RAPORT ȘTIINȚIFIC proiect complex ReTeRom, etapa II - noiembrie 2019

Proiectul 2: TEPROLIN

Activitatea 2.9

Denumire activitate: Prelucrarea părții textuale a corpusului bimodal colectat în proiectul 1. Validarea și corectarea (dacă este necesară) a erorilor de prelucrare

Autori: ICIA

REZUMATUL ETAPEI

Această activitate a proiectului 2, TEPROLIN, are rolul de a asigura prelucrarea transcrierilor corespunzătoare înregistrărilor audio colectate în cadrul proiectului COBILIRO. Prelucrarea se face cu un lanț de procesare care presupune 7 tipuri de operații, dezvoltat în cadrul proiectului TEPROLIN. Corpusurile textuale procesate sunt disponibile partenerilor în platforma COBILIRO.

Rezultatele activității: (i) Raport asupra prelucrării părții textuale a corpusului bimodal colectat în proiectul 1;

(ii) corpusul textual este prelucrat, validat și corectat și accesibil partenerilor.

În raportul științific asupra activității A1.2. Inventarierea colecțiilor de date lingvistice românești disponibile la parteneri sau în terțe coaliții din etapa I/2018 a proiectului COBILIRO, erau inventariate și descrise 11 corpusuri colectate în consorțiul ReTeRom. Datele au fost descrise conform unui set de trăsături unitar, care să faciliteze armonizarea și integrarea lor în platforma COBILIRO rezultată în urma activităților 2.1, 2.2., 2.3., 2.4 din cadrul Proiectului 1, în etapa 2019.

Activitatea de prelucrare a transcrierilor presupune, în contextul proiectului TEPROLIN, mai multe etape de procesare succesivă:

- segmentare la nivel de propoziție (atunci când este necesar);
- segmentare la nivel de cuvânt: separarea cuvintelor și punctuației din text ca unități de segmentare distincte (sau tokeni);
- introducerea diacriticelor, atunci când lipsesc;
- adnotare morfo-sintactică a cuvintelor;
- lematizarea sau identificarea formei de dicționar a cuvintelor;
- adnotarea sintactică a cuvintelor;
- detectarea entităților denumite;

Aceste procesări au fost efectuate cu lanțul de prelucrare TEPROLIN (ION, 2018), parte din platforma RELATE (Păiș et al., 2019), despre care am vorbit și în raportul asupra Activității 2.8. Corpusurile au fost încărcate în platforma COBILIRO (Cristea et al., 2019), care integrează lanțul TEPROLIN și îl execută de fiecare dată când o resursă nouă este introdusă și prelucrarea cu TEPROLIN este bifată în interfață.

În continuare prezentăm informații cu privire la numărul de fișiere, propoziții, leme și forme unice, verbe, adjective, adverbe și substantive, corespunzător fiecărui corpus adnotat în cadrul acestei activități.

Corpus	Nr. fișiere/prop	Nr. cuvinte	Verbe	Adverbe	Adjective	Substantive	Nr. leme unice	Nr. forme unice
IIT	7001	146633	23590	10866	6743	30710	13853	23603
RACAI	3827	87433	14103	3655	4989	27015	17452	26474
RADOR	12551	396804	62286	24554	21085	103692	21642	39394
RASC	2177	30520	4909	1375	2839	9044	6186	9816
SWARA	17	257684	43690	11984	10234	55115	4023	5529
Adevărul	1871/ 30391	694681	109689	33229	32875	196299	34061	51033
Mara	1/5573	116387	23028	10878	3772	17573	4592	7486
RSC	1/918	13160	2640	928	505	2385	2319	3509
SSC-eval	1/3035	35947	5029	2438	2139	9096	4603	7290
SSC-train	1/53898	241006	37289	18163	13307	64814	8452	16829
SSC-train2	1/170292	983844	122525	63291	56526	256677	12851	26743
SOROES	1/166	896	187	67	13	169	175	223
TOTAL	27450/447894	3581242	551914	215695	196839	944449	127306	233504

Referințe

Ion, R. (2018) TEPROLIN: An Extensible, Online Text Preprocessing Platform for Romanian. In Proceedings of the International Conference on Linguistic Resources and Tools for Processing Romanian Language (ConsILR 2018), November 22-23, 2018, Iași, Romania.

V. Păiș, D. Tufiș, R. Ion, ”Integration of Romanian NLP tools into the RELATE platform”. In Proceedings of the International Conference on Linguistic Resources and Tools for Processing Romanian Language – CONSILR 2019, pages 181-192

Dan Cristea, Cristian Pădurariu, Șerban Boghiu, Daniela Gîfu, Mihaela Onofrei, Diana Trandabăț, Ionuț Cristian Pistol, Anca Bibiri and Andrei Scutelnicu, The CoBiLiRo project: Building and Distributing a Bimodal Corpora for the Romanian Language In Proceedings of the International Conference on Linguistic Resources and Tools for Processing Romanian Language – CONSILR 2019, pages 13-24.

Diseminarea activităților din proiectul component TEPROLIN a fost :

1. Dan Tufiș. 2019. Language Technology and Digital Culture. invited talk at European Conference on Exposing Online the European Cultural Heritage: the Impact of Cultural Heritage on the Transformation of the Society, Iași, 17-18 aprilie, 2019
2. Dan Tufiș. 2019. Language technologies and the challenges for Digital Single Market. invited pannelist at IMPACT 2019, 21-22 May, Kracow, Poland
3. Dan Tufiș, Resources and Technologies for Speech and Language Processing of Romanian, invited talk at the 10th Conference on Speech Technology and Human-Computer Dialogue, 10-12 Oct 2019, Timișoara.
4. Dan Tufiș. 2019. “Inteligența artificială și provocările ingineriei lingvistice”. Plenary talk to the international workshop BACStud 2019 (5th edition), 17-19 October, 2019

5. 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.
6. Vasile Păiș, Dan Tufiș, Radu Ion, ”Integration of Romanian NLP tools into the RE-LATE platform”. In Proceedings of the International Conference on Linguistic Resources and Tools for Processing Romanian Language – CONSILR 2019, pages 181-192