

RAPORT ȘTIINȚIFIC proiect complex ReTeRom, etapa II - noiembrie 2019

Proiectul 1: CoBiLiRo

Activitatea A2.5 Management și diseminare

Denumire Etapă: **Activități suport - Diseminare și participare la manifestări tehnico-științifice, inclusiv în mass-media**

Responsabil și contribuții din colectivul Universității “Alexandru Ioan Cuza” din Iași (UAIC): **Daniela Gîfu**, Mihaela Onofrei.

Responsabil și contribuții din colectivul Universității Politehnica din București (UPB): **C. Burileanu**, D. Burileanu, H. Cucu.

Responsabil din colectivul Universității Tehnice din Cluj-Napoca (UTCN): **Adriana Stan**.

Responsabil din colectivul Institutului de Cercetări pentru Inteligență Artificială “Mihai Drăgănescu” București (ICIA): **Verginica Mititelu**.

REZUMATUL ETAPEI

A cincea etapă a proiectului 1 CoBiLiRO din anul 2019 prevede procesul de *Diseminare* planificat în Agenda ReTeRom, centrat pe prezentarea rezultatelor obținute în etapele precedente, totodată a inițiativelor partenerilor către actori-cheie vizați (ex. mass-media).

Rezultate Etapă: Raport D1.3.5 privind diseminarea proiectului pe diverse canale de comunicare (științifică și jurnalistică). Instituția implicată în realizarea acestei etape a fost UAIC, sprijinită de toți ceilalți parteneri, răspunzători pentru celelalte 3 componente (TEPROLIN, TADARAV, SINTERO).

În această etapă procesul de diseminare a implicat aducerea la cunoștința grupului țintă (lingviști, informaticieni, studenți etc.) a rezultatelor proiectului de până acum. Pentru diseminarea eficientă

a acestora, s-a stabilit o procedură, agreată de întreg consorțiul ReTeRom, care s-a bazat pe un dialog permanent în diverse momente ale anului în curs (întâlniri de proiect, ședințe on-line), și centrată pe setul de întrebări-cheie (*ce, cum, de ce, când, către cine și unde* vor fi împărtășite rezultatele obținute pe fiecare componentă a proiectului complex).

Introducere

Activitatea este inclusă în categoria *Activități suport - Diseminare* și participare la manifestări tehnico-științifice. Indicatorii de realizare au fost planificați după cum urmează: *diseminarea corpusului bimodal, valorizare și utilizare; cel puțin 5 articole la conferințe indexate ISI și alte 2 la reviste de prestigiu, stagii de practică, actualizarea paginii web a proiectului; promovarea întâlnirilor de proiect în mass-media.*

1. Diseminarea corpusului bimodal, valorizare și utilizare

Pentru asigurarea promovării, vizibilității proiectului și, totodată, a diseminării rezultatelor obținute în cadrul acestuia, s-au actualizat continuu următoarele platforme online:

1.1. Proiectare paginii web ReTeROM

Pagina web pentru un nou corpus folosit în sinteza text-vorbire cu expresivitate:
<https://speech.utcluj.ro/corpora/mara.html>

1.2. Proiectare paginii web de control automat al prozodiei

Pagina web / demonstrator conține prima versiune a modului de control automat al prozodiei:
https://speech.utcluj.ro/sintero/prosody_examples/

1.3. Proiectare paginii web demonstrator de sinteză text vorbire

Pagina web pentru demonstrarea sintezei folosește rețele DNN de tip Tacotron-GST:

<https://speech.utcluj.ro/sintero/dnn-samples/>

1.4. Proiectare paginii web demonstrator de modelare a prozodiei și adaptare la noi vorbitori

Pagina web conține mostre audio privind adaptarea sistemelor de sinteză text-vorbire la vorbitori și la stilul de vorbire, folosind Tacotron – GST: https://speech.utcluj.ro/sintero/spkadapt_prosody_examples/

2. Evenimente organizate

(a) **18 noiembrie 2019 - ReTeRom Workshop 3** - în cadrul celei de-a 14-a ediții a Conferinței Internaționale “Linguistic resources and tools for processing natural language” - <https://profs.info.uaic.ro/~consilr/>. Workshopul a fost găzduit de Universitatea Tehnică Cluj-Napoca și a fost promovat inclusiv pe site-ul Universității Babeș-Bolyai din Cluj-Napoca, <https://news.ubbcluj.ro/cea-de-a-xiv-a-editie-a-conferintei-internationale-resurse-lingvistice-si-instrumente-pentru-prelucrarea-limbii-romane-organizata-la-ubb/>.

În cadrul acestui atelier de lucru au fost prezentate rezultatele pentru cele 4 subproiecte componente (CoBiLiRO, TEPROLIN, TADARAV și SINTERO). Discuțiile s-au axat pe 6 direcții tematice, în care membrii proiectului au interacționat prezentând probleme de mare interes curent, discutând chestiuni care privesc evoluția proiectului și propunând soluții: (1) Integrare unelte software; (2) Copyright și GDPR date audio și text; (3) Sinteza cu expresivitate; (4) Procesare text (lexicoane, noi algoritmi); (5) Alinieri audio și text (metode, demonstratoare); (6) Noi tendințe în domeniu (deep learning, text, audio, supervised/unsupervised).

Lista care urmează conține conferințe prezentate de membri ai consorțiului ReTeRom pe subiecte apropiate de tematicile proiectului:

(a) Dan Tufiș: *Resources and Technologies for Speech and Language Processing of Romanian*, conferință invitată la cea de a 10-a conferință *Speech Technology and Human-Computer Dialogue*, 10-12 Oct 2019, Timișoara.

(b) 2 participări (Beata Lorincz și Maria Nuțu) la **Eastern European Machine Learning Summer School (EEML)**, 1-6 iulie 2019, București, România - <https://www.eeml.eu/>.

(c) Dan Cristea: *The COBILIRO Project: Building and Distributing a Bimodal Corpus for Romanian Language*, în cadrul celei de a 14-a conferințe internaționale *Linguistic Resources and Tools for Natural Language Processing*, ConsILR-2019, 19 Nov. 2019, Cluj-Napoca.

3. Publicații care menționează proiectul ReTeRom

- (1) Pistol, Ionuț Cristian, and Andrei Arusoai. AIM: Designing a language for AI models. In: *Procedia Computer Science* 159 (2019): 202-211
<https://www.sciencedirect.com/science/article/pii/S1877050919313547>.
- (2) Dan Cristea, Cristian Pădurariu, Șerban Boghiu, Daniela Gîfu, Mihaela Onofrei, Diana Trandabăț, Ionuț Cristian Pistol, Anca-Diana Anca Bibiri, Andrei Scutelnicu. The COBILIRO Project: Building and Distributing a Bimodal Corpus for Romanian Language. In: *Proceedings of the 14th International Conference Linguistic Resources and Tools for Processing The Romanian Language, ConsILR-2019, 18-20 Nov. 2019, Cluj-Napoca, Romania, Mihaela Onofrei et al.(eds.), "Alexandru Ioan Cuza" University Publishing House, Iași, 2019, pp. 13-23.*
- (3) Vasile Păiș, Dan Tufiș, Radu Ion. Integration of Romanian NLP tools into the RELATE platform. In *Proceedings of the 14th International Conference on Linguistic Resources and Tools for Natural Language Processing (ConsILR 2019)*, Mihaela Onofrei *et al.*(eds.), "Alexandru Ioan Cuza" University Publishing House, Iași, 2019, pp. 181-192.
- (4) Mihaela Onofrei, Anca-Diana Bibiri, Constantin Dragoș Nicolae, Dan Tufiș, Dan Cristea. *Proceedings of the 14th International Conference "Linguistic Resources and Tools for processing of the Romanian language"*, Cluj-Napoca, 2019, ISSN: 1843-911.
- (5) Daniela Gîfu, Alex Moruz, Cecilia Bolea, Anca Bibiri, Maria Mitrofan. The Methodology of Building CoRoLa In: *Revue Roumaine de Linquistique (Romanian Review of Linguistics)*, Vol. 64, 2019, Publishing House of the Romanian Academy, ISSN: 0035-3957, indexed by ISI Web of Science, (impact factor = 0.189/2018) -
http://www.lingv.ro/index.php?option=com_content&view=article&id=342%3Arrl-arhiva-2019&catid=36%3Areviste-ilb&Itemid=1.
- (6) Alexandru-Lucian Georgescu, Horia Cucu, Corneliu Burileanu, "Progress on automatic annotation of speech corpora using complementary ASR systems," in the *Proceedings of the 42nd International Conference on Telecommunications and Signal Processing (TSP)*, 2019, Budapest, Hungary.

- (7) Gheorghe Pop, Serban Mihalache, Dragoş Burileanu, "Forensic Recognition of Narrowband AMR Signals," in Proceedings of the 10th Conference on Speech Technology and Human-Computer Dialogue (SpeD), Timișoara, Romania, 2019.
- (8) Alexandru-Lucian Georgescu, Horia Cucu, Corneliu Burileanu, "Kaldi-based DNN architectures for speech recognition in Romanian," in the Proceedings of the 10th Conference on Speech Technology and Human-Computer Dialogue (SpeD), Timișoara, Romania, 2019.
- (9) Gheorghe Pop and Dragoş Burileanu, "Speech Enhancement for Forensic Purposes," in UPB Scientific Bulletin, Series C, Vol. 81, Issue 3, pp. 41-52, 2019.
- (10) Florin Iordache, Alexandru-Lucian Iordache, Dan Oneață, Horia Cucu, "Romanian Automatic Diacritics Restoration Challenge", in the Proceedings of the 14th International Conference on Linguistics Resources and Tools for Natural Language Processing, Cluj-Napoca, Romania, 2019.
- (11) Adriana Stan, "Input Encoding for Sequence-to-Sequence Learning of Romanian Grapheme-to-Phoneme Conversion", In Proceedings of the 10th IEEE International Conference on Speech Technology and Human-Computer Dialogue (SpeD), Timișoara, Romania, 2019
- (12) Beáta Lőrincz, Maria Nuțu, Adriana Stan, "Romanian Part of Speech Tagging using LSTM Networks", In Proceedings of the IEEE 15th International Conference on Intelligent Computer Communication and Processing, Cluj-Napoca, Romania, 2019.
- (13) Maria Nuțu, Beáta Lőrincz, Adriana Stan, "Deep Learning for Automatic Diacritics Restoration in Romanian", In Proceedings of the IEEE 15th International Conference on Intelligent Computer Communication and Processing, Cluj-Napoca, Romania, 2019.
- (14) David A. Braude, Matthew P. Aylett, Caoimhin Laoide-Kemp, Simone Ashby, Kristen M. Scott, Brian O Raghallaigh, Anna Braudo, Alex Brouwer, Adriana Stan, "All Together Now: The Living Audio Dataset", In Proceedings of Interspeech, Graz, Austria, 2019.
- (15) Verginica Barbu Mititelu, Mihaela Cristescu, Mihaela Onofrei, The Romanian Corpus Annotated with Verbal Multiword Expression. In: Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019), August 2019, Florence, Italy, ACL, pp. 13-21.

4. Întreținerea paginii web a proiectului complex ReTeROM

Pentru asigurarea promovării, vizibilității proiectului și diseminarea rezultatelor obținute în cadrul acestuia s-a actualizat continui platforma online (<http://dev.racai.ro/ReTeRom/>), compusă dintr-un “landing zone” și patru secțiuni: “DESCRIERE”, “RAPOARTE”, “ECHIPA” și “CONTACT”.

Meniul este unic și poate fi accesat din orice locație a paginii, întregul conținut fiind cuprins în aceste legături. Regăsim aici cele 4 subproiecte componente: CoBiLiRo (Corpus bimodal pentru limba română adnotat pe multiple niveluri); TEPROLIN (Tehnologii pentru procesarea limbajului natural – text); TADARAV (Tehnologii pentru adnotarea automată a datelor audio și pentru realizarea interfețelor de recunoaștere automată a vorbirii); SINTERO (Tehnologii de realizare a interfețelor om-mașină pentru sinteza text-vorbire cu expresivitate).

5. Comunicate de presă

Confluente literare, ediția nr. 3244, Anul IX, 18 noiembrie 2019.
https://confluente.org/daniela_gifu_1574084975.html.

6. Concluzii

Livrabilul de față descrie modalitățile folosite de diseminare a proiectului ReTeRom pe parcursul anului 2019: proiectarea și realizarea de pagini web, organizarea de evenimente științifice special dedicate sau în care proiectul a fost făcut cunoscut participanților, publicații științifice, comunicate de presă.

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