RAPORT ŞTIINŢIFIC proiect complex ReTeRom, etapa I - iunie 2018
Proiectul 1: COBILIRO
Activitatea A1.1
Denumire Etapă: **Studiu *state-of-the-art* asupra realizării corpusurilor bimodale**

Autori: UAIC, ICIA, UTCN, UPB

REZUMATUL ETAPEI

Prima etapă a proiectului 1 COBILIRO pregăteşte cu date şi modele realizarea prototipului, ce va fi realizat pe parcursul etapei a II-a. În esenţă, activităţile prevăzute în acest an au ca principal obiectiv stabilirea convenţiilor de adnotare lingvistică şi adnotarea unui eşantion de texte, ce vor fi apoi luate drept model pentru un proces automat, propunerea unui inventar de date geografice şi specificaţii pentru realizarea interfeţelor utilizator. Alte activităţi, urmăresc realizarea comunicării în cadrul consorţiului, diseminarea rezultatelor, coordonare şi management şi raportare.

Rezultate Etapa: Raport asupra progreselor privind realizarea corpusurilor modale

În această etapă au fost identificate, analizate şi 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, procesarea şi adnotarea pe diferite niveluri lingvistice a corpusului bimodal generat în cadrul Proiectului 1 COBILIRO.

1. INTRODUCERE

1.1 Mediu, modalitate, corpus

În semiotică, prin modalitate[1] se înţelege modul prin care informaţiile trebuie codificate pentru a fi prezentate oamenilor, adică tipul de semn şi statutul realităţii atribuite sau revendicate de un semn, fie el text sau alt gen. Modalităţile senzoriale vor fi: vizuale, auditive, tactile, olfactive, gustative, kinestezice etc.

O listă a tipurilor de semne ar include: scrierea, simbolul, indexul, imaginea, harta, graficul, diagrama etc. Unele combinaţii de semne pot fi multi-modale, adică diferite tipuri de semne pot apărea grupate împreună pentru un efect amplificat. Trebuie făcută distincţia dintre mediu şi modalitate. Modalitatea se referă la un anumit tip de informaţie şi / sau la formatul de reprezentare în care sunt stocate informaţiile. Mediul este mijlocul prin care această informaţie este transmisă simţurilor interpretului uman.

Limbajul natural este modalitatea principală, având proprietăţi invariabile în mediul auditiv - ca limbă vorbită, în mediul vizual - ca limbă scrisă, în mediul tactil - ca Braille şi în mediul cinetic - ca limbaj al semnelor. Din acest punct de vedere, termenul de bi-modal (care înseamnă utilizarea combinată a două modalităţi) are un înlocuitor preferat, care este

[1] https://en.wikipedia.org/wiki/Modality_(semiotics)

bi-medial (care înseamnă, utilizarea combinată a două medii). Cele două medii de comunicare ale limbajului, în cazul proiectului nostru, sunt textual şi auditiv.

În contextul interacţiunii om-calculator şi în interesul proiectului nostru vom considera că ***modalitatea*** este clasificarea unui canal, independent de tipul de intrare / ieşire senzorială între om şi calculator. Un sistem este proiectat unimodal dacă are o singură modalitate implementată şi multimodal dacă are mai multe (Karray *et al.*, 2015). În cazul în care sunt disponibile mai multe modalităţi pentru o sarcină, se spune că sistemul are modalităţi redundante. Modalităţi multiple pot fi utilizate în combinaţie pentru a oferi metode complementare care, deşi redundante, pot transmite informaţii mai eficient (Palanque *et al.*, 2001, p. 43).

Modalităţile pot fi implementate în ambele direcţii: dintre om spre calculator şi invers. În aceste transferuri informaţionale se utilizează o gamă largă de tehnologii: (a) modalităţi comune (vizuale - grafică computerizată, de obicei, printr-un ecran; auditive - ieşiri audio; tactile - vibraţii sau alte mişcări), (b) modalităţi sofisticate (gustative; olfactive; termoreceptive; nonperceptive; de echilibru). Modalităţile vizuale şi auditive sunt cel mai frecvent utilizate deoarece sunt capabile să transmită informaţii cu o viteză mai mare faţă de altele: de la 250 la 300 cuvinte pe minut (WPM) (Ziefle, 1998) şi, respectiv, de la 150 la 160 cuvinte pe minut (Williams, 1998). Deşi implementarea în direcţia de modalitate calculator => om este rară, tactilul poate atinge o medie de 125 WPM prin utilizarea unui afişaj Braille actualizabil. Alte forme comune ale tactilului sunt vibraţiile de smartphone şi controlerul jocului.

În cazul comunicării om => calculator, calculatoarele pot fi echipate cu diferite tipuri de dispozitive de intrare şi senzori pentru a le permite să primească informaţii de la oameni. Dispozitivele de intrare comune sunt adesea interschimbabile dacă au o metodă de comunicare standardizată cu computerul şi permit ajustări practice utilizatorului. Anumite modalităţi pot oferi o interacţiune mai bogată în funcţie de context, iar opţiunile de implementare permit sisteme mai robuste (Bainbridge, 2004: 483). Vorbim de modalităţi simple (tastatură, dispozitiv de indicare, ecran tactil) şi modalităţi complexe (recunoaşterea imaginilor, a vorbirii, interfeţe haptice, orientare etc.).

Odată cu creşterea popularităţii smartphone-urilor, modalităţile complexe devin tot mai pe plac publicului larg. Recunoaşterea vorbirii, după introducerea lui Siri (Epstein, 2015), a reprezentat un punct de vânzare important al produselor Apple. Această tehnologie oferă utilizatorilor un mod alternativ de a comunica cu calculatoarele atunci când tastarea este mai puţin dezirabilă. Cu toate acestea, într-un mediu cu zgomot puternic, modalitatea auditivă nu este destul de eficientă. Acest lucru justifică faptul că anumite puncte forte ale modalităţilor diferă în funcţie de situaţie (Kurkovsky, 2009: 210-211). Alte modalităţi complexe, cum ar fi cele legate de vizual (de exemplu, Kinect Microsoft, sau alte tehnologii similare) pot face ca sarcinile sofisticate să fie mai uşor de comunicat unui calculator, mai ales sub forma unei mişcări tridimensionale (Kurosu, 2013: 366).

Desigur, un sistem cu multiple modalităţi oferă mai multe posibilităţi utilizatorilor şi robusteţe sistemului. De asemenea, permite o mai mare accesibilitate utilizatorilor. De reţinut rămâne faptul că modalităţile multiple pot fi folosite ca rezervă atunci când anumite forme de comunicare nu sunt posibile. Acest lucru este valabil mai ales în cazul

modalităţilor redundante în care două sau mai multe modalităţi sunt utilizate pentru a comunica aceleaşi informaţii.

Există şase tipuri de cooperare între modalităţi şi ele ajută la definirea modului în care o combinaţie sau fuziune a modalităţilor cooperează pentru a transmite mai eficient informaţiile (Grifoni, 2009: 37): (1) echivalenţa (informaţiile sunt prezentate în mai multe moduri şi pot fi interpretate ca aceleaşi informaţii), (2) specializarea (când un anumit tip de informaţii este prelucrat întotdeauna prin aceeaşi modalitate), (3) redundanţa (mai multe modalităţi procesează aceleaşi informaţii), (4) complementaritatea (multiplele modalităţi iau informaţii separate şi le combină), (5) transferul (o modalitate consumă o altă modalitate pentru a produce noi informaţii), (6) concurenţa (mai multe modalităţi iau informaţii separate care nu sunt fuzionate).

În accepţiunea proiectului ReTeRom, prin corpus bimodal[2] vom înţelege o colecţie de înregistrări orale însoţite de transcrierile lor şi de metadatele corespunzătoare. Un corpus bimodal este găzduit pe o platformă specializată, împreună cu serviciile şi aplicaţiile web de acces, dezvoltare şi întreţinere ale lui, unde sunt specificaţi algoritmii de utilizare ai corpusului şi, în unele cazuri, de unde pot fi descărcate exemple de aplicaţii care utilizează corpusul.

Un corpus bimodal este un caz particular de corpus multimodal, care, la rândul lui, reprezintă un tip particular de corpus[3]. Corpusurile pot conţine texte scrise, înregistrări orale sau ambele modalităţi de redare a unei limbi naturale. În proiectul de faţă ne interesează ultimul caz.

Se remarcă o evoluţie a conceptului de corpus: prima generaţie de corpusuri conţinea doar transcrieri (ex. *British National Corpus*), a doua generaţie conţinea şi textul oral şi pe cel scris (ex. *The Michigan Corpus of Academic Spoken English*). A treia generaţie conţine în plus alinierea dintre componenta scrisă şi cea orală. Există şi o ultimă etapă de evoluţie, caracterizată prin includerea dimensiunii video în corpusuri (Rasso and Mello, 2014:29). Obiectul interesului nostru este a treia generaţie de corpusuri.

Uneori[4] se face distincţia dintre corpus de vorbire (engl. *speech corpus*) şi corpus oral (engl. *spoken corpus*), deşi, adesea, cel puţin în engleză, termenii sunt utilizaţi neglijent, fără o distincţie clară între ei. În continuarea acestei secţiuni vom arăta cum pot fi departajate aceste două categorii de corpusuri, pentru ca mai departe să ne referim la amândouă prin termenul de corpus oral.

Prin **corpus de vorbire**, în general, se înţelege o bază de date de fişiere audio şi de transcrieri textuale ale acestora, într-un format care poate fi folosit pentru a crea modele acustice ce devin motorul cercetărilor asupra recunoaşterii vorbirii. Un exemplu sunt transcrierile Switchboard revizuite la Institutul pentru Procesarea Semnalelor şi Informaţiilor (ISIP) începând cu anul 1997[5] (Godfrey and Hollman, 1997). Fişierele audio şi transcrierile lor pot fi aliniate (la nivel de fonem, silabă şi cuvânt). În cadrul sistemelor de recunoaştere vocală, modelele prozodice sunt folosite în principal pentru predicţia evenimentelor

[2] V. şi http://www.aclweb.org/anthology/R09-1044

[3] Uneori este utilizat termenul de *corpus lingvistic*, evident - un pleonasm, pentru că prin corpus se înţelege oricum o colecţie de date despre limbă, în format scris sau audio.

[4] http://www.ilc.cnr.it/EAGLES96/spokentx/spokentx.html

[5] https://catalog.ldc.upenn.edu/ldc97s62

prozodice (accente sintactice şi semantice) asociate unui text. Implementarea prozodiei în procesarea vocală a permis obţinerea de semnal vocal sintetizat cu conţinut retoric/pragmatic. Corpusul de vorbire interesează cercetările în care accentul se pune pe semnalul sonor, pe proprietăţile sale acustice şi proprietăţile articulatorii ale tractului vocal. Reprezentarea simbolică în acest caz se face utilizând alfabetul fonetic.

În schimb, **corpusul oral** prezintă interes, cu precădere, pentru cercetările în care se studiază modul de utilizare a limbii şi caracteristicile diverselor niveluri lingvistice: lexical, morfologic, sintactic, semantic, analiza discursului, a conversaţiei, studii de pragmatica comunicării, sociolingvistică, dialectologie etc., dar şi în tehnologia vorbirii, atât pentru sisteme de înţelegere cât şi de generare a vorbirii. Transcrierea marchează, într-un sistem ortografic îmbogăţit, diverse fenomene auditive care însoţesc pronunţia: bâlbâieli, pauze, tuse, râs etc., uneori dublat şi de un alfabet fonetic. Şi aici înregistrarea verbală este aliniată la interpretarea sa textuală. Ideal ar fi atât cercetătorii lingvişti cât şi cei din tehnologia limbajului natural să folosească acelaşi set de date, cu condiţia ca acesta să îndeplinească cerinţele ambelor comunităţi în ceea ce priveşte colectarea, transcrierea, codificarea şi adnotarea datelor.

1.2 Tipuri de corpusuri orale

Aşa cum am precizat mai sus, ne referim aici la baze de date de înregistrări vocale dublate de transcrierile lor textuale, utile în crearea de *modele acustice*, care, la rândul lor sunt folosite în aplicaţii de recunoaştere a vorbirii (*speech recognition engines* sau *speech-to-text systems*, S2T). Pe de o parte, corpusul oral înseamnă un corpus de conversaţie informală, improvizată, fără implicare mediatică (Chafe, 1995), iar pe de altă parte, el este folosit pentru a înţelege limba a cărei formă de manifestare predominantă este cea orală.

Corpusurile orale sunt clasificate în două tipuri:

1) **voce-în-citire** (*read speech*), care include:

- lecturi din cărţi;
- ştiri;
- liste de cuvinte;
- secvenţe de numere.

2) **vorbire spontană** (*spontaneous speech*), care include[6]:

- dialoguri - între două sau mai multe persoane;
- narative - o persoană redă o întâmplare sau spune o poveste;
- trasee pe hartă - o persoană explică alteia un traseu pe o hartă, persoanele nu se văd, ambele au în faţă aceeaşi hartă, dar doar harta celei care explică are marcat un traseu, cealaltă urmând să refacă traseul pe harta proprie;
- stabilirea unei întâlniri - două persoane încearcă să fixeze parametrii unei întâlniri, în funcţie de restricţiile individuale din calendare;
- simulări "Vrăjitorul din Oz" - simularea unei situaţii din viaţa reală, ca de exemplu, dialogul pentru rezervarea unui zbor între un client şi un funcţionar al unei agenţii de turism;

[6] A se vedea, ca exemplu, corpusul SRI:
https://web.stanford.edu/dept/linguistics/corpora/material/X_Speech_Corpora.pdf

Limba engleză, este cel mai bine reprezentată de corpusuri orale, dar pe web pot fi găsite şi corpusuri ale altor limbi: poloneză, arabă, bulgară, cantoneză, cehă, farsi, franceză, germană, hindi, japoneză, coreeană, mandarină, portugheză, rusă, spaniolă, tamil, vietnameză etc.

Avantajele utilizării unui corpus oral sunt:

1. economie de timp la antrenarea sistemelor, nu este necesară colectarea spontană, înregistrările sunt pre-procesate, pot fi utilizate în mod repetat;
2. corpusurile pot conţine cantităţi mari de date, necesare în procesele de învăţare;
3. sunt interogabile;
4. pot fi create sub-corpusuri pe bază de selecţii;
5. exemplele de limbaj din corpusurile de vorbire spontană sună natural, pentru că în colectarea datelor se pune accent pe spontaneitate.

Dificultăţi/provocări în utilizarea unui corpus oral:

1. uneori, calitatea înregistrării este mai scăzută decât într-un laborator fonetic;
2. sunt greu de procurat, procesul de introducere de metadate şi de prelucrare este laborios, consumator de timp;
3. dacă se lucrează cu voluntari, mai ales într-un regim *crowdsourcing*, datele trebuie verificate, validate, curăţate, pentru că, uneori, voluntarii sunt neglijenţi ori chiar rău intenţionaţi;
4. înregistrările trebuie asigurate prin contracte de drepturi de autor şi de permisiune de utilizare (*data privacy*) cu vorbitorii.

1.3 Transcrierile

Transcrierile sunt redările textuale ale înregistrărilor vocale. În realizarea corpusurilor, transcrierile pot urma sau precede înregistrările vocale, dar ele sunt componente obligatorii ale corpusurilor bimodale. 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, la rândul lor clasificate în semne grafice care au drept corespondente sunete „primare” şi semne grafice care au drept corespondente sunete marcate de unul sau mai multe fenomene fonetice (Bejinariu *et al.*, 2007).

2. ACHIZIŢIA CORPUSURILOR BIMODALE

Cercetările dedicate recunoaşterii automate a vorbirii au cunoscut o creştere spectaculoasă începând cu anii ‘60 (Halle şi Stevens 1962; Denes & Mathews 1960; Denes, 1960), deşi utilizarea de corpusuri orale ca depozitare a limbii ce ar trebui interpretată sau generată de maşină este de dată mult mai recentă. În vederea achiziţionării de astfel de corpusuri au început să fie create interfeţe specializate. Am considerat util să prezentăm în această secţiune caracteristicile unei aplicaţii dedicate achiziţiei corpusurilor bimodale, care

concentrează principalele funcţionalităţi necesare construcţiei de corpusuri aliniate voce-text de tip voce-în-citire, independente de limbă.

Google for Android (Hughes *et al*., 2010)[7]

Arhitectura sistemului este una *client-server*, cu clienţi rulând o aplicaţie ecran sub Google Android, serverul rulând Java şi tratând cereri HTTP simultane de la clienţi.

Aplicaţia server pune la dispoziţie clienţilor rânduri de text, care sunt scurte segmente de text (propoziţii, cuvinte). Aceste prompt-uri sunt generate la iniţializarea aplicaţiei din cereri de căutare Google Search, filtrate pentru a elimina cuvinte pornografice şi cuvinte tipărite greşit; limba cererii se detectează automat, inclusiv utilizând informaţii de localizare; la primirea unei cereri de date (prompt) de la un client, serverul selectează la întâmplare din această listă de texte un rând şi îl trimite clientului (fără protecţia de a nu trimite duplicate); el primeşte apoi înregistrarea sonoră şi metadatele aferente ei via HTTP POST request; tripletul text-metadate-înregistrare este apoi stocat într-o bază de date şi clientul este informat că înregistrarea poate fi ştearsă din cardul SD; înregistrărilor sonore li se adaugă automat o margine scurtă de sunet de fundal la început şi la sfârşit.

În *aplicaţia client*, la deschiderea unei sesiuni de lucru, clientului i se deschide o interfaţă în care i se cere mai întâi să completeze metadatele următoare:
- mediul acustic (în casă, afară, în maşină, nivelul de zgomot);
- genul vorbitorului;
- vârsta (pe decade) a vorbitorului;
- accentul vorbitorului;
- un nume utilizator, pentru identificarea sesiunii.

În completare la acestea, aplicaţia mai include automat:
- data şi ora/minutul curent;
- versiunea hard a telefonului şi versiunea de SO Android;
- numărul IMEI al telefonului (un ID unic de identificare a dispozitivului mobil);
- locaţia geografică, determinată de GPS-ul mobilului (atenţie, o parte din aceste informaţii pot ridica probleme în contextul curent al legii intimităţii informatice!).

Urmează o serie de interacţiuni, în care utilizatorul primeşte în ecran textul şi trebuie să-l pronunţe. Vocea utilizatorului este stocată şi trimisă pe server împreună cu metadatele completate de client. Pentru a permite utilizarea în locuri fără conexiune la Internet, clientul necesită doar conexiuni intermitente la server pentru a descărca şi încărca înregistrările. Sesiunea este terminată de utilizator la dorinţă sau după un număr fixat de intrări completate de utilizator.

Sistemul a fost utilizat până acum pentru a colecta peste 3000 de ore de transcriere audio în 17 limbi, inclusiv limba rusă.

3. O TRECERE ÎN REVISTĂ A UNOR CORPUSURI BIMODALE

Oferim în această secţiune o seamă de informaţii relative la corpusuri bimodale existente în lume. Informaţiile sunt grupate în 10 categorii, care au fost identificate în majoritatea descrierilor.

[7] https://static.googleusercontent.com/media/research.google.com/ro//pubs/archive/36801.pdf

Santa Barbara Corpus of Spoken American English[8] (Du Bois *et al.*, 2000-2005)

1. **limba**: engleză;
2. **tip corpus**: în principal vorbire spontană dar şi voce-în-citire, predomină conversaţia faţă în faţă, dar conţine şi convorbiri telefonice, discuţii în timpul jocului de cărţi, al gătitului, discuţii la serviciu, prelegeri la clasă, predici, spunere de poveşti, întâlniri profesionale, prezentări din timpul ghidajelor turistice, texte citite etc.;
3. **dimensiune**: înregistrări totalizând 249.000 cuvinte;
4. **vorbitori**: oameni din diverse regiuni geografice de pe întreg teritoriul SUA, de vârste diferite, ocupaţii diferite, de sexe diferite, etnii şi categorii sociale diferite;
5. **informaţii private relative la vorbitori**: numele participanţilor la conversaţii şi alte date personale sunt anonimizate;
6. **metadate**: pentru fiecare vorbitor, sunt specificate: ID nume, sex, vârstă, localitatea de origine, statul de origine, statul în care trăieşte la momentul înregistrării, educaţie, ani de educaţie, ocupaţie, etnie;
7. **conţinut şi mod de realizare**: înregistrare sonoră, transcriere, aliniere voce-timp-text (engl. *timestamps*) la nivelul unităţilor intonaţionale;
8. **formatul înregistrărilor**: MP3 şi WAV;
9. **suport pentru realizare**: -
10. **distribuţie şi copyright**: cu licenţă CC-BY-ND.

The Buckeye Speech Corpus[9] (Pitt *et al.*, 2007)

1. **limba**: engleză
2. **tip corpus**: vorbire spontană în dialog, temele conversaţiilor sunt cotidiene (politică, sport, transport, şcoală);
3. **dimensiune**: 40 de ore şi 300.000 de cuvinte;
4. **vorbitori**: 40, intervievaţi fiecare câte o oră, provenind din Columbus, Ohio (născuţi aici sau care au ajuns aici la o vârstă mai mică de 10 ani, ceea ce face din acest corpus unul dialectal); stratificare pe vârstă şi sex;
5. **informaţii private relative la vorbitori**: -
6. **metadate**: -
7. **conţinut şi mod de realizare**: înregistrările sunt transcrise (folosind Soundscriber); aliniere automată (folosind Entropics Aligner) la nivel de cuvânt; ulterior, s-a corectat manual această aliniere;
8. **formatul înregistrărilor**: -
9. **suport pentru realizare**: -
10. **distribuţie şi copyright**: accesibil gratuit pentru cercetare (pe bază de cont); se poate descărca prin semnarea unei licenţe.

The Spoken Turkish Corpus[10] (Ruhi, 2011)

1. **limba**: turcă;
2. **tip corpus**: multimodal (conţine şi înregistrări video, nu doar audio); conversaţii faţă în faţă sau interacţiune mediată (la telefon);

[8] http://www.linguistics.ucsb.edu/research/santa-barbara-corpus
[9] https://buckeyecorpus.osu.edu/
[10] https://std.metu.edu.tr/en/

3. **dimensiune**: 10 milioane de cuvinte (în dezvoltare);
4. **vorbitori**: reprezentând limba turcă contemporană, în varianta sa standard, dar şi forme dialectale;
5. **informaţii private relative la vorbitori**: vorbitorii sunt anonimizaţi;
6. **metadate**: -
7. **conţinut şi mod de realizare**: textul este adnotat morfologic şi la acte de vorbire; se foloseşte EXMARaLDA pentru transcriere şi interogare;
8. **formatul înregistrărilor**: -
9. **suport pentru realizare**: realizat prin finanţare, dar şi prin voluntariat;
10. **distribuţie şi copyright**: parţial accesibil gratuit pentru cercetare (demo), descărcare - prin semnarea unei licenţe.

CORP-ORAL (Freitas and Santos, 2008)

1. **limba**: portugheza europeană;
2. **tip corpus**: dialoguri faţă în faţă, vorbire spontană în sensul că nu se impune o temă, dar vorbitorii ştiu că sunt înregistraţi;
3. **dimensiune**: 53h, dintre care transcrise ortografic: 32h, dintre care transcrise fonetic: 1h;
4. **vorbitori**: 60, din zona Lisabonei, cu vârste între 17-74 ani, nivel educaţional: de la 9 clase la doctorat;
5. **informaţii private relative la vorbitori**: -
6. **metadate**: -
7. **conţinut şi mod de realizare**: transcriere ortografică cu ELAN (cu marcarea pauzelor, respiraţiei, râsului etc.), transcrierea fonetică cu Praat, adnotare prozodică;
8. **formatul înregistrărilor**: -
9. **suport pentru realizare**: -
10. **distribuţie şi copyright**: poate fi interogat la http://spock.iltec.pt/

Spoken Portuguese Corpus

1. **limba**: portugheza;
2. **tip corpus**: comunicare orală spontană, pe subiecte legate de viaţa curentă;
3. **dimensiune**: 86 de înregistrări: Portugalia (30), Brazilia (20), Angola (5), Cap Verde (5), Guinea-Bissau (5), Mozambic (5), Sao Tome and Principe (5), Macao (5), Goa (3), East-Timor (3). Total: 8h 44min, 153.588 tokeni;
4. **vorbitori**: diversitate sociolingvistică, perioada: 1970-2001;
5. **informaţii private relative la vorbitori**: -
6. **metadate**: -
7. **conţinut şi mod de realizare**: transcrieri aliniate în format XML Exmaralda şi transcrieri în text simplu, adnotate la POS;
8. **formatul înregistrărilor**: fişiere audio WAV;
9. **suport pentru realizare**: -
10. **distribuţie şi copyright**: ELRA, gratuit pentru cercetare.

Bavarian Archive for Speech Signals Corpora (*Siemens Synthesis Corpus*)[11]

1. **limba**: germană

[11] http://www.bas.uni-muenchen.de/forschung/Bas/BasKorporaeng.html

2. **tip corpusuri**: voce-în-citire (dictări, numere de la 1 la 100, nume de străzi, coduri postale, nume de oraşe, numere de telefon etc.) şi vorbire spontană (dialoguri ale unui dispecer de taxi vorbitor de germană şi un client vorbitor de engleză înregistrat prin conexiune telefonică fixă şi GSM, în traducere; monologuri spontane pe linii de telefon publice, dialoguri om-maşină purtate pe motocicletă în mers etc.), unele cu zgomot de fond;
3. **dimensiune**: peste 500.000 fraze
4. **vorbitori**: peste 10.000, din toate categoriile de vârstă şi sociale, inclusiv adolescenţi (13-20 ani), inclusiv vorbitori profesionişti, inclusiv vorbitori nenativi de germană, din toate regiunile Germaniei, inclusiv înregistrări istorice ale unor vorbitori de germană saxonă din Transilvania din perioada anilor 1970[12] etc.;
5. **informaţii private relative la vorbitori**: -
6. **metadate**: -
7. **conţinut şi mod de realizare**: parţial adnotat zgomotul de fond, parţial adnotarea prozodiei, segmentări automate la nivel de cuvânt şi fonem, segmentare MAUS, segmentare în concordanţă cu Standardul CLIPS - BPF, TextGrid, Emu;
8. **formatul înregistrărilor**: BAS Partitur Format files, compatibilitate Verbmobil, SpeechDat Database Format etc.
9. **suport pentru realizare**: Siemens, PhonDat[13], proiectul Verbmobil, Erlanger Bahnansage[14], SPINA[15] (Robot Commands) etc.
10. **distribuţie şi copyright**: unele sub ELRA, altele direct la producător, contact: bas@bas.uni-muenchen.de.

GermaParl[16] (Blätte and Blessing, 2018)

1. **limba**: germană;
2. **tip corpus**: protocoale plenare din Bundestag;
3. **dimensiune**: 100 de mil. de tokeni;
4. **vorbitori**: parlamentari, perioada 1996-2002;
5. **informaţii private relative la vorbitori**: numele vorbitorilor, afilierea la un grup parlamentar sau la un partid politic;
6. **metadate**: numele vorbitorului, afilierea la un grup parlamentar sau la un partid politic, teme de discuţii;
7. **conţinut şi mod de realizare**: se specifică dacă enunţul este o interjecţie sau vorbire;
8. **formatul înregistrărilor**: TEI sau pachet de date R;
9. **suport pentru realizare**: -
10. **distribuţie şi copyright**: domeniul politic nu ridică probleme de copyright.

C-ORAL-ROM[17] (Nascimento *et al.*, 2002**)**

1. **limbi**: italiană, franceză, portugheză, spaniolă (patru sub-corpusuri comparabile);
2. **tip corpus**: vorbire spontană (vorbirea formală: conversaţii, inclusiv telefonice, media, conferinţe, dezbateri politice, predare; vorbire informală: conversaţii, monologuri şi dialoguri publice şi private, interviuri, întâlniri, experienţe profesionale; amintiri personale;

[12] http://www.bas.uni-muenchen.de/forschung/Bas/BasASDeng.html
[13] http://www.bas.uni-muenchen.de/forschung/Bas/BasPD1eng.html
[14] http://www.bas.uni-muenchen.de/forschung/Bas/BasERBAeng.html
[15] http://www.bas.uni-muenchen.de/forschung/Bas/BasSPINAeng.html
[16] http://www.lrec-conf.org/proceedings/lrec2018/pdf/1024.pdf
[17] http://www.elda.fr/en/proj/coralrom.html

transcrieri de filme reprezentative; eşantioane de înregistrări provenite din înregistrări de radio şi TV; interacţiuni om-maşină; dialoguri extrase dintr-un sistem automat de rezervări de bilete de tren etc.);
3. **dimensiune**: 772 de texte din 121:43:07 de ore de înregistrări audio, aproximativ 300.000 de cuvinte pentru fiecare limbă => total: 1.200.000 de cuvinte;
4. **vorbitori**: 1,427, având caracteristici diferite de vârstă, nivel de educaţie, origini sociale şi geografice;
5. **informaţii private relative la vorbitori**: -
6. **metadate**: fiecare sesiune conţine metadate cu informaţii despre vorbitori, situaţia înregistrării, calitatea acustică a înregistrării, sursa, informaţii despre subiectul sesiunii;
7. **conţinut şi mod de realizare**: sunt evidenţiate principalele evenimente paralingvistice și nonlingvistice, pauzele prozodice și evenimentele verbale, identificarea acestora din urmă fiind o caracteristică a acestui corpus; textele sunt lematizate şi adnotate cu părţi de vorbire; fiecare text transcris a trecut prin cinci stadii de validare: transcriere, revizuire, adnotare prozodică, aliniere; fiecare etapă a fost făcută de către alt lingvist;
8. **formatul înregistrărilor**: pentru sincronizarea text-sunet s-a folosit formatul WinPitch, cele patru corpusuri au fost transcrise conform standardelor ortografice de transcriere CHAT format (MacWhinney, 1994); sunetul este aliniat la text; marcaje XML;
9. **suport pentru realizare**: -
10. **distribuţie şi copyright**: corpusul este disponibil în două formate: 8 dvd-uri care conţin toate materialele sau 1 dvd care conţine o versiune criptată a corpusului; are acordul fiecărui vorbitor.

Corpus de Référence du Français Parlé (CRFP) (Teston-Bonnard and Véronis, 2004)

1. **limba**: franceză;
2. **tip corpus**: interviuri grupate pe vârstă, educaţie, tipul conversaţiei (privată, profesională sau publică);
3. **dimensiune**: 137 de înregistrări (37 de ore), 440.000 de cuvinte;
4. **vorbitori**: diferite categorii de vârstă şi nivel de educaţie;
5. **informaţii private relative la vorbitori**: -
6. **metadate**: -
7. **conţinut şi mod de realizare**: -
8. **formatul înregistrărilor**: -
9. **suport pentru realizare**: -
10. **distribuţie şi copyright**: -.

Nagoya University conversation corpus (Fujimura *et al*. 2012)

1. **limba**: japoneză
2. **tip corpus**: conversaţii spontane între prieteni, membri ai unei familii sau colegi;
3. **dimensiune**: 129 de fişiere, durata: 30-60 min; 1,5 milioane de morfeme;
4. **vorbitori**: 198 de persoane de vârste diverse, nivel de educaţie variat, la fiecare conversaţie participă 2-4 persoane; distribuţie neuniformă pe sexe (predomină cel feminin) şi pregătire profesională înaltă (mulţi absolvenţi filologi);
5. **informaţii private relative la vorbitori**: -
6. **metadate**: -
7. **conţinut şi mod de realizare**: o parte din corpus (97 de dialoguri) a fost adnotată manual cu acte de vorbire şi *sympathy* (vorbitorul arată interes pentru subiectul

conversaţiei); transcriere ortografică (nu există accent, intonaţie, proeminenţă; există doar intonaţia urcătoare din întrebări);
8. **formatul înregistrărilor**: -
9. **suport pentru realizare**: -
10. **distribuţie şi copyright**: disponibil gratuit.

JAIST Annotated Corpus (Shirai and Fukuoka, 2018[18])

1. **limba**: japoneză
2. **tip corpus**: conversaţii spontane între două persoane despre diverse subiecte;
3. **dimensiune**: 97 de dialoguri; durata: 100h;
4. **vorbitori**: la fiecare conversaţie participă 2 persoane; distribuţie neuniformă;
5. **informaţii private relative la vorbitori**: -
6. **metadate**: -
7. **conţinut şi mod de realizare**: o parte din corpus (97 de dialoguri) a fost adnotată manual cu acte de vorbire (act de dialog) şi *sympathy* (vorbitorul arată interes pentru subiect);
8. **formatul înregistrărilor**: -
9. **suport pentru realizare**: -
10. **distribuţie şi copyright**: disponibil gratuit.

4. STANDARDIZAREA CORPUSURILOR BIMODALE

Descriem în această secţiune iniţiative de standardizare care pot fi aplicate, cel puţin parţial, în proiectul nostru.

Li and Yin (2006; 2007) descriu o iniţiativă chineză de standardizare, care a fost aplicată la constituirea corpusului RASC863. Corpusul cuprinde un număr de 2200 de fraze echilibrate fonetic şi 460 de fraze frecvent folosite în viaţa zilnică, pronunţate de 80 de vorbitori, aleşi astfel încât reprezentativitatea pe categorii de vârstă, gen, studii să fie echilibrată. Vorbitorii reprezintă 10 pronunţii dialectale diferite, culese de pe toată întinderea Chinei. Înregistrările vocale sunt dublate de transcrieri. Corpusul e utilizat la antrenarea sistemelor text-vorbire. Lucrarea prezintă şi proceduri standard de producere a corpusurilor orale, pe care nu le vom detalia aici.

În contextul proiectului ReTeRom, ne interesează informaţiile codificate în metadate şi transcrieri aliniate. Ele sunt următoarele (Li and Yin, 2007):

- specificaţii relative la vorbitori: vârstă, gen, educaţie, calitate a vocii, dialect, accent, dar şi un ID ales de vorbitor sau generat automat;
- specificaţii 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;
- specificaţii 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 maşină etc.), nivelul zgomotului de fond (în db), detalii

[18] http://www.lrec-conf.org/proceedings/lrec2018/pdf/179.pdf

asupra digitalizării semnalului (numărul de biţi al eşantioanelor, rata de eşantionare, formatul de înregistrare al formei de undă), data, momentul şi locul înregistrării, un ID al înregistrării (ales sau generat) pus în legătură cu numele fişierului[19];

- specificaţii de adnotare: formatul convenţiilor de adnotare ale transcrierii sunet-text (aliniere semnal-timp şi text-timp sau direct semnal-offset caracter), adnotări ortografice, fonetice, prozodice, sintactice etc., instrumentele de segmentare şi aliniere utilizate;
- criterii de validare: dacă au fost utilizate anumite criterii de validare a corpusului, la nivel de înregistrare dar şi global, la nivelul corpusului ca întreg;
- specificaţii de distribuţie şi stocare: planul de distribuţie, mediul de stocare, condiţii de backup etc.

TEI (*Text Encoding Initiative*)[20] (Sperberg-McQueen and Burnard, 2018) dezvoltă şi menţine (prin consorţiul W3C[21]) un standard pentru reprezentarea textelor în format digital. Produsul principal este un set de linii directoare care specifică metode de codificare a textelor în formate interpretabile de către maşină (*machine-readable texts*) cu aplicaţii în domeniul umanioarelor digitale (*digital humanities*). Începând cu anul 1994, ghidul TEI a fost utilizat pe scară largă de biblioteci, muzee, editori şi cercetători pentru a pregăti texte în vederea conservării lor de lungă durată şi cercetării. În plus faţă de liniile directoare, consorţiul oferă o varietate de resurse şi evenimente de instruire pentru învăţarea TEI, informaţii despre proiectele care utilizează TEI, o bibliografie a publicaţiilor legate de TEI şi software dezvoltat pentru TEI sau adaptat acestuia.

Normele TEI P5 includ un capitol numit *Transcriptions of Speech*, în care sunt descrise problemele cele mai comune întâlnite în transcrierile de vorbire, informaţiile general utile a fi plasate în metadatele care însoţesc transcrierile vocale, precum şi probleme cu caracter specific care pot interveni în transcrierile vocale. Pe baza acestor informaţii pot fi create standarde ori convenţii de reprezentare care să facă faţă oricăror particularităţi ale corpusurilor dedicate transcrierilor vocale. Proiectul nostru exploatează nu numai transcrierile în sine, ci pe acestea în pereche cu semnalul vocal original din care au fost generate transcrierile, în vederea dezvoltării de tehnologii de recunoaştere vocală şi generare a vocii.

5. ASPECTE LEGALE

Problemele etice în înregistrarea şi publicarea în spaţiul public a ceea ce a fost produs într-un cadru privat şi destinat unei audienţe limitate sunt întâlnite mai frecvent în tratarea textelor vorbite decât în cele scrise (TEI[22]). Ca urmare, aspectele legale la constituirea unui corpus bimodal trebuie să aibă în vedere următoarele:

- drepturile de autor ale proprietarilor textelor pronunţate (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 informaţii de natură personală,
- convenţii legale semnate între producătorii, distribuitorii şi utilizatorii corpusului.

[19] Lungimea numelui fişierului trebuie să respecte specificaţiile atributelor de fişiere descrise în ISO-9960.
[20] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-sp.html
[21] https://www.w3.org/
[22] https://quod.lib.umich.edu/cgi/t/tei/tei-idx?type=pointer&value=TSOV

Convenţia de colaborare trebuie semnată între producător şi vorbitor înainte de începerea înregistrărilor.

6. CONCLUZII

Colecţia de resurse ce urmează a fi organizată (dezvoltată, standardizată) în cadrul proiectului ReTeRom nu se adresează strict membrilor consorţiului proiectului. Această activitate are un orizont ştiinţific de durată, cu un pronunţat caracter de generalitate. Ea deschide calea pentru formarea unei viziuni ştiinţifice care să ghideze colectarea, adnotarea şi distribuţia de corpusuri de acest gen în România.

Raportul de faţă lămureşte câteva chestiuni de terminologie a domeniului, sintetizează aspectele relevante în crearea corpusurilor bimodale, incluzând modalităţi de achiziţie, prezintă un număr de exemple notorii de astfel de corpusuri, cu sintetizarea a 10 trăsături care au putut fi revelate din literatură, şi prezintă preocupări de standardizare în realizarea corpusurilor bimodale, precum şi aspecte legale.

Referinţe

Bainbridge, William (2004). Berkshire Encyclopedia of Human-computer Interaction. alle

Godfrey, John J. and Edward Hollman (1997). *Switchboard I, Release 2*, Linguistic Data Consortium, Philadelphia.

Grifoni, Patrizia (2009). Multimodal Human Computer Interaction and Pervasive Services. IGI Global. p. 37.

Halle, M., & Stevens, K. (1962). *Speech recognition: A model and a program for research*. IRE Transactions on Information Theory, 8(2), pp. 155-159.

Hughes, Thad, Kaisuke Nakajima, Linne Ha, Atul Vasu, Pedro Moreno and Mike LeBeau (2010). *Bruilding transcribed speech corpora quickly and cheeply for many languages*. In: Proceedings of INTERSPEECH 2010.

Karray, Fakhreddine, Alemzadeh, Milad, Abou Saleh, Jamil, Nours Arab, Mo. (2008). *Human-Computer Interaction: Overview on State of the Art*. International Journal on Smart Sensing and Intelligent Systems. 1 (1).

Kurkovsky, Stan (2009). Multimodality in Mobile Computing and Mobile Devices: Methods for Adaptable Usability. IGI Global. pp. 210–211.

Kurosu, Masaaki (2013). Human-Computer Interaction: Interaction Modalities and Techniques, Springer, p. 366.

Li, A. and Y. Zu (2006). *Corpus Design and Annotation for Speech Synthesis and Recognition.* In Advances in Chinese Spoken Language Processing. Lee, C., Li, H., Lee, L., Wang, R., & Huo, Q. (eds.) World Scientific Publishing Co.

Li, Ai-jun and Zhi-gang Yin (2007). *Standardization Of Speech Corpus*. In: Data Science Journal, Volume 6, Supplement, 18 November.

MacWhinney, B. (2000). The CHILDES Project: Tools for Analyzing Talk. 3rd Edition. Mahwah, NJ: Lawrence Erlbaum Associates, copie online aici: http://repository.cmu.edu/cgi/viewcontent.cgi?article=1181&context=psychology.

Nascimento, Bacelar do, M. F., E. Cresti, M. Moneglia, A. Moreno Sandoval, J. Veronis, P. Martin, K. Choucri, V. Mapelli, D. Falavigna, A. Cid and C. Blum (2002). *The C-ORAL-ROM Project. New methods for spoken language archives in a multilingual romance corpus LREC*. In: M. C. RODRIGUES e C. SUAREZ ARAUJO, Proceedings of the *Third International Conference on Language Resources and Evaluation*, Paris: ELRA, vol. 1, pp. 2-10.

Palanque, Philippe and Paterno, Fabio (2001). *Interactive Systems. Design, Specification, and Verification*. Springer Science & Business Media, p. 43.

Pitt, M.A., L. Dilley, K. Johnson, S. Kiesling, W. Raymond, E. Hume and E. Fosler-Lussier (2007). *Buckeye Corpus of Conversational Speech (2nd release)* [www.buckeyecorpus.osu.edu] Columbus, OH: Department of Psychology, Ohio State University.

Rasso, T. and H. Mello (eds.) (2014). *Spoken Corpora and Linguistic Studies*, John Benjamins.

Ruhi, Ş. (2011). *Creating a sustainable large corpus of spoken Turkish for multiple research purposes (Ulusal Konuşma ve Dil Teknolojileri Platformu Kuruluşu: Türkçede Mevcut Durum Çalıştayı'nda sunulan görüş bildirisi)*. TÜBİTAK-TÜSSİDE, TÜBİTAK-BİLGEM, 6-7 Ekim 2011, Gebze.

Sperberg-McQueen, C.M. and Lou Burnard (2018). Original editors, revised and expanded under the supervision of the Technical Council of the TEI Consortium. TEI P5: Guidelines for Electronic Text Encoding and Interange, version 3.3.0, last update: 31st January 2018, revision: f4d8439.

Shirai, K. and Fukuoka, T. (2018). *JAIST Annotated Corpus of Free Conversation*. In: Proceedings of LREC 2018, pp. 741-748.

Williams, J. R. (1998). Guidelines for the use of multimedia in instruction. In: Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting, pp. 1447-1451.

Ziefle, M. (December 1998). *Effects of display resolution on visual performance*. In: Human factors. 40 (4), pp. 554-680.