, cel cu extensia .phn conține alinierile la nivel de fonem.

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, fiind adnotat cu etichete HTS în vederea folosirii sale pentru generarea de limbaj. Astfel, el 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. Detalierea etichetelor este prezentată în livrabilul 1.2.

Corpusul Cartea Sonoră – Mara inventariat de UTCN include doar transcrierea ortografică și intervalele de timp.

Corpusul IIT inventariat de UAIC este dezvoltat în cadrul proiectului CoRoLa, astfel având aceeași structură ca cea a corpusului inventariat de ICIA. Al doilea corpus inventariat de UAIC, SoRoEs conține fișiere audio în format .wav, un fișier TextGrid și un fișier .txt. Fișierul TextGrid conține segmentarea vocalelor din fișierul audio și identificarea vocalelor. 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, sunt marcate momentele în timp la care au fost citite valorile.

Se constată o destul de mare diversitate a formatelor utilizate de parteneri în codificarea corpusurilor bimodale, ceea ce impune aducerea lor la un format unitar, care va urma să fie folosit pe platforma ReTeRom. Convertirea de formate pentru aderarea la formatul comun reprezintă activități viitoare în proiect. **Toate obiectivele incluse în plan la această activitate au fost realizate.**

2.3 *Activitatea 1-1-3 Proiectarea funcțională și arhitecturală a infrastructurii care va găzdui resursele și instrumentele de prelucrare și acces ale consorțiului și realizarea unui prototip*

În această activitate ne-am propus să proiectăm arhitectura și funcționalitatea unui Portal care să permită stocarea resurselor bimodale din proiect și care să permită accesul la ele. Activitatea s-a materializat prin raportul livrabil: *Proiectul arhitectural și funcțional al infrastructurii* http://www.racai.ro/p/reterom/rapoarte/1.3.pdf). Instituțiile implicate în realizarea acestei etape au fost toți cei 4 parteneri din proiectul complex: UAIC, UPB, UTCN, ICIA. Activitatea inclusă în această etapă, pe care o vom rezuma mai jos, se include în categoria de activitate *cercetare fundamentală*.

S-a descris structura unui Portal ReTeRom-COBILIRO, care va găzdui resursele și instrumentele de prelucrare și acces ale consorțiului. Un prototip al acestei platforme va fi prezentat în workshop-ul ReTeRom din 23 noiembrie 2018, desfășurat în tandem cu a 13-a ediție a conferinței internaționale a Consorțiului de Informatizare pentru Limba Română - ConsILR-2018, *Linguistic Resources and Tools for Processing Romanian Language*, care se va desfășura între 22-23 noiembrie, la Filiala Iași a Academiei Române (https://profs.info.uaic.ro/~consilr/).

Portalul va putea fi accesat din pagina proiectului complex ReTeRom și cea a subproiectului COBILIRO, secțiunile lui principale fiind următoarele 6: acasă - o zonă de informații generale despre proiect și participanți, scurte prezentări ale instituțiilor partenere, secțiunea principală de servicii, secțiunea de comunicări și publicații, partenerii externi cu care proiectul are legături și informații de contact. Secțiunea de resurse și servicii de procesare a limbajului natural va conține: lista resurselor existente, posibil de accesat în căutare pe bază de cuvinte cheie din descriere sau pe bază unui formular referitor la metadate, o zonă dedicată încărcării unei noi resurse, o zonă de corectare/actualizare a unei resurse și o zonă de exploatare a resurselor. Secțiunea de comunicări susținute și articole publicate va oferi informații despre: rapoarte de cercetare (cu acces doar pentru persoanele din consorțiu), calendarul evenimentelor interne și de interes pentru consorțiu, documentații, liste de lucrări de interes științific din domeniul prelucrării limbajului natural, evenimente organizate de institutele consorțiului, un forum de discuții, o zonă a presei cu testimoniale asupra activităților derulate prin Portal, ecouri mediatice și, opțional, un newsletter.

Din punctul de vedere al funcționalității, Portalul împarte utilizatorii în următoarele categorii: administrator, curator de resurse (persoană care dă acceptul de publicare pe site, monitorizează și gestionează resursele), contributor (persoană care donează noi resurse) și utilizatorul de resurse.

Interogarea bazei de date se va putea face: prin cuvinte cheie care să se potrivească pe descrierea resurselor, prin completarea unui formular de interogare care încearcă potriviri asupra câmpurilor din metadate și direct pe conținut, prin interfețe specializate web.

Alte servicii oferite de Portal au în vedere: lansarea în execuție a unei resurse ca serviciu web, încărcarea unei resurse, ștergerea unei resurse, corecția/actualizarea unei resurse, conversia de format a unei resurse, operațiuni de salvare periodică a bazei de date a Portalului (back-up), precum și administrare securizată, formular de contact, forum de discuții/chat, RSS, Google Analytics etc.

Pentru compatibilitate cu resurse internaționale, formatul de reprezentare al datelor propus este foarte apropiat de standardul TEI P5 (Sperberg-McQueen and Burnard, 2018). Acest format standard este necesar pentru uniformizarea contribuțiilor membrilor consorțiului ReTeRom și în vederea proiectării tehnologiei de conversie de metadate/înregistrări sonore/înregistrări textuale. La nivelul metadatelor, propunerea combină informații preluate din (Li and Yin, 2007) cu rezultatul analizei făcute în secțiunea 3 a livrabilului A1.1 și în livrabilul A1.2, pe un schelet TEI P5. La nivelul adnotării înregistrărilor, convențiile noastre urmează îndeaproape indicațiile TEI, cu câteva adaosuri necesare descrierii explicite a corpusului bimodal, tip de document care nu are o secțiune dedicată în ghidul TEI.

În interesul proiectului ReTeRom este ca înregistrările vocale să fie prezentate în corpus în pereche cu transcrierile lor textuale. Unitatea de bază în alinierea vorbire-text este fraza (ori întinderea ei minimală - propoziția). Inferior limitei de frază/propoziție, vorbirea poate fi segmentată în unități morfologice (cuvânt), fonologice (fonem), prozodice (*pitch*, creștere ori descreștere a frecvenței fundamentale) ori sintactice (grup nominal, clauză etc.), deși ghidul TEI nu precizează nume pentru astfel de segmente.

Un document (element <TEI>) trebuie să conțină o secțiune <teiHeader>, care furnizează informații contextuale detaliate, cum ar fi: sursa obiectului, identitatea participanților (în condițiile respectării constrângerilor de confidențialitate), tipul de vorbire (spontană sau voce-în-citire), aspecte privind condițiile tehnice ale înregistrării etc. Fiecare obiect are în constituția sa o secvență de perechi <speech> <text>, unde <speech> reprezintă înregistrarea sonoră, iar <text> - transcrierea ei textuală. Înregistrările sonore pot fi: conversații între un număr mic de persoane, prelegeri, piese de teatru, emisiuni Radio sau TV difuzate, interviuri efectuate pe stradă, în casă, în mijloace de transport etc, înregistrări speciale realizate în condiții de laborator (camere anecoide), audiobook-uri etc.

TEI Guidelines recomandă ca un document să conțină un <text> coeziv, să conțină o înregistrare care să acopere o întindere de timp contiguă, fără întreruperi semnificative (microfon deschis o singură dată) și să poată fi descrisă printr-un singur <teiHeader>. Adesea însă aceste cerințe nu pot fi îndeplinite simultan (ca de exemplu, în înregistrările audio-book-urilor, în înregistrări provenite din piese de teatru și chiar în unele emisiuni radio/TV, care sunt realizate în sesiuni de lucru succesive), deci încalcă cerința întinderii de timp contigue. Alte documente pot include culegeri de fraze disparate, fără legătură între ele, prea multe pentru a li se atribui fiecăruia un header propriu, deci încalcă cerința textului coeziv. Ca urmare, cerințele formulate trebuie considerate doar orientative și nu vor fi impuse în corpusul ReTeRom.

Livrabilul A1.3 descrie o propunere de structură și funcționare a Portalului proiectului ReTeRom/COBILIRO, care să facă față cerințelor de stocare și acces a resurselor bimodale deținute de partenerii proiectului, precum și în perspectiva utilizării lui deschise pentru activități de cercetare dedicate prelucrării limbajului natural. Se propunem apoi un standard de reprezentare a datelor corpusului bimodal, inspirat din standarde internaționale, care să ofere un echilibru între completitudinea informațiilor și simplitate. **Toate obiectivele incluse în plan la această activitate au fost realizate.**

2.4 Activitatea 1-1-4 Management și diseminare

Această activitate s-a concentrat pe organizarea eforturilor de cercetare din primul an în cadrul proiectului COBILIRO, precum și pe diseminarea rezultatelor și s-a materializat prin raportul livrabil: *D1.1.4: Diseminarea proiectului în mass-media și pe internet*. Instituția implicată în realizarea acestei etape a fost UAIC. Activitatea inclusă în această etapă se include în categoria de activitate *Activități suport - Diseminare și*

participare la manifestări tehnico-științifice. Indicatorul de realizare planificat a fost: pagina web a proiectului este funcțională; partenerii din consorțiu au făcut cunoscut proiectul în cel puțin o intervenție în mass-media.

Membrii colectivului COBILIRO au elaborat mai multe lucrări (13) care au menționat proiectul ReTeRom, prezentând corpusuri lingvistice și de vorbire, precum și modele și instrumente de prelucrare a limbajului. Au fost organizate două workshop-uri ReTeRom, primul la Universitatea „Politehnica" din București în 8 iunie 2018 și al doilea în cadrul celei de-a 13-a ediții a conferinței internaționale *Linguistic Resources and Tools for Processing Romanian Language*. A fost proiectată și realizată pagina COBILIRO, iar fiecare partener și-a realizat propria pagină web. Lucrările în limba engleză au inclus textul: "This work is supported by a grant of the Ministry of Research and Innovation CCCDI-UEFISCDI, project cod PN-III-P1-1.2-PCCDI-2017-0818 within PNCDI III."

3. Structura ofertei de servicii de cercetare și tehnologice

Partenerul UAIC/FII va oferi pe platforma ERRIS serviciile de cercetare și tehnologice enumerate în tabelul următor (https://erris.gov.ro/UAIC-FII).

Serviciu	**Detalii**
UAIC Romanian Part of Speech Tagger	http://nlptools.info.uaic.ro/WebPosRo/
UAIC Romanian Noun Phrase Chunker	http://nlptools.info.uaic.ro/WebNpChunkerRo/
UAIC Romanian FDG parser	http://nlptools.info.uaic.ro/WebFdgRo/
Graphical Grammar Studio	http://sourceforge.net/projects/ggs/

4. Locuri de muncă susținute prin program

Pe lângă cercetătorii cu experiență (Dan Cristea -profesor, Diana Trandabăț - conferențiar, Daniela Gîfu – CS2, Ionuț Pistol - lector, Anca Bibiri - CS, Mihaela Onofrei - CS) în proiect a fost recent angajat pe poziție de doctorand Cristian Pădurariu, care își va începe activitatea la 1 decembrie 2018.

5. La nivelul proiectului component COBILIRO cecurile nu au fost valorificate în 2018.

PROIECTUL TEPROLIN

Verginica Barbu Mititelu, Maria Carp, Eric Curea, Radu Ion, Elena Irimia, Vasile Păiș, Dan Tufiș

Responsabil: Institutul de Cercetări pentru Inteligență Artificială ”Mihai Drăgănescu”, Academia Română, coordonator Acad. Dan Tufiș

Rezumatul etapei

Etapa avută în vedere a avut menirea să analizeze instrumentele de prelucrare a textelor în limba română disponibile la parteneri, să analizeze extensiile și/sau modificările necesare pentru a răspunde necesităților proiectului complex și a propune o arhitectură pentru platforma de servicii de prelucrare a textelor în limba română. Pe baza acestei analize, au fost definite specificațiile funcționale și arhitecturale (http://www.racai.ro/p/reterom/rapoarte/1.5.pdf) ale platformei RETEROM de prelucrare a textelor în limba română, au fost definite modalitățile de comunicare între diversele module funcționale existente sau viitoare (http://www.racai.ro/p/reterom/rapoarte/1.6.pdf), precizându-se chiar limbajele de programare recomandate pentru realizarea viitoarelor module. Modulele NLP existente au fost testate, evaluate iar unele dintre ele modificate pentru a se asigura integrarea lor facilă în fluxul de prelucrări TEPROLIN. În situația în care aceeași funcționalitate a fost propusă de două sau mai multe module existente, performanțele acestora au fost comparate (acuratețe, viteză, ușurință de integrare în fluxul de prelucrare anvizajat) și modulul cel mai performant a fost ales pentru integrarea in platforma TEPROLIN. Au fost realizate adaptările necesare (activitatea 1.7 - http://www.racai.ro/p/reterom/rapoarte/1.7.pdf) pentru integrarea într-un flux de prelucrare coerent a tuturor modulelor NLP selectate ca cele mai performante dintre cele existente și analizate în activitatea 1.6.

Cele trei activități au fost documentate în rapoartele 1.5, 1.6 și 1.7, încărcate pe pagina web a proiectului. O altă prioritate în această etapă a fost construcția lexiconului ce va fi folosit de toate proiectele componente ale proiectului RETEROM. Un lexicon își dovedește utilitatea atât în recunoașterea vorbirii, cât și în sinteza sa. El conține forma scrisă a cuvintelor și pronunția lor. Transcrierile (componenta textuală a) corpusurilor bimodale disponibile la toți partenerii proiectului au fost colectate și analizate: de la ICIA și UAIC - **CoRoLa – oral**, de la UPB - corpusurile **RSC, SSC-train și SSC-eval**, de la UTCN – **SWARA**, **MARA** și **Adevărul.ro**. Corpusurile sunt diferite din mai multe puncte de vedere: domeniul în care se înscriu, convențiile de transcriere, corectitudinea transcrierilor, număr de vorbitori. Din toate aceste corpusuri au fost extrase cuvintele ocurente. Dacă ele nu existau în lexiconul existent deja la ICIA (lexicon numit tblwordform, care conține 1,2 milioane de forme), au intrat într-un proces de validare automată, iar atunci când nu a fost posibil, manuală. Pentru orice formă ocurentă validată a fost generată paradigma flexionară din care face parte, astfel încât lexiconul rezultat să acopere, de fapt, mai mult decât corpusurile de la care s-a pornit. Includerea în lexicon presupune adăugarea lemei, a etichetei morfosintactice, a despărțirii în silabe, a accentului și transcrierea fonetică. Această activitate, extrem de laborioasă a fost descrisă în livrabilul activității 1.8 http://www.racai.ro/p/reterom/rapoarte/1.8.pdf). Activitatea 1.9 de diseminare a fost realizată prin proiectarea și implementara site-ului proiectului (http://www.racai.ro/p/reterom), popularea sa cu informații relevante, precum și prin publicarea unor articole științifice la conferințe internaționale sau în revista cu largă audiență națională Market Watch. Publicațiile considerate în raportul nostru sunt doar acelea care au descris realizări în cadrul proiectului RETEROM și au inclus în textul apărut mulțumirile adresate instituției finanțatoare a proiectului.

1. Gradul de realizare a obiectivelor specifice pentru Etapa I-a

Ob. Pr2.1.5: ***Definirea specificațiilor funcționale și arhitecturale ale platformei integrate și configurabile de prelucrare a textelor***

Grad realizare: Obiectiv realizat integral

Rezultate:
- Raportul 1.5 care pune în evidență cerințele minimale obligatorii pentru realizarea unei platforme integrate și configurabile de prelucrare a textelor cu referire directă la obiectivele proiectului RETEROM.

Ob. Pr2.1.6: ***Definirea modulelor software și a serviciilor oferite de proiect; identificarea adaptărilor pentru modulele NLP existente și a modulelor noi necesare***

Grad realizare: Obiectiv realizat integral

Rezultate:
- Au fost identificate toate modulele operaționale utile pentru obiectivele proiectului RETEROM ale partenerilor, au fost testate, evaluate și selectate în forma inițială, și au fost evidențiate modificările necesare modulelor selectate pentru a funcționa în platforma obiectiv a proiectului RETEROM.

Ob. Pr2.1.7: ***Realizarea adaptărilor necesare pentru modulele NLP existente, identificate în activitățile 1.5 și 1.6***

Grad realizare: Obiectiv realizat integral

Rezultate:
- Au fost efectuate modificările sau extensiile codificărilor modulelor selectate in activitățile anterioare și integrate într-un prototip funcțional.

Ob. Pr2.1.8: ***Crearea și validarea (eventual cu corectările manuale necesare) a unui lexicon specific corpusului bimodal și încorporarea sa în lexiconul existent***

Grad realizare: Obiectiv realizat integral

Rezultate:
- A fost creat un lexicon mare destinat aplicațiilor de prelucrare a vorbirii (de peste 354.000 de intrări) conținînd pe lângă forma ocurență, lema, descrierea morfo-lexicală, informația specifică aplicațiilor de prelucrare a vorbirii: silabificare, plasare accent, transcriere fonetică; acest rezultat constituie o resursă fundamentală de largă utilitate

Ob. Pr2.: ***Diseminare (inclusiv site-ul proiectului complex și al proiectelor componente)***

Grad realizare: Obiectiv realizat integral

Rezultate:
- realizarea și actualizarea web site-ului proiectului[4]
- 2 articole la conferința CONSILR 2018, 1 articol la Conferința LREC 2018, 1 articol în revista Market Watch.
- 1 livrabil referitor la activitățile de diseminare (D1.9).

Descrierea științifică și tehnică

1. Introducere

Corpusul bimodal pentru limba română adnotat pe multiple niveluri (ce urmează a fi colectat, gestionat și exploatat prin platforma realizată în proiectul component COBILIRO) este în esență un corpus de text și voce în care fragmentele de text sunt împerecheate cu fișierele care înregistrează rostirea lor de către diverși vorbitori nativi ai limbii române. Un astfel de corpus este resursa fundamentală pentru învățarea automată a sistemelor de sinteză a vorbirii care produc automat rostirea în limba română a unor fraze arbitrare și a sistemelor de recunoaștere a vorbirii care traduc automat semnalul vocal în text. Aceste sisteme folosesc adnotări ale frazelor corpusului la diverse niveluri (fonetic, morfologic, sintactic și semantic) împreună cu alinieri ale cuvintelor cu semnalul vocal, pentru a învăța automat fie să sintetizeze o voce umană, fie să transcrie automat un semnal vocal. O platformă de prelucrare a textelor care să ofere prelucrările utile într-un corpus de vorbire trebuie să respecte cel puțin următoarele cerințe:

- Să fie standardizată, astfel încât noi adnotări să poată fi ușor adăugate;

[4] http://www.racai.ro/p/reterom/

- Să fie modulară și configurabilă, astfel încât programatorul să poată ușor substitui module în lanțurile de prelucrare, module care efectuează anumite operații cu o acuratețe mai bună;
- Să fie accesibilă ca serviciu web, încât mai multe echipe să o poată folosi simultan, fără a o instala.

2. Specificații arhitecturale ale platformei de prelucrare a textelor

Pentru că platforma de prelucrare a textelor pentru COBILIRO (dezvoltată în proiectul TEPROLIN) este o colecție eterogenă de aplicații de prelucrare a limbajului natural, scrise în diverse limbaje de programare, și pentru că dorim ca această platformă să fie disponibilă pe orice calculator cu o conexiune la Internet, fără a trebui să fie instalată[5], platforma *va fi oferită utilizatorului sub forma unui serviciu web de tip REST* (Fielding, 2000).

Serviciul web va expune programatorului o serie de funcții care realizează prelucrările oferite de platformă, funcții care pot fi folosite individual sau ca parte a unui lanț de prelucrare. Fiecare funcție va primi parametri de intrare fie de la utilizator, fie de la o altă funcție a platformei și va furniza un rezultat care va putea fi utilizat imediat sau ca un parametru pentru următoarea funcție din lanțul de prelucrare. Dacă utilizatorul nu dorește să utilizeze prelucrarea oferită de o funcție, acesta va putea specifica acest lucru la apelul funcției (printr-un parametru special) urmând ca funcția să prelucreze parametrii de intrare doar în sensul asamblării acestora în formatul rezultatului oferit.

Pentru a putea compune funcțiile PLN implementate de aplicațiile componente ale platformei TEPROLIN, *trebuie să standardizăm parametrii de I/O și rezultatele calculate de funcții*, standardizare care însă va fi transparentă utilizatorului. Vom folosi un format textual și tabular pentru parametrii de I/O și pentru rezultate, pentru a nu pierde timp de procesare cu parsarea unor formate structurate de tip XML. Formatul intern al parametrilor de I/O și al rezultatelor pe care l-am ales este CoNLL-X (Buchholz și Marsi, 2006) care va fi extins cu coloane suplimentare pentru a acomoda rezultatele furnizate de aplicațiile PLN componente.

În cadrul TEPROLIN *vom defini o serie de lanțuri de prelucrare (funcții compuse)* pe care le considerăm utile pentru proiectele componente ale ReTeRom (TADARAV - Tehnologii pentru adnotarea automată a datelor audio și pentru realizarea interfețelor de recunoaștere automată a vorbirii și SINTERO - Tehnologii de realizare a interfețelor om-mașină pentru sinteza text-vorbire cu expresivitate) cum ar fi de exemplu adnotarea cu etichete morfo-sintactice a unei fraze sau analiza gramaticală cu arbori de dependențe sintactice (pentru care existența adnotării cu etichete morfo-sintactice este o cerință strictă). Va exista de asemenea *un lanț de prelucrare complet* care va include toate tipurile de procesări pe care platforma le poate realiza.

O provocare pe care o întâlnim la realizarea unei platforme de prelucrare a textelor în limba română alcătuită dintr-o colecție eterogenă de aplicații de prelucrare a limbajului natural (PLN) este viteza de execuție a lanțurilor de prelucrare, mai ales a lanțului de prelucrare complet, în condițiile în care *platforma va fi utilizată și pentru prelucrări de text în timp real*. Funcțiile serviciului web sunt „wrapper”-e[6] ale aplicațiilor componente iar lanțul de prelucrare presupune preluarea rezultatelor unei funcții ca parametri de intrare pentru următoarea funcție din lanț. Aplicațiile componente pot fi executate doar în procese separate (fiecare cu interpretorul potrivit sau direct dacă sunt compilate) însă aceste procese au, de obicei, un cost ridicat (în timp de execuție) de pornire (au nevoie de timp pentru a încărca diverse resurse pe care le accesează cum ar fi dicționare, modele statistice, etc.) care nu permite pornirea lor la fiecare apel de funcție. Acest lucru implică păstrarea proceselor rezidente în memoria platformei și comunicarea cu acestea cu unul dintre mecanismele IPC[7] cum ar fi de exemplu „named pipes”, o metodă de comunicare inter-proces în care un fișier rezident în

---

[5] Instalarea ar fi anevoioasă pentru că, fiind compusă din aplicații dezvoltate independent, în diverse limbaje de programare interpretate, utilizatorul ar trebui să instaleze separat interpretoarele și bibliotecile de programe necesare, să citească manuale de instalare, etc.

[6] https://en.wikipedia.org/wiki/Wrapper_function

[7] https://en.wikipedia.org/wiki/Inter-process_communication

memoria RAM are rol de canal de comunicare (uni- sau bi-direcțional) sincronizat accesibil proceselor active. În acest context, *accesul concurent (pe mai multe fire de execuție sau de la mai mulți clienți HTTP) la platforma de prelucrare a textelor trebuie programat în serviciul web* astfel încât procesele aplicațiilor componente ale platformei să servească cereri secvențial. Astfel, în funcție de specificațiile serverului pe care găzduim platforma și serviciul web (număr de nuclee, memorie RAM) putem porni mai multe procese de același tip și servi cereri în ordinea în care apar (cererile sunt plasate într-o coadă de așteptare).
Rezumând cele expuse mai sus, platforma de prelucrare a textelor TEPROLIN:

a. Este un serviciu web REST cu suport pentru acces concurent pentru un număr fix de clienți (cu atât mai mare cu cât resursele serverului care găzduiește serviciul web permit);
b. Rulează aplicații PLN în procese rezidente în memorie, minimizând timpul de execuție al lanțurilor de prelucrare prin utilizarea mecanismelor IPC;
c. Definește lanțuri de prelucrare a textelor formate din compunerea funcțiilor dar permite utilizatorului și apelarea de funcții individuale prin standardizarea parametrilor de I/O.
d. Rezultatele lanțurilor de prelucrare vor fi oferite utilizatorului în format standardizat și interoperabil.

3. Specificații funcționale ale platformei de prelucrare a textelor

Platforma de prelucrare a textelor TEPROLIN va reuni o mare parte a aplicațiilor de PLN care sunt deja disponibile (și funcționale), dezvoltate de partenerii proiectului ReTeRom. Provocarea proiectului TEPROLIN este armonizarea tuturor acestor aplicații în platforma pe care am descris-o în secțiunea anterioară, printr-o soluție ușor extensibilă care să accepte module noi PLN și definiții ale unor noi lanțuri de prelucrare. Inventarul modulelor PLN existente care vor fi integrate în platforma TEPROLIN și modificările pe care le operăm pe fiecare modul pentru platforma TEPROLIN fac obiectul *Activității 1.6 - Definirea modulelor software și a serviciilor oferite de proiect; identificarea adaptărilor necesare pentru modulele PLN existente și a modulelor noi necesare platformei integrate și configurabile de prelucrare a textelor*.

Platforma TEPROLIN va include următoarele tipuri de preprocesări textuale atomice (nu neapărat în această ordine):

1. **Normalizarea textelor în limba română**: acest modul presupune identificarea diacriticelor românești non-standard (anume „ş” și „ţ”) și înlocuirea lor automată cu variantele standard (cele cu virgulă, anume „ș” și „ț”), normalizarea cratimei (înlocuirea tuturor variantelor Unicode ale cratimei cu cratima standard „-”), normalizarea spațiilor (înlocuirea tuturor variantelor Unicode ale cratimei cu caracterul spațiu standard „ ”), și alte tipuri de normalizări;
2. **Inserarea automată a diacriticelor românești în texte**: acest modul va identifica cuvintele care nu conțin diacritice într-o frază și *va insera automat variantele acestora cu diacritice, în context*.
3. **Despărțirea în silabe a cuvintelor**: acest modul va despărți în silabe cuvintele unei fraze și va atașa despărțirea în silabe fiecărei unități lexicale a frazei. De exemplu pentru „camion”, despărțirea în silabe este „ca-mi-on”;
4. **Poziționarea accentului**: acest modul va identifica silaba care poartă accentul cuvântului.
5. **Transcriere fonetică**: acest modul va „traduce” automat din forma scrisă a cuvântului în reprezentarea fonetică a sa.
6. **Expandarea numerelor în termeni numerali**: acest modul va transcrie automat orice numeral scris în cifre în echivalentul său literal.
7. **Expandarea abrevierilor**: acest modul va transcrie automat, în context, o serie de abrevieri în formele lor extinse.
8. **Segmentare la nivel de frază**: acest modul va despărți un text dat (paragraf sau document) în frazele componente.

9. **Segmentare la nivel de unitate lexicală**: acest modul va identifica toate cuvintele și semnele de punctuație dintr-o frază și va oferi lista (ordonată) în care acestea apar în frază.
10. **Adnotare cu etichete morfo-sintactice („POS tagging”)**: acest modul va face analiza morfo-sintactică, în context, a fiecărei unități lexicale a frazei.
11. **Lematizare**: acest modul identifică formele standard de dicționar ale cuvintelor frazei.
12. **Identificarea constituenților sintactici („Chunking”)**: acest modul detectează limitele (indecșii cuvintelor care le alcătuiesc) grupurilor nominale („o floare albastră”, „mașina de cusut”), verbale („au fost făcute”, „se vor fi dus”), adjectivale („foarte frumoasă”) și adverbiale („destul de repede”) în frază.
13. **Analiza sintactică cu relații de dependență („Dependency parsing”)**: acest modul construiește automat un arbore de analiză sintactică a frazei în care se precizează subiectul propoziției, complementele predicatului, etc.

Operațiile atomice de mai sus pot fi definite ca niște metode Java cu parametri descriși pe larg în http://www.racai.ro/p/reterom/rapoarte/1.5.pdf. Lanțul lexical complet se va face prin apelarea înlănțuită a tuturor operațiilor de mai sus, de la operația nr. 13 la operația nr. 1. În continuare prezentăm informațiile relevante pentru modulele incluse în prototipul fluxului de prelucrare propus și aflat în curs de implementare în proiectul component TEPROLIN.

DiacriticsRestoration[8] și javaNLP2[9]

Aceste două module Java sunt puse la dispoziția proiectului de Universitatea „Politehnica” din București (UPB) și implementează următoarele operații:

- Inserarea automată a diacriticelor românești în texte: este realizată de metoda
  public static void diacritics_process(String wordName)
  din clasa DiacriticsProcess.java a modulului DiacriticsRestoration (Ivan, 2016). Această metodă trebuie adaptată platformei TEPROLIN pentru a elimina un apel SRILM[10] care încarcă modelul de limbă pentru fiecare utilizare a metodei diacritics_process. De asemenea, metoda trebuie să poată lucra și cu fraze, nu numai cu fișiere, așa cum lucrează în prezent.
- Normalizarea textelor în limba română: este realizată de clasa CharactersNormalizer.java din modulul javaNLP2 dar care trebuie adaptată să utilizeze diacriticele românești actuale.
- Expandarea numerelor în termeni numerali: este realizată de clasa NumbersHandler.java din modulul javaNLP2; se poate utiliza în platforma TEPROLIN fără adaptări.
- Expandarea abrevierilor: este realizată de clasa AbbreviationsReplacer.java din modulul javaNLP2; se poate utiliza în platforma TEPROLIN fără adaptări.

Romanian TTS[11]

Acest modul Python (3.6) a fost furnizat proiectului de Universitatea Tehnică din Cluj-Napoca (UTCN) și implementează următoarele operații:

- Despărțirea în silabe a cuvintelor.
- Poziționarea accentului.
- Transcriere fonetică.

Toate aceste apeluri se fac pe niște structuri de date care au fost pregătite în prealabil de alte apeluri de metode. Operațiile menționate mai sus sunt implementate în scriptul text_processing_av.py iar adaptarea acestui script la platforma TEPROLIN va include studierea conținutului structurilor de date necesare acestor metode astfel încât acesta să poată fi pregătit de platformă.

---

[8] http://git.speed.pub.ro/iivan/DiacriticsRestoration.git

[9] http://git.speed.pub.ro/common/javaNLP2.git

[10] http://www.speech.sri.com/projects/srilm/

[11] http://www.romaniantts.com/

NLP-Cube[12]

NLP-Cube (Boroș et al., 2018) este un modul open-source, scris în Python (3.6) care realizează procesarea textelor la următoarele niveluri:

- Segmentare la nivel de frază.
- Segmentare la nivel de unitate lexicală.
- Adnotare cu etichete morfo-sintactice („POS tagging").
- Analiza sintactică cu relații de dependență („Dependency parsing").

Operațiile de mai sus sunt implementate cu ajutorul unor rețele neuronale recursive care sunt antrenate pe treebank-ul românesc, adnotat cu relații sintactice de dependență universale în cadrul proiectului SSPR[13] (Barbu Mititelu et al., 2016), disponibil pe Internet la adresa http://universaldependencies.org/. Singura îmbunătățire a acestui modul pentru includerea sa în platforma TEPROLIN constă în adăugarea unui lexicon românesc cu leme astfel încât lematizarea învățată automat să intervină doar în cazul în care cuvântul care se lematizează nu se află în lexicon. În cazul în care cuvântul se află în lexicon, împreună cu eticheta morfo-sintactică atribuită, lema se va recupera din acest lexicon.

TTL[14]

TTL (Ion, 2007) este un modul Perl (5.14) care implementează aceleași operații de procesare a textelor ca NLP-Cube dar care adaugă o operație suplimentară:

- Identificarea constituenților sintactici („Chunking"). TTL poate fi integrat în platforma TEPROLIN fără alte modificări.

4. Adaptarea modulelor PLN pentru platforma TEPROLIN

Având în vedere că cele mai multe operații ale platformei TEPROLIN sunt implementate în Python 3, platforma TEPROLIN va fi scrisă în limbajul de programare Python 3 iar serviciile web REST vor fi oferite de serverul Flask[15].

Modulele care nu sunt scrise în Python vor implementa o interfață care le va permite să-și încarce resursele de care au nevoie și să comunice cu platforma printr-un canal de comunicare de tip „named pipe" (vezi raportul tehnic al Activității 1.5 pentru descrierea arhitecturii platformei).

5. Lexicon specific corpusului bimodal

Un lexicon își dovedește utilitatea atât în recunoașterea vorbirii, cât și în sinteza sa. El conține forma scrisă a cuvintelor și pronunția lor. Transcrierile (componenta textuală a) corpusurilor bimodale disponibile la toți partenerii proiectului au fost colectate și analizate: de la ICIA și UAIC - **CoRoLa – oral**, de la UPB - corpusurile **RSC, SSC-train și SSC-eval**, de la UTCN – **SWARA**, **MARA** și **Adevărul.ro**. Corpusurile sunt destul de diferite din mai multe puncte de vedere: domeniul în care se înscriu, convențiile de transcriere, corectitudinea transcrierilor, număr de vorbitori. Din toate aceste corpusuri au fost extrase cuvintele ocurente (pentru detalii, vezi raportul științific al activității 1.8). Dacă ele nu existau în lexiconul existent deja la ICIA (lexicon numit tblwordform, care conține 1,2 milioane de forme), au intrat într-un proces de validare automată, iar atunci când nu a fost posibil, manuală. Pentru orice formă ocurentă validată a fost generată paradigma flexionară din care face parte, astfel încât lexiconul rezultat să acopere, de fapt, mai mult decât corpusurile de la care s-a pornit. Includerea în lexicon presupune adăugarea lemei, a etichetei morfosintactice, a despărțirii în silabe,

---

[12] https://github.com/adobe/NLP-Cube

[13] http://dev.racai.ro/ti/wordpress/index.php/project/

[14] http://ws.racai.ro/ttlws.wsdl

[15] http://flask.pocoo.org/

a accentului și transcrierea fonetică (ultimele trei realizate cu instrumentul Romanian TTS, descris mai sus). Au fost identificate peste 354.000 de intrări pentru acest lexicon (http://www.racai.ro/p/reterom/rapoarte/1.8.pdf), dintre care doar 8000 au corespuns cuvintelor absente din tblwordform. Pe măsură ce vom mai aduna și alte texte pentru acest proiect, lexiconul va fi îmbogățit, în etapele viitoare.

6. Activități în celelalte proiecte componente

Cercetătorii ICIA au participat la proiectul component COBILIRO astfel: la activitatea 1.1 cu documentarea despre corpusurile bimodale existente la nivel mondial, caracteristicile acestora și redactarea raportului științific asupra acestei activități; la activitatea 1.2 au contribuit cu informații despre corpusul oral existent la ICIA, conform formatului întocmit de cercetătorii UAIC.

Cercetătorii ICIA au participat la proiectul component SINTERO astfel: la activitatea 1.15 cu documentarea și redactarea unei secțiuni din raportul aferent, și anume „Manifestarea prozodiei la nivel lingvistic". S-a discutat aici despre trăsăturile prozodice (intonația, accentul, pauza), despre rolul lor în comunicare (tipuri de informație transmisă, clasificarea tiparelor intonaționale, funcțiile intonației) și despre modul în care acestea contribuie la dezambiguizarea lingvistică. Pentru activitatea 1.16 cercetătorii ICIA au selectat și au pus la dispoziția partenerilor UTCN mostre de date din corpusul COROLA clasificate pe stiluri literare (Juridic-Administrativ, Beletristic, Memorialistic, Publicistic și Științific), în total 5 milioane de cuvinte (forme ocurență, incluzând punctuația), câte 1 milion din fiecare stil.

Rapoartele detaliate prevăzute pentru primul an sunt disponibile pe site-ul proiectului.

7. Diseminarea rezultatelor.

Site-ul proiectului este funcțional, ca singur punct de acces la http://www.racai.ro/p/reterom/, de unde se pot accesa informații despre fiecare proiect component din site-ul propriu. Din motive de securitate s-a optat pentru soluția implementarii câte unui site la fiecare instituție participantă în proiect, aceste situri fiind accesibile din situl instituției coordonatoare. Arhitectura celor patru situri este unitară, completarea cu informații specifice fiind în sarcina fiecărei instituții partenere.

Proiectul ReTeRom și primele sale rezultate au fost diseminate prin publicații la conferințe relevante și reviste, fiind mentionată finațarea publică prin contractul **73PCCDI/2018.** Lucrările în limba engleză au inclus textul "This work is supported by a grant of the Ministry of Research and Innovation CCCDI-UEFISCDI, project cod PN-III-P1-1.2-PCCDI-2017-0818 within PNCDI III."

Dan Tufiș, Dan Cristea (2018) "A Bird's-eye View of Language Processing Projects at the Romanian Academy", in Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)", Miyazaki, Japan, pp. 2445-2451. (conferință ISI)

Dan Tufiș (2018) "CoRoLa Primul corpus computațional de referință pentru limba română contemporană", Market Watch, no. 205, iunie 2018, pp. 28-29.

Vasile Păiș, Dan Tufiș (2018) "More Romanian word embeddings from the ReTeRom project". In Proceedings of the 13th International Conference CONSILR 2018, Iași,22-23 November

Radu Ion (2018) "TEPROLIN: An Extensible, Online Text Preprocessing Platform for Romanian". In Proceedings of the 13th International Conference CONSILR 2018, Iași,22-23 November

8. Structura ofertei de servicii de cercetare şi tehnologice

Institutul de Cercetări pentru Inteligență Artificială (ICIA/RACAI), va oferi pe platforma ERRIS serviciile de cercetare şi tehnologice enumerate în tabelul următor (https://erris.gov.ro/RACAI-ICIA):

Serviciu	Detalii
Interogare corpus de referință al limbii române	corola.racai.ro
TTL	http://ws.racai.ro/ttlws.wsdl
Modular Language Processing for Lightweight Applications (MPLA) - cu prelucrări pentru mai mult de 40 de limbi	http://slp.racai.ro/index.php/mlpla-new/
Platforma TEPROLIN	www.racai.ro/p/reterom
Sistem de detecție de cuvinte cheie în conversații înregistrate	http://heimdall.racai.ro/
Romanian Spoken Language Processing	rsp.racai.ro
Romanian Anonymous Speech Corpus	rasp.racai.ro
Sistemul online de sinteza a vorbirii de la RACAI	http://www.racai.ro/about-us/sinteza-vorbirii-pornind-de-la-text/

9. Locuri de muncă susţinute prin program

Pe lângă cercetătorii cu experiență (Dan Tufiș -CS1, Verginica Barbu Mititelu – CS2, Radu Ion -CS3, Elena Irimia CS3, Eric Curea – CS, Maria Carp -ACS) la proiect a participat cu un aport semnificativ **Vasile Păiș,** nou angajat pe pozitie de doctorand. Din păcate, la începutul proiectului au părasit, din motive personale, echipa de cercetare doi cercetători foarte valoroși (Tiberiu Boroș - CS2 și Ștegan Dumitrescu – CS3). Responsabilitățile lor au fost preluate de unii dntre cercetători cu experiență și de noul angajat.

10. Valorificarea şi îmbunătăţirea competenţelor şi resurselor existente la nivelul consorțiului

La nivelul proiectului TEPROLIN nu au fost valorificate cecurile anului 2018

Referințe bibliografice

Barbu Mititelu, Verginica, Ion, Radu, Simionescu, Radu, Irimia, Elena and Perez, Cenel-Augusto (2016). The Romanian Treebank Annotated According to Universal Dependencies. In Proceedings of HrTAL2016, Dubrovnik, Croatia, 29 September - 1 October 2016.

Ștefan Daniel Dumitrescu, Tiberiu Boroș, Dan Tufiș (2017). RACAI's Natural Language Processing pipeline for Universal Dependencies. Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pages 174–181, Vancouver, Canada, August 3-4, 2017. http://universaldependencies.org/conll17/proceedings/

Tiberiu Boroș, Ștefan Daniel Dumitrescu and Ruxandra Burtica. (2018). NLP-Cube: End-to-end raw text processing with neural networks. Proceedings of the CoNLL-2018 Shared Task "Multilingual Parsing from Raw Text to Universal Dependencies", pages 171—179,Brussels, October 31 - November 1, 2018, Belgium, https://doi.org/10.18653/v1/K18-2017

Buchholz, Sabine and Marsi, Erwin (2006). CoNLL-X shared task on Multilingual Dependency Parsing. In Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL-X), pages 149–164, New York City, June 2006. © 2006 Association for Computational Linguistics

Fielding, Roy Thomas (2000). "Chapter 5: Representational State Transfer (REST)". *Architectural Styles and the Design of Network-based Software Architectures* (Ph.D.). University of California, Irvine.

Radu Ion. 2007. Word Sense Disambiguation Methods Applied to English and Romanian. PhD thesis (in Romanian). Romanian Academy, Bucharest.

Andra-Irina Ivan. 2016. RESTAURAREA DE DIACRITICE ÎN FIȘIERE TEXT COMPLEXE. Lucrare de diplomă, Facultatea de Electronică, Telecomunicații și Tehnologia Informației, Universitatea Politehnica București.

Tufiș, Dan (2000). Using a Large Set of EAGLES-compliant Morpho-Syntactic Descriptors as a Tagset for Probabilistic Tagging. International Conference on Language Resources and Evaluation LREC'2000, Athens, 2000, pp. 1105-1112

PROIECTUL TADARAV

Alexandru-Lucian Georgescu, Cristian Manolache, Gheorghe Pop, Dan Oneaţă,

Horia Cucu, Dragoş Burileanu, Corneliu Burileanu

Responsabil: Universitatea Politehnica București, coordonator Prof. Corneliu Burileanu

1 Rezumatul etapei

Prima etapă a proiectului TADARAV a avut două obiective principale: (i) actualizarea cunoştinţelor consorţiului privind ultimele descoperiri şi invenţii în domeniul adnotării automate a datelor audio şi (îi) evaluarea posibilităţii de utilizare a unor sisteme de recunoaştere automată a vorbirii (RAV) complementare în vederea adnotării automate a semnalului de vorbire. Cele două obiective au fost realizate în proporţie de 100%, în urma activităţilor întreprinse rezultând toate livrabilele asumate de consorţiu la începutul acestei etape.

Concret, etapa 1 / 2018 a proiectului TADARAV a debutat cu cele trei activităţi de studiu al metodelor din literatură privind adnotarea automată a datelor audio. Activităţile 1.10, 1.11 şi 1.12 au fost finalizate în luna iunie cu cele trei rapoarte ştiinţifice asumate, după cum urmează:

- Studiu privind starea artei: Sisteme complementare de recunoaştere automată a vorbirii
- Studiu privind starea artei: Alinierea transcrierilor aproximative cu semnalul de vorbire
- Studiu privind starea artei: Estimarea scorurilor de încredere pentru sistemele de recunoaştere automată a vorbirii

Etapa a continuat cu activitatea 1.13 şi anume proiectarea şi implementarea unei soluţii de bază de adnotare automată a semnalului de vorbire utilizând sisteme de RAV complementar. În cadrul acestei activităţi echipa de implementare a achiziţionat din mediul *online* câteva sute de ore de documente audio ce conţin vorbire şi le-a adnotat în mod automat folosind metodologia de adnotare proiectată şi implementată. În acest moment, soluţia se află la nivelul tehnologic asumat prin propunerea de proiect (TRL3), fapt certificat de rezultatele experimentale de transcriere a vorbirii obţinute folosind noile date audio adnotate automat (a se vedea secţiunea 2 pentru detalii).

Diseminarea rezultatelor proiectului a fost realizată în cadrul consorţiului în cele două workshop-uri organizate pe parcursul acestei etape (Bucureşti: 7 – 8 iunie 2018, şi Iaşi: 22 – 23 noiembrie 2018) şi în comunitatea ştiinţifică la *The 41st International Conference on Telecommunications and Signal Processing (TSP)*. Articolul publicat la această conferinţă, ale cărui date complete sunt indicate mai jos, este în prezent indexat *IEEE Xplore* şi în curs de indexare *Thompson Reuters CPCI* (ISI), iar numele finanţatorului este menţionat în secţiunea *Acknowledgement*, conform indicaţiilor din contractul de finanţare. Chiar dacă articolul nu este încă indexat ISI, avem certitudinea că el va fi indexat în lunile următoare, la fel cum au fost indexate ediţiile 2008 – 2017 ale acestei conferinţe de prestigiu.

- Alexandru-Lucian Georgescu, Horia Cucu, "Automatic annotation of speech corpora using complementary GMM and DNN acoustic models," în *the Proceedings of the 41st International Conference on Telecommunications and Signal Processing (TSP)*, 2018, Athens, Greece.

2 Descrierea ştiinţifică şi tehnică a activităţilor

2.1 Activitatea 1.10 – Studiul metodelor din literatură privind utilizarea sistemelor de RAV complementare pentru generarea automată de adnotări

În cadrul acestei activităţi, echipa de cercetare a UPB a studiat peste 30 de lucrări ştiinţifice relevante, dintre care a selectat aproximativ 20 de lucrări care vizează direct subiectul utilizării sistemelor de RAV complementare în vederea generării automate de adnotări. Aceste lucrări au fost analizate în detaliu şi, în

urma analizei bibliografice, a rezultat un raport de 21 de pagini despre starea cunoaşterii în acest domeniu (www.racai.ro/p/reterom/rapoarte/1.10).

Ideea utilizării sistemelor RAV complementare este una fiabilă în contextul adnotării automate a corpusurilor de vorbire. Diversitatea sistemelor poate fi dată de mai mulţi factori, cei mai importanţi fiind compoziţia seturilor de antrenare acustică/lingvistică, tipul trăsăturilor extrase, tipul modelelor antrenate, tipul algoritmilor utilizaţi la antrenare/decodare. Nu se poate afirma cu certitudine că vreuna dintre metode oferă complementaritate mai bună, fiecare putând fi eficientă într-o situaţie particulară.

În ceea ce priveşte metodele de combinare şi selecţie a transcrierilor obţinute de RAV, se observă faptul că metoda *Recognizer Output Voting Error Reduction* (ROVER) este o procedură clasică, ce a fost aplicată cu succes în multe situaţii. O altă metodă clasică este cea a utilizării scorurilor de încredere, deşi nu întotdeauna sistemele RAV scot la ieşire scoruri corelate cu corectitudinea transcrierii. Totuşi, alternativele ce se bazează pe tehnici de învăţare automată (clasificatori de tip *Conditional Random Field* – CRF, reţele neurale etc.) au început să fie utilizate în ultima vreme, iar rezultatele sunt mai bune în comparaţie cu metodele clasice. Metoda stabilităţii acustice (A-stabil) este derivată din ROVER şi a fost propusă pentru situaţia în care se folosesc sisteme RAV pentru mai multe limbi sursă diferite faţă de limba ţintă.

2.2 Activitatea 1.11 – Studiul metodelor din literatură pentru alinierea transcrierilor aproximative cu semnalul de vorbire

În cadrul acestei activităţi, echipa de cercetare a UPB a studiat peste 50 de lucrări ştiinţifice relevante, dintre care a selectat aproximativ 30 de lucrări care vizează direct subiectul alinierii transcrierilor aproximative cu semnalul de vorbire. Aceste lucrări au fost analizate în detaliu şi, în urma analizei bibliografice, a rezultat un raport de 25 de pagini despre starea cunoaşterii în acest domeniu (www.racai.ro/p/reterom/rapoarte/1.12).

Sarcina de aliniere transcrieri aproximative – vorbire este abordată cu două clase de metode: cele care aliniază transcrierea aproximativă direct la semnalul vocal (aliniere text-audio) şi cele care aliniază transcrierea aproximativă la reprezentarea textuală a semnalului vocal, obţinută cu un sistem de RAV (aliniere text-text). Aceste clase nu sunt exclusive în sensul că există şi metode care combină ambele idei. Cele mai multe metode constau într-o serie de paşi de aliniere succesivă şi etape de segmentare a semnalului vocal prin împărţirea în unităţi mai mici. Alinierea este adesea bazată pe algoritmi de aliniere generici care sunt optimizaţi cu algoritmi de programare dinamică. Din păcate, este dificil de concluzionat care dintre metode funcţionează cel mai bine pentru că seturile de date şi metodologia de lucru variază – aceste motive fac imposibilă o comparaţie echitabilă. Din punctul de vedere al proiectului, cele mai promiţătoare lucrări sunt cele care folosesc transcrierile foarte imprecise, în principal pentru că genul acesta de transcrieri sunt cele mai uzuale în practică: date încărcate de utilizatori şi date descărcate de pe internet.

2.3 Activitatea 1.12 – Studiul metodelor din literatură pentru generarea scorurilor de încredere (SI) pentru recunoaşterea automată a vorbirii (RAV)

În cadrul acestei activităţi, echipa de cercetare a UPB a studiat peste 60 de lucrări ştiinţifice relevante, dintre care a selectat aproximativ 45 de lucrări care vizează direct subiectul generării scorurilor de încredere pentru RAV. Aceste lucrări au fost analizate în detaliu şi, în urma analizei bibliografice, a rezultat un raport de 24 de pagini despre starea cunoaşterii în acest domeniu (vezi livrabilul www.racai.ro/p/reterom/rapoarte/1.11).

Estimarea încrederii (EI) prezintă o importanţă majoră în îmbunătăţirea sistemelor de RAV, contribuind la procesul de detecţie şi corectare a erorilor. De asemenea metodele de estimare a încrederii care generează scoruri de încredere pentru transcrierile RAV sunt utilizate pe scară largă pentru auto-antrenarea modelelor acustice pentru RAV. Metodele de generare a scorurilor de încredere se încadrează în trei mari categorii:

abordări bazate pe clasificare, pe probabilităţi posteriori şi pe verificarea enunţurilor. Pentru evaluarea scorurilor de încredere se folosesc curbele ROC (*Receiver Operating Characteristic*) şi DET (*Detection Error Tradeoff*) precum şi indicele *Normalized Cross-Entropy* (NCE).

Dintre toate metodele de EI studiate, cele mai slabe rezultate au fost înregistrate de abordările bazate pe probabilităţile aposteriori, întrucât metodele actuale nu reuşesc să estimeze corect distribuţia aposteriori. Cele mai bune performanţe au fost obţinute în contextul utilizării unor trăsături predictive, ce pot fi extrase din laticea de recunoaştere, din modelul de limbă, sau din informaţia acustică propriu-zisă. Aceste trăsături sunt ulterior introduse la intrarea unui clasificator. În literatura de specialitate au fost propuşi o serie de clasificatori, dintre care s-au remarcat sistemele bazate pe CRF şi cele bazate pe reţele neurale recurente. Cea mai promiţătoare direcţie de dezvoltare în prezent o reprezintă utilizarea de reţele neurale recurente bidirecţionale (BiDRNN), ce folosesc în procesul de decizie un context stânga-dreapta. Sistemele BiDRNN au depăşit performanţele celor bazate pe CRF, demonstrând o putere mai mare de generalizare.

2.4 Activitatea 1.13 – Proiectarea şi implementarea unei soluţii de bază de adnotare automată a semnalului de vorbire utilizând sisteme de RAV complementar

Modelele acustice bazate pe reţele neurale profunde (*Deep Neural Network* – DNN) obţin performanţe direct proporţionale cu cantitatea de date folosite la antrenarea reţelei. Prin urmare, dat fiind faptul că adnotarea manuală a resurselor audio presupune o investiţie consistentă de efort şi timp, interesul faţă de tehnicile de adnotare automată a vorbirii a crescut semnificativ. Adnotarea automată a vorbirii presupune colectarea de vorbire în format brut şi folosirea unei metode automate pentru a produce transcrieri cât mai precise pentru cel puţin o parte din corpusul iniţial.

Una dintre metodele de adnotare automată presupune folosirea unui număr de sisteme de RAV pentru generarea mai multor transcrieri aproximative, urmând ca apoi transcrierile obţinute să fie aliniate iar părţile identice să fie considerate corecte. Aceasta este şi metoda folosită în etapa curentă. Două sisteme de RAV complementare au fost antrenate cu acelaşi corpus de vorbire în limba română, de două tipuri: vorbire citită şi spontană. Diferenţa dintre sisteme constă însă în modelul acustic folosit: HMM-GMM (*Hidden Markov Model – Gaussian Mixture Model*), respectiv HMM-DNN. În acest fel, s-a plecat de la presupunerea că erorile de transcriere vor fi necorelate, probabilitatea ca ambele sisteme să greşească identic fiind scăzută. Această ipoteză experimentală a fost validată ulterior cu rezultate concrete.

2.4.1 Descrierea metodei

Metoda utilizată în această etapă are ca scop obţinerea într-un mod automat, nesupervizat, a unei adnotări cât mai precise pentru un corpus de vorbire. Corpusul nou obţinut s-a dorit a fi utilizat pentru antrenarea sistemelor de RAV existente, crescând astfel variabilitatea acustică a modelelor, îmbunătăţind implicit şi acurateţea transcrierilor.

Ideea principală a acestei metode de adnotare automată constă în utilizarea a două sisteme RAV pentru a produce transcrieri pentru un corpus neadnotat, urmând ca apoi transcrierile să fie aliniate, iar părţile identice să fie selectate ca fiind corecte. În final, transcrierile selectate şi segmentele de vorbire corespunzătoare sunt folosite pentru a forma un nou corpus adnotat de vorbire.

Pentru ca această metodă să funcţioneze este esenţial ca cele două sisteme RAV să fie complementare. Mai exact, erorile celor două sisteme RAV trebuie să fie necorelate. Există câteva opţiuni care fac ca acest lucru să fie posibil: modelele acustice sau modelele lingvistice să fie antrenate pe date diferite, algoritmii de decodare să fie diferiţi etc.

În abordarea curentă, complementaritatea sistemelor constă în:

- Tipul modelului acustic (HMM-GMM vs. HMM-DNN);

- Dimensiunea vocabularului (64k cuvinte vs. 200k cuvinte);
- Modelul de limbă folosit la decodare (3-gram vs. 2-gram);
- Utilizarea tehnicii de reevaluare lingvistică (fără reevaluare vs. reevaluare folosind model de limbă 4-gram).

În acest context, primul pas a presupus crearea sistemelor RAV complementare, aşa cum a fost menţionat anterior. Al doilea pas a fost colectarea de vorbire neadnotată. Materialele produse zilnic de *mass-media* (emisiuni, ştiri, interviuri, reportaje) sunt o foarte bogată sursă de vorbire. Publicarea lor în mediul *online* permite o implementare automată a procesului; strângerea unei cantităţi semnificative de date devine doar o chestiune de timp, fără vreun efort suplimentar. Al treilea pas a constat în obţinerea transcrierilor folosind cele două sisteme RAV. Transcrierile ipotetice au fost aliniate folosind un algoritm bazat pe *Dynamic Time Warping* (DTW), fiind astfel selectate părţile identice. Aşa cum experimentele arată, probabilitatea ca ambele sisteme să producă erori identice este foarte mică. Secvenţe consecutive de cuvinte, ce conţin un număr de cuvinte mai mare decât un prag determinat experimental, sunt considerate a fi corect transcrise. Un alt criteriu utilizat la selecţia transcrierilor este durata secvenţelor audio, fiind necesar ca aceasta să depăşească un anumit prag. De asemenea, distanţa în timp între două cuvinte este luată în calcul pentru a asigura faptul că nu există cuvinte intermediare ne-transcrise. La final, după ce secvenţele de cuvinte corecte au fost selectate, pe baza ştampilelor de timp oferite de cel mai bun dintre cele două sisteme RAV, s-a recurs la tăierea din datele audio iniţiale a părţilor corespunzătoare secvenţelor aliniate.

2.4.2 Evaluarea metodei

Utilitatea metodelor de adnotare automată poate fi măsurată folosind două criterii:

- Cantitatea de vorbire obţinută în urma alinierii, raportat la dimensiunea totală a corpusului iniţial;
- Calitatea adnotării, măsurabilă în eroarea la nivel de cuvânt (WER) şi/sau caracter (ChER).

Aceste metrici de performanţă pot fi calculate folosind un corpus de vorbire pentru care există deja o adnotare de referinţă, împreună cu ştampile de timp la nivel de cuvânt. Această transcriere de referinţă poate fi obţinută în două feluri: prin crearea ei în mod manual sau prin folosirea unui sistem RAV care să realizeze alinierea forţată a corpusului de evaluare. Deşi a doua metodă pare să fie predispusă la a genera erori, dacă sistemul RAV este destul de performant, alinierea forţată are o acurateţe suficient de bună. Astfel, corpusul este considerat la început a fi un corpus neadnotat şi va fi transcris cu ambele sisteme RAV, fiind obţinute transcrieri ipotetice ce vor fi aliniate. Pe baza ştampilelor de timp din părţile aliniate, sunt selectate părţile corespunzătoare din transcrierea de referinţă. În acest mod, WER şi ChER pot fi calculate între cele două seturi de text. De asemenea, tot pe baza ştampilelor de timp pentru părţile aliniate, se poate calcula durata segmentelor de vorbire selectate.

2.4.3 Pregătirea experimentelor

Această secţiune prezintă resursele principale utilizate la implementarea metodei de adnotare automată propusă. Sunt oferite detalii despre corpusurile de vorbire folosite, atât pentru antrenare cât şi pentru evaluarea sistemelor RAV, dar şi despre corpusul de vorbire neadnotat. Cele două sisteme RAV sunt prezentate din punct de vedere al performanţelor şi al modelelor acustice şi lingvistice.

Seturile de date de vorbire

Pentru antrenarea şi evaluarea sistemelor RAV, au fost folosite două mari corpusuri de vorbire în limba română: *Read Speech Corpus* (RSC), ce conţine vorbire citită, colectată în condiţii de laborator, fără zgomot de fundal şi *Spontaneous Speech Corpus* (SSC), ce conţine vorbire continuă, spontană, preluată de la posturi de radio şi TV, uneori afectată de zgomot. Ambele corpusuri cuprind fişiere audio şi transcrieri corespunzătoare şi sunt divizate în seturi de antrenare şi seturi de evaluare. RSC-train este setul de antrenare din RSC, ce conţine 100 ore de vorbire citită, cuvinte izolate sau fraze de la 157 de vorbitori diferiţi. RSC-eval este setul de

evaluare din RSC; acesta conţine vorbire de la 22 de vorbitori diferiţi, însumând 5.5 ore de vorbire. SSC-train este setul de antrenare din SSC şi conţine 130 ore de vorbire spontană, majoritatea din emisiuni de ştiri şi *talkshow*-uri. SSC-eval este setul de evaluare din SSC şi însumează 3.5 ore de vorbire.

Primul corpus de vorbire neadnotat a fost achiziţionat din mass-media românească, mai exact de la 3 *website*-uri de ştiri şi un post de radio într-o perioadă de o lună calendaristică. Prin parcurgerea *feed*-urilor RSS al acestor *website*-uri, au fost extrase fişierele audio, eşantionate la 16 kHz, 16 biţi pe eşantion.

Al doilea corpus de vorbire neadnotat a fost achiziţionat de asemenea din cele 3 surse din mass-media românească, într-o perioadă de nouă luni calendaristice.

A. Modelele acustice

Sistemele RAV utilizate în această abordare sunt bazate pe modele acustice ce aparţin de două paradigme diferite: HMM-GMM şi HMM-DNN. Ambele sisteme folosesc trăsături acustice de tipul coeficienţilor cepstrali, mai exact vectori de 13 elemente, împreună cu derivatele lor de ordinul 1 şi 2. Coeficienţii au fost extraşi cu o fereastră Hamming de 25 ms, cu suprapuneri 50%. Modelul acustic bazat pe DNN este de asemenea antrenat pe bazat unor i-vectori de dimensiune 100. Ambele sisteme modelează 36 de foneme dependente de context din limba română.

Modelul acustic bazat pe HMM-GMM cuprinde 4.000 de stări acustice (senone), fiecare dintre acestea fiind modelată de un număr de 128 densităţi Gaussiene. Modelul acustic bazat pe DNN este construit folosind alinierile obţinute cu un model HMM-GMM. Arhitectura DNN folosită este de tipul *Time Delay Neural Network*, modelând dinamica temporală a semnalului. Stratul de intrare al reţelei constă în 3.500 neuroni, ce procesează câte 9 cadre de semnal la un moment de timp. Reţeaua are 6 straturi ascunse şi 1200 neuroni pe fiecare strat. Stratul de ieşire cuprinde 350 neuroni. Modelul a fost antrenat pe parcursul a 5 epoci, cu o rată de învăţare iniţială de 0.015 şi a rata finală de învăţare de 0.00015. Setul de antrenare a fost divizat în mini-loturi de mărime 512.

B. Modelele lingvistice

Modelele lingvistice folosite în ambele sisteme RAV sunt de tip probabilistic – modele de limbă *n*-gram. Textele folosite la crearea acestora provin din ştiri *online* şi transcrieri ale unor conversaţii. Acestea au fost interpolate cu o pondere de 0.5. Primul corpus conţine 315M cuvinte, în timp ce al doilea conţine 40M cuvinte. Sistemul RAV bazat pe HMM-GMM foloseşte un model de limbă 3-gram ce conţine 64k cuvinte, în timp ce sistemul RAV bazat pe DNN foloseşte un model de limbă 2-gram cu 200k cuvinte pentru decodare, respectiv un model de limbă 4-gram cu 200k cuvinte pentru reevaluarea lingvistică.

2.4.4 Rezultate experimentale

A. Evaluarea metodei de adnotare automată

Evaluarea metodei de adnotare automată a fost realizată prin compararea transcrierii ipotetice rezultată în urma aplicării metodei pe corpusul de evaluare, cu transcrierea de referinţă. Transcrierile generate automat nu acoperă întreg corpusul de evaluare. În consecinţă, comparaţia propusă anterior trebuie efectuată pe o selecţie a corpusului de referinţă. Corpusul de evaluare a fost aliniat forţat, iar ştampilele de timp rezultate au fost utilizate (împreună cu ştampilele de timp ale transcrierii ipotetice) pentru a selecta părţile corespunzătoare din transcrierea de referinţă.

Rezultatele arată că metoda de adnotare automată produce corpus de calitate înaltă: 99% din cuvintele selectate sunt corecte. Acest fapt confirmă eficienţa celor două sisteme RAV: erorile făcute de acestea sunt identice numai într-o mică măsură, iar părţile transcrise identic sunt aproape (99%) corecte. În concluzie, utilizarea celor două sisteme RAV care diferă prin tipul modelului acustic (HMM-GMM vs. HMM-DNN) este de departe mult mai eficientă decât utilizarea de sisteme RAV ce diferă prin alte caracteristici (seturile de date folosite sau metoda de decodare).

B. Adnotarea seturilor de vorbire nou achiziţionate

Metoda de adnotare automată a fost aplicată pentru seturile de date nou achiziționate. Pentru primul set, aproape 50% din cantitatea totală de date din sursa #1 a fost adnotată. Pentru sursa #2 au putut fi adnotate doar 20% din date, în timp ce pentru sursa #3, aproape 31% din date au fost aliniate şi adnotate automat. Aceleaşi procente de adnotare s-au obţinut şi pentru cel de-al doilea set.

Recunoaşterea automată a vorbirii

Noile seturi de date de vorbire adnotată (49 ore, respectiv 280 de ore, după cum a fost prezentat anterior) au fost adăugate la seturile de antrenare deja existente. Ambele sisteme RAV au fost reantrenate şi evaluate

Adăugarea primului corpus adnotat la setul de antrenare, a fost avantajoasă numai pentru sistemul bazat pe HMM-GMM: pe ambele corpusuri de evaluare, noul sistem bazat pe HMM-GMM a obţinut îmbunătăţiri relative minore. Situaţia diferă însă în cazul sistemului bazat pe HMM-DNN, care a obţinut rezultate mai slabe decât sistemul iniţial. Adăugarea celui de-al doilea corpus adnotat s-a utilizat numai pentru reantrenarea sistemului ASR bazat pe HMM-DNN. Pe ambele corpusuri de evaluare au fost obţinute uşoare îmbunătăţiri.

În această etapă a fost utilizată o metodă de adnotare automată a corpusurilor de vorbire ce presupune utilizarea transcrierii de la două sisteme RAV complementare. Experimentele au arătat că sistemele RAV bazate pe HMM-GMM, respectiv HMM-DNN, fac semnificativ mai puţine erori comparativ cu situaţia când ambele sisteme folosesc modele acustice de tip HMM-GMM. În prima situaţie, eroarea la nivel de cuvânt (WER) este de aproximativ 1%, în timp ce rezultatele prezentate într-o lucrare anterioară indică o rată de eroare la nivel de cuvânt de aproximativ 10%.

2.4.5 Concluzii

Sistemele de RAV iniţiale au o precizie destul de bună, fiind antrenate cu o foarte mare cantitate de date. Acesta poate fi un motiv pentru care reantrenarea sistemelor adăugând noile corpusuri obţinute prin această abordare nesupervizată nu aduce îmbunătăţiri semnificative. Din cantitatea totală de date a fiecărui nou corpus achiziţionat, un procent de aproximativ 36% au fost adnotate corect. După reantrenarea modelelor acustice adăugând primul corpus achiziţionat, sistemul de RAV bazat pe HMM-GMM a obţinut o mică îmbunătăţire, în timp ce sistemul de RAV bazat pe HMM-DNN a pierdut puţin în ceea ce priveşte performanţa. Reantrenarea sistemului HMM-DNN adăugând ambele corpusuri nou achiziţionate a condus la o uşoară îmbunătăţire a performanţelor.

2.5 Activitatea 1.14 – Diseminare

Diseminarea rezultatelor proiectului a fost realizată în cadrul consorţiului în cele două workshop-uri organizate pe parcursul acestei etape (Bucureşti: 7 – 8 iunie 2018, şi Iaşi: 22 – 23 noiembrie 2018) şi în comunitatea ştiinţifică la *The 41st International Conference on Telecommunications and Signal Processing (TSP)*. Articolul publicat la această conferinţă, este în prezent indexat *IEEE Xplore* şi în curs de indexare *Thompson Reuters CPCI* (ISI), iar numele finanţatorului este menţionat în secţiunea *Acknowledgement*, conform indicaţiilor din contractul de finanţare.

- Alexandru-Lucian Georgescu, Horia Cucu, "Automatic annotation of speech corpora using complementary GMM and DNN acoustic models," în *The Proceedings of the 41st International Conference on Telecommunications and Signal Processing (TSP)*, 2018, Athens, Greece.

3 Structura ofertei de servicii de cercetare şi tehnologice

Serviciu	Detalii
Serviciu şi aplicaţie web de transcriere de documente ce conţin vorbire în limba română	https://transcriptions.speed.pub.ro
Serviciu şi aplicaţie web de identificare de cuvinte cheie în documente ce conţin vorbire în limba română	https://keywords.speed.pub.ro
Serviciu şi aplicaţie web de restaurare de diacritice în limba română	https://diacritics.speed.pub.ro
Proiectarea şi implementarea de aplicaţii personalizate de transcriere a vorbirii continue	La cerere
Proiectarea şi implementarea de aplicaţii personalizate de identificare de cuvinte şi termeni de interes	La cerere
Proiectarea şi implementarea de aplicaţii personalizate de sinteză de vorbire pornind de la text	La cerere
Proiectarea şi implementarea de sisteme de recunoaştere de pattern-uri folosind inteligenţă artificială	La cerere

Laboratorul de cercetare *Speech and Dialogue* (SpeeD) este prezent pe platforma ERRIS la adresa https://erris.gov.ro/SpeeD---UPB. Laboratorul de cercetare *Speech and Dialogue* (SpeeD) din cadrul Universităţii Politehnica din Bucureşti (UPB), reprezentantul UPB în proiectul TADARAV, oferă pe platforma ERRIS serviciile de cercetare şi tehnologice enumerate în tabelul de mai sus.

4 Locuri de muncă susţinute prin program

Echipa de cercetare UPB

Nr.	Nume	Calitatea	Poziţia	Normă
1	Horia CUCU	Conf. Univ.	Responsabil proiect component	Parţială
2	Corneliu BURILEANU	Prof. Univ.	Membru cercetător	Parţială
3	Dragoş BURILEANU	Prof. Univ.	Membru cercetător	Parţială
4	Alexandru-Lucian GEORGESCU	ACS	Membru cercetător	Parţială
5	Dan Theodor ONEAŢĂ	CS	**Membru cercetător nou**	**Întreagă**
6	Gheorghe POP	ACS	**Membru cercetător nou**	**Întreagă**
7	Cristian MANOLACHE	ACS	**Membru cercetător nou**	**Întreagă**

Cei trei noi membri cercetători au fost angajaţi începând cu data de 25.04.2018 cu funcţiile de cercetător ştiinţific (CS), respectiv asistenţi cercetare ştiinţifică (ACS), cu normă întreagă.

5 Valorificarea şi îmbunătăţirea competenţelor şi resurselor existente la nivelul consorțiului

La nivelul proiectului component TADARAV CEC-urile nu au fost valorificate.

PROIECTUL **"SINTERO: Tehnologii de realizare a interfețelor om-mașină pentru sinteza text-vorbire cu expresivitate"**

Mircea Giurgiu, Adriana Stan

Responsabil: Universitatea Tehnică Cluj-Napoca, coordonator Prof. Mircea Giurgiu

Rezumatul etapei

Acest document prezintă o sinteză a realizărilor de natură științifică și tehnică obținute in prima etapă de implementare a sub-proiectului SINTERO din cadrul proiectului PCCDI ReTeRom. Realizările se referă la:

- identificarea pattern-urilor prozodice și corelațiile între text și semnal vocal
- identificarea metodelor de clasificare automată a stilului de exprimare
- analiza metodelor de control și adaptare a expresivității în sistemele de sinteză
- implementarea modulului de control automat al prozodiei

Activitățile de cercetare desfășurate în prima etapă de implementare au condus la obținerea rezultatelor așteptate și ele sunt în concordanță cu obiectivele specifice ale etapei. Astfel, rezultatele raportate în acest document și descrise detaliat în cele 5 livrabile aferente perioadei de raportare, pregătesc cadrul etapei viitoare pentru implementarea componentelor de modelare a prozodiei și adaptare la noi vorbitori a vocilor sintetice. De asemenea, acest raport prezintă detalii referitoare la oferta de servicii de cercetare și tehnologice, activitățile de management și comunicare, modul de valorizare a resursei umane și dezvoltarea acesteia prin activități colaborative la nivelul consorțiului.

2. Activitățile etapei de raportare în contextul general al proiectului

În prima etapa a proiectului SINTERO (2018), etapă cu denumirea *„Metode de modelare și control a expresivității în sistemele de sinteză text-vorbire"*, s-a pornit de la resurse si module software deja existente la partenerii UTCN și ICIA și au fost desfășurate o serie de activități pentru: **a)** identificarea pattern-urilor prozodice și propunerea unei soluții pentru de modelare a prozodiei**, b)** identificarea metodelor de clasificare automata a stilului de exprimare și implementarea algoritmilor pentru reprezentarea vectorială a surselor de date text și audio, **c)** analiza a 3 metode de control și adaptare a expresivității vorbirii artificiale (concatenativ, statistic, neuronal), **d)** implementarea modulului de control automat al prozodiei pe baza unor noi corpusuri de date audio cu diferite stiluri de exprimare (de exemplu stil jurnalistic și stil narativ), cu controlul intonației frazei (3 pattern-uri: declarativ, exclamativ și interogativ), și cu demonstrarea online a rezultatelor, *https://speech.utcluj.ro/sintero/prosody_examples/*.

În etapele următoare ale proiectului SINTERO vom realiza *„Integrarea componentelor pentru modelare prozodie și adaptare la noi vorbitori a vocilor sintetice"* (2019) și în final *„Dezvoltarea unei noi tehnologii pentru sinteza text-vorbire cu expresivitate"* (2020).

3. Gradul de realizare a obiectivelor specifice pentru Etapa I-a

Ob. Pr4.1.15: *Identificarea pattern-urilor prozodice*

Grad realizare: Obiectiv realizat integral

Rezultate:

- 1 set de înregistrări audio pentru evaluare pattern-uri prozodice
- măsurări cantitative și calitative a pattern-urilor accent, întonație, pauze și ritm în fraze declarative, exclamative sau interogative
- evaluarea frecvenței fundamentale, a formanților si duratei vocalelor, diftongilor si triftongilor în funcție de contextul prozodic

- identificarea a 7 pattern-uri prozodice și concluzii privind modul de variație a prozodiei în funcție de gradul de expresivitate a textului.
- un livrabil (D1.15) cu titlul *„Identificarea pattern-urilor prozodice și evidențierea corealțiilor între txt și semnal vocal”.*

Ob. Pr4.1.16: *Identificarea metodelor de clasificare automată a expresivității (text/audio)*

Grad realizare: Obiectiv realizat integral

Rezultate:

- 3 metode candidat pentru reprezentarea vectorială a textelor
- 1 implementare și rezultate preliminare pentru clasificarea stilului din text
- lista cu parametrii acustici relevanți pentru clasificare stil vorbire
- 1 implementare si rezultate preliminare privind clasificarea stilului de vorbire (emotivitate) din datele audio
- 1 livrabil (D1.16) cu titlul *„Identificarea metodelor de clasificare automată a stilului de exprimare din surse de date text si audio”.*

Ob. Pr4.1.17: *Analiza metodelor de control și adaptare automată a expresivității*

Grad realizare: Obiectiv realizat integral

Rezultate:

- raport cu metodele de control a expresivității în sisteme concatenative
- raport cu metodele de control a expresivității în sisteme statistice HMM
- raport cu metodele de control a expresivității în sisteme DNN
- 1 livrabil (D1.17) cu titlul *„Analiza metodelor de control și adaptare automată a expresivității vorbitorilor în sistemele de sinteză text-vorbire”.*

Ob. Pr4.1.18: *Realizarea unui modul de control automat al prozodiei*

Grad realizare: Obiectiv realizat integral

Rezultate:

- 1 metodă și interfață funcțională pentru controlul manual al prozodiei prin modificarea liniară a duratei și a intonației în propoziție
- 3 noi voci sintetice pentru prozodie cu expresivitate neutră, stil jurnalistic de prezentator de știri, stil narativ de tip audio book
- implementarea metodei de adaptare CSMALPR pentru adaptarea vocii neutre la 2 stiluri de vorbire (jurnalistic, narativ)
- 1 demonstrator online pentru controlul automat al prozodiei[16]
- 1 livrabil (D1.18.) *„Implementarea modulului de control automat al prozodiei”.*

Ob. Pr4.1.19: *Diseminarea rezultatelor intermediare*

Grad realizare: Obiectiv realizat integral

Rezultate:

- realizarea și actualizarea web site-ului proiectului[17]
- 1 pagină cu demonstrator online pentru control și adaptare prozodie
- 1 articol la conferința CONSILR 2018.
- 1 livrabil referitor la activitățile de diseminare (D1.19).

4. Rezultatele etapei și descrierea lor științifică și tehnică

4.1. Identificarea pattern-urilor prozodice și corelațiile între text și semnal vocal

Rezultatele raportate în aceasta secțiune corespund obiectivului Pr4.1.15, iar ele sunt descrise in extenso in livrabilul D1.15 (www.racai.ro/p/reterom/rapoarte/1.15). Ca fundament pentru cercetările raportate în acest

[16] *http://speech.utcluj.ro/sintero/prosody_examples*

[17] http://speech.utcluj.ro/sintero/

livrabil sunt rezultatele anterioare obținute de partenerii CO-ICIA (procesarea limbajului natural) și P1-UTCN (analiza unităților acustice din semnalul vocal), care pun în evidență principalii factori de natură lingvistică prin care se manifestă modificările prozodice în forma de undă: accentul, intonația în vorbire, silabificarea, pauzele, ritmul vorbirii, respectiv elemente de morfologie și sintaxă în interacțiune. Pornind de aici s-au ramificat două direcții de cercetare: identificarea modului de manifestare a prozodiei în parametrii semnalului vocal, respectiv corelația parametrilor prozodici cu caracteristici extrase din text.

În primul rând sunt prezentate rezultatele experimentale privind variația parametrilor prozodici frecvență fundamentală pentru vocale, frecvența fundamentală în funcție de accent, frecvență fundamentală în funcție de intonația din propoziție, variația frecvenței formanților pentru diferiți vorbitori, respectiv rolul duratei și a pauzelor în modelarea pattern-urilor prozodice. Analiza s-a realizat pe un corpus de semnal vocal înregistrat în acest scop.

De exemplu, pentru unitățile acustice diftongi, pattern-urile prozodice indica faptul ca frecvențele fundamentale suferă variații atunci când diftongii (respectiv vocalele) sunt încadrați în cuvinte; F0 maxim scade atunci când avem grupuri de vocale încadrate împreună în cuvânt, iar energia acestor diftongi încadrați în cuvinte este sensibil mai mică decât cea a diftongilor, triftongilor izolați. Similar s-au obținut rezultate pentru diferite categorii de unități acustice. Un alt exemplu este pentru accent.

Una din concluziile importante ale studiului se referă la o creștere a frecventei fundamentale pentru silabele (sau vocalele) accentuate, față de cele neaccentuate în medie cu 5%..20% (în 90% din cazuri creșterea s-a plasat în intervalul 9%..12%). Conform studiilor realizate până acum s-a arătat ca în general silabele accentuate au tendința de a avea frecventa fundamentala, durata și amplitudine mai ridicata decât silabele neaccentuate. Însă, exista și cazuri în care numai unul sau doi din acești parametri este mai ridicat, precum și situații în care tendința silabelor accentuate este de a-si reduce fundamentala sau ceilalți parametrii. Merită făcută și observația ca au existat și câteva cazuri în care accentuarea unei silabe nu a adus nici un fel de diferențiere din punctul de vedere al valorii F0. Similar sunt prezentate rezultate pentru formanți, respectiv evaluarea duratei unităților acustice în funcție de accent.

În al doilea rând sunt prezentate rezultate privind analiza caracteristicilor de natura lingvistică ce afectează prozodia, în special la nivel de intonație de propoziție. Sunt identificate un set de 7 pattern-uri intonaționale la nivel de propoziție, dar și efectul prozodic al semnelor de punctuație.

Cercetările demonstrează faptul că pattern-urile prozodice manifestate la nivelul semnalului vocal au legătură directă și prezintă strânse corelații pe termen scurt sau pe termen lung cu atribute de morfologie și sintaxă aferente textului. Principalele atribute se referă la poziționare accent în cuvinte, silabificare, părți de vorbire, sintaxă, respectiv punctuație. Aceste rezultate, documentate prin grafice și tabele (www.racai.ro/p/reterom/rapoarte/1.15) prezintă fundamentul pentru dezvoltarea unor noi metode de sinteză expresivă a vorbirii prin intermediul unor module de analiză a expresivității textului (în componenta software de procesare de text), respectiv de modificare automată a prozodiei (în componenta software de sinteză de semnal).

4.2. Identificarea metodelor de clasificare automată a stilului de exprimare

Rezultatele raportate în această secțiune corespund obiectivului Pr.4.1.16, iar ele sunt descrise în extenso în livrabilul D1.16 (www.racai.ro/p/reterom/rapoarte/1.16). Acest livrabil prezintă atât rezultate de natură teoretică ce au în vedere identificarea unor soluții de clasificare automată a stilului de exprimare din surse de date text și audio, precum și implementarea modulelor software aferente. Identificarea si clasificarea stilului de exprimare din text este necesară în modulul de procesare a textului din cadrul unui sistem de sinteză text – vorbire cu scopul de a informa generatorul de semnal vocal despre expresivitatea pe care trebuie să o incorpoze la sinteză. Aceasta expresivitate este determinată de conținutul semantic al textului și de polaritatea (ca sentiment +/-) acestuia.

În primă etapă au fost identificate cîteva metode de reprezentare vectorială a textelor. Acestea se referă la reprezentări de tipul Bag of Words, VSM (Vector Space Models) și LSA (Latent Semantic Analysis). Din punct de vedere practic s-au implementat și apoi testat experimental fluxurile de procesări care realizează reprezentările amintite și prin care s-a verificat posibilitatea de clasificare a mai multor stiluri de vorbire similar identificarii automate a topicurilor din discursul de tip text. Rezultatele preliminare s-au obținut pe un corpus redus, dar avem in vedere utilizarea corpusurilor (belestristic, științific, jurnalistic, narativ) obținute de la Partenerul ICIA.

Similar metodelor de clasificare a textelor s-au identificat paramterii acustici care ar fi relevanți în clasificarea stilului de vorbire numai din date audio. Tonul din voce, aparte de mesajul lingvistic, este un bun indicator. Ca exemplu, în raportul extins ilustrăm modul de variație a 2 dintre acești parametri (frecvența fundamentală, respectiv parametrul LSF1) pentru 2 voci cu emotivități diferite. Acești parametri au un potențial înalt de discriminare între diferitele stiluri de vorbire.

Cele mai frecvente metode de clasificare aplicate pentru recunoașterea stilului de vorbire și a expresivitații (inclusiv pentru recunoașterea emoțiilor) sunt arborii de decizie, clasificatorii SVM sau rețelele neuronale. În aplicația prototip s-a utilizat un corpus cu 5 stiluri de expresivitate, corespunzând la 5 clase de emoții. În total s-au folosit 500 de fișiere audio pentru fiecare emoție, în total un set de 2500 de fișiere. Întreg setul a fost împărțit în două, un set pentru antrenare și unul pentru testare.

Identificarea parametrilor acustici relevanți pentru clasificarea expresivității vocale

Parametri	Utilizare
Parametri spectrali pe termen lung	*Media spectrului, spectral flatness measure, centroidul spectral*
Parametri spectrali pe termen scurt	*MFCC (Mel frequency Cepstral Coefficients), LSF (Line Spectrum Frequency), LPC-PLP (Linear Predictive Coefficients – Perceptual Linear Prediction)*
Pitch	*Media, deviația standard, skewness, kurtosis, maximum, minimum, quartiles, diferențe între quartile, coeficienții de regresie liniară și quadratică*
Rata vorbirii	*Media si deviația standrad pentru durata silabelor, raportul dintre durata segmentelor sonore și nesonore*
Parametri in timp	*Intensitatea, RMS, numarul de treceri prin zero, TEO (Teager Energy Operator)*
Parametri tonali	*Coeficientii CHROMA, CENS*
Calitatea vocii	*HNR (harmonic to Noise Ratio), Jitter, Schimmer*

Parametrii acustici au fost extrași cu aplicația GlottHMM și printr-o procedură de selecție a parametrilor bazată pe "information gain", s-au generat vectorii specifici fiecărui stil. Rezultate se prezintă pentru setul de parametri (F0. NAQ, LSF1, LSF2, LSF3, LSF4, HNR1, HNR2, HNR3, HNR4, HNR5) pentru care s-au inclus în vector media și deviația standard.

Prezentam doar rezultatele globale de clasificare obținute prin 3 metode standard,

J48-arbori de decizie - *83,67%;* **Logistic Model Tree** - *95,40%;* **MLP** - *97,95%*

Pe baza acestei metodologii, în următoarea etapă vom considera colectarea unui set de date audio si text relevante pentru aplicația finală, iar pe baza acestora vom desfășura experimente elaborate pentru testare în condiții mult mai complexe (volum date, vectori mari).

4.3. Analiza metodelor de control și adaptare a expresivității în sistemele de sinteză text-vorbire

Rezultatele raportate în această secțiune corespund obiectivului Pr.4.1.17. Problema variabilității și expresivității vocilor sintetice este de mare actualitate (vezi Livrabil D1.17

www.racai.ro/p/reterom/rapoarte/1.17), în special prin prisma faptului că expresivitatea și/sau prozodia nu pot fi evaluate în mod obiectiv printr-un set de parametri și de cele mai multe ori depind de starea emoțională a persoanei care o evaluează, precum și de fondul cultural, etnic sau educațional. Diferitele tipuri de sisteme de sinteză: concatenative, parametric-statistice sau cele ce modelează direct forma de undă, permit un control al expresivității și prozodiei specific, în funcție de arhitectura sistemului.

Problema sistemelor concatenative este faptul că informația audio nu este parametrizată sub nicio formă, astfel că pentru controlul expresivității este necesară manipularea formei de undă. La modul cel mai simplu, controlul expresivității este făcut prin selectarea segmentelor audio de concatenat pe baza unei traiectorii prozodice predefinite sau estimate din text, prin intermediul unei funcții de selecție a unităților. Totuși, înregistrarea aceluiași vorbitor în condiții de expresivitate sau emotivitate variabile este greu de realizat. Modificarea formei de undă se face prin metoda PSOLA (Pitch Synchronous Overlap and Add), doar că acest tip de modificare introduce artefacte ne-naturale în vocea sintetică. Deși au fost dezvoltate și implementate multiple metode de control a emoțiilor și a expresivității în sinteza concatenativă, faptul că această tehnologie se bazează pe forma de undă în sine, cu anumite modificări parametrice ale rezultatului vocal, face ca orice modificare adusă semnalului să introducă erori de sinteză nedorite.

În sistemele de sinteză bazate pe modele Markov fonemele sunt modelate printr-un anumit set de parametri (de exemplu coeficienți Mel-cepstrali, coeficienți de aperiodicitate, frecvența fundamentală, F0, și durata), iar pentru controlul expresivității se pot astfel adapta în mod independent modelele acestor parametri. Problema este că naturaleța vocii sintetizate este condiționată de utilizarea unor înregistrări audio cu prozodie cât mai variată, pentru ca modelele statistice să poată utiliza un număr cât mai mare de exemple fonetice pentru același context. Principalele modalități de adaptare:

- adaptarea prozodică a informției textuale prin etichete prozodiec ce tip ToBI, utilizarea varianței globale a parametrilor, utilizarea unor adnotări la nivel suprasegmental.
- controlul modelelor acustice prin date care conțin starea de emoție sau expresivitate specifică vorbitorului, utilizarea caracteristicilor la nivel suprasegmental si aplicarea Multipe Regression HMM, adnotări la nivel articulator.
- adaptarea modelelor acustice pornind de la un set foarte mare de date de la foarte multi vorbitori, crearea unei voci eigen, iar apoi adaptarea modelelor către vocea sintetica printr-un set redus de date și aplicarea unor metode de factorizare a modelelor Markov.

În sistemele de sinteză bazate pe rețele neuronale abordările pentru adaptarea expresivității includ extinderea setului de caracteristici de intrare cu un set de caracteristici de prozodie sau de stil de vorbire.

Problema majoră a acestor sisteme este necesitatea existenței unui corpus de voce de dimensiuni imense (sute de ore de vorbire) pe baza căruia să se realizeze antrenarea rețelei. Învățarea prin transfer poate fi o soluție pentru a compensa indisponibilitatea unui astfel de corpus. Chiar și cu resurse de date disponibile, modul de adnotare a caracteristicilor de expresivitate (de exemplu emoții) este esențial. Adesea se introduc codificări suplimentare cu creșterea imensionalității datelor de intrare.

Un exemplu de sistem comercial (Tacotron) este cel de la Google. Foarte recent acest sistem a fost extins cu tehnologia Global Style Tokens pentru a învăța automat expresivitatea latentă în semnalul de intrare. Pe de altă parte, sistemul EMPHASIS de la Baidu modelează dependențele lingvistic – acustic printr-o rețea de regresie. Ambele sisteme comerciale generează voce sintetică de calitate, dar generarea automată a etichetelor de expresivitate și prozodice din text, precum și transferul stilului de vorbire, rămân probleme deschise.

4.4. Implementarea modulului de control automat al prozodiei

Rezultatele raportate în această secțiune corespund obiectivului Pr.4.1.18. Modulul software pentru controlul prozodiei (vezi Livrabil D1.18 (www.racai.ro/p/reterom/rapoarte/1.18) permite două moduri de operare. În modul manual, prin intermediul unei interfețe grafice intuitive accesibile chiar si pentru utilizatori non-experți, este posibilă modificarea de către utilizator a conturului frecvenței fundamentale (F0) și a duratei segmentelor sonore de vorbire, dar cu alinierea automată a datelor audio. Fișierul de configurare păstreaza valorile medii ale parametrilor.

În modul automat de funcționare a modulului, controlul prozodiei se bazează pe 3 seturi de date audio înregistrate în studio, cu stiluri de exprimare diferite (neutru -1h și 43 de minute, jurnalistic – 48 de minute, narativ – 11 ore din audio book).

Experimentele noastre (vezi http://speech.utcluj.ro/sintero/prosody_examples/) descrise în raportul extins, arată că valorile medii pentru F0 sunt apropiate (200-220Hz), dar au deviații standard destul de diferite, în relație cu stilul de vorbire. De asemenea, am observat duratele diferite ale fonemelor în cele 3 stiluri de vorbire.

De exemplu, în stilul jurnalistic durata fonemelor este mai mică. Acest corpus, deși are o durată mai mică și este aliniat doar la nivel de propoziție, este suficient de bogat în informație prozodică, datorită expresivității verbale a prezentatoarei de știri. În primă fază au fost implementate 3 voci sintetice pentru aceste 3 corpusuri (modelare HMM, 5 stări, vocoder WORLD) pentru a putea compara ulterior vocea sintetică, adaptată la aceste stiluri, cu vocea sintetică originală. S-a implementat metoda de adaptare automată a prozodiei bazată pe algoritmul Constrained Structural Maximum Aposteriori Linear Regression - CSMAPLR). Noul mod de variație a F0 și duratei pentru vocea sintetică adaptată a fost prezentată grafic.

Versiunile următoare ale acestui sistem vor avea în vedere detecția automată a stilului de exprimare pornind de la textul de intrare, precum și modalități de adaptare a stilului folosind un set redus de date audio.

5. Oferta de servicii de cercetare, locuri de muncă susținute și valorificarea resurselor

Tabel 4.8. Sinteză privind oferta de servicii, locuri de muncă și valorificarea resurselor în UTCN

Oferta de servicii în UTCN	● oferta unei tehnologii de sinteză text-vorbire în limba română ● servicii de adnotare automată a resurselor de date audio ● servicii de înregistrare audio de înaltă fidelitate ● servicii software pentru dezvoltarea modelelor bazate pe învățare automată. *ERRIS: https://erris.gov.ro/speech.utcluj.ro*
Locuri de muncă susținute în UTCN	1 x CS I, 1 x CS II, 1 x CS III, 1 x Tehnician 2 x ACS pentru noii angajați
Resursa umană nou angajată în UTCN	În iunie 2018 au fost demarate procedurile în UTCN pentru scoaterea la concurs a 2 noi posturi de ACS. Anunțul a fost publicat în 12.09.2018, iar concursul a fost planificat pentru 26.09.2018 (cf anunt România Libera, Monitorul Oficial, site ANCS/Euraxes, site UTCN). Nu s-a prezentat nici un doctorand,asa cum s-a solicitat în anunț. Ulterior s-au făcut demersuri, cf Legii 319/2003 pentru angajarea pe aceste 2 posturi a 2 masteranzi, doar ca UTCN dorește ca ocuparea postului de ACS să fie făcută de un doctorand. În aceste condiții se caută candidați cu profil de doctorand.
Valorificare resurse în parteneriat	● UTCN a preluat de la ICIA resurse de date text (5 corpusuri) pentru clasificarea stilurilor de exprimare ● UTCN a furnizat pentru ICIA si UAIC corpusurile de date audio disponibile și adnotările acestora ● UAIC a furnizat pentru UTCN o metodă de clasificare a textului dezvoltată în limbajul R.
Cecuri	● UTCN a oferit un cec pentru înregistrare corpusuri audio, dar încă nu a fost folosit de parteneri.

6. Management si comunicare

Activitățile de management au fost orientate în special către managementul proiectului complex în vederea integrării diferitelor grupuri de cercetare și a resurselor tehnice ale acestora. Este de notat faptul ca s-a asigurat o bună comunicare și coordonare și pentru realizarea planului de achiziții global, respectiv pentru documentația de raportare etapă. Din punct de vedere administrativ s-au primit 4 tranșe de avans cu o regularitate adecvată. Nu toate resursele financiare alocate UTCN au fost folosite integral.

7. Diseminarea rezultatelor

O preocupare a Consorțiului în etapa de raportare a fost implementarea și îndeplinirea cu succes a obiectivelor stabilite în strategia de diseminare a rezultatelor elaborată în cadrul propunerii de proiect. Astfel, adecvat acestei etape inițiale s-a acționat pe următoarele direcții: a) crearea paginii web a proiectului SINTERO *(http://speech.utcluj.ro/sintero/),* b) publicarea conform planului a unui articol la conferința CONSILR 2018 (vezi mai jos), c) crearea unei pagini web dedicate pentru demonstrarea online a modulului de control a prozodiei (*https://speech.utcluj.ro/sintero/prosody_examples/*).

[1] A. Stan, M.Giurgiu , „A comparison between traditional machine learning approaches and deep neural networks for text processing in Romanian", In Proc. of The The 13th International Conference on Linguistic Resources and Tools for Processing Romanian Language, Iasi, 22-23 November 2018.

8. Concluzii

Activitățile de cercetare desfășurate în etapa I-a de implementare a proiectului (2018) au condus la obținerea rezultatelor asteptate și ele sunt în concordanță cu obiectivele specifice ale etapei. Astfel, rezultatele raportate în acest document și descrise detaliat în cele 5 livrabile aferente perioadei de raportare (vezi Secțiunea 8 a acestui raport), pregătesc cadrul pentru etapa a doua.

9. Referințe la livrabilele aferente etapei 2018 (Anexe la raport)

[1] Livrabil D1.15:	*„Identificarea pattern-urilor prozodice si evidențierea corelațiilor între text și semnal vocal",* Mai 2018.
[2] Livrabil D1.16:	*„Identificarea metodelor de clasificare automată a stilului de exprimare din surse de date text și audio",* Mai 2018.
[3]Livrabil D1.17	*„Analiza metodelor de control și adaptare automată a expresivității vorbitorilor în sistemele de sinteză text-vorbire",* Noiembrie 2018.
[4] Livrabil D1.18:	*„Implementarea modulului de control automat al prozodiei",* Noiembrie 2018.
[5] Livrabil D1.19:	*„Diseminare",* Noiembrie 2018.