RAPORT ȘTIINȚIFIC proiect complex ReTeRom

Etapa II - noiembrie 2019

Proiectul 2: TEPROLIN

Activitatea 2.8

Implementarea prototipului de platformă integrată și configurabilă, testarea, evaluarea și validarea prototipului

V. Păiș

Institutul de Cercetări pentru Inteligență Artificială, „Mihai Drăgănescu”

Academia Română

Cuprins

Rezumat

În cadrul activității 2.8 a proiectului 2, TEPROLIN, s-a urmărit realizarea unei platforme integrate care să permită procesarea automată a unui corpus textual de mari dimensiuni, în vederea adnotării pe diferite nivele lingvistice. În acest context, corpusul avut în vedere este cel colectat în cadrul Proiectului 1 (COBILIRO), conform cu obiectivul general al proiectului TEPROLIN.

Rezultatele activității:

(i) Prototipul platformei integrate, disponibil online la adresa: http://relate.racai.ro

(ii) Raport asupra implementării prototipului de platformă integrată și configurabilă, asupra testării, evaluării și validării prototipului.

Introducere

În cadrul etapei anterioare, au fost identificate o serie de module software care permit adnotarea textelor în limba română pe diferite nivele lingvistice, incluzând lematizare, identificare părți de vorbire, despărțire în silabe, transcriere fonetică. Acestea au fost colectate de la parteneri, armonizate și integrate în cadrul unui serviciu web denumit ”TEPROLIN”, după numele acestui proiect. Aceste rezultate au fost prezentate în (Ion, 2018) și în rapoartele corespunzătoare activităților etapei I, disponibile pe site-ul proiectului RETEROM[1] în secțiunea ”rapoarte”.

Pornind de la această primă integrare, s-a urmărit realizarea unei integrări la un nivel superior a modulelor disponibile prin intermediul serviciul web TEPROLIN cu diferitele resurse adiționale existente, cum ar fi interfețele de căutare din cadrul corpusului CoRoLa (Barbu et al., 2018), reprezentările distribuționale ale cuvintelor antrenate automat pe corpusul CoRoLa (Păiș, Tufiș. 2018), WordNet-ul românesc (Tufiș, Mititelu, 2014). Totodată, platforma a urmărit expunerea funcționalităților prin intermediul unei interfețe grafice disponibilă online de tip ”portal web”, care să permită interacțiunea facilă a utilizatorilor cu tehnologiile expuse, fără a fi necesare cunoștințe de programare.

Având în vedere necesitatea prelucrării automate a unui corpus de mari dimensiuni, s-a urmărit și paralelizarea lanțurilor de procesare utilizate, în scopul reducerii timpului total de prelucrare. Acest lucru a condus la implementarea unui mecanism de control al procesărilor sub formă de ”job”-uri care pot fi apoi distribuite la nivelul unui server pe mai multe procese sau la nivelul unei rețele de calcul pe mai multe noduri.

[1] http://www.racai.ro/p/reterom/#rapoarte

Cerințe funcționale

O primă etapă a realizării prototipului de platformă a constat în formularea unor cerințe funcționale care au stat apoi la baza implementării. De asemenea, acestea au fost apoi urmărite în vederea evaluării funcționale și validării prototipului realizat.

Următoarele aspecte au fost avute în vedere, formulate din perspectiva unui utilizator al platformei:

- Expunerea funcționalităților deja integrate în serviciul web TEPROLIN (Ion, 2018). Utilizatorii ar trebui să poată interacționa cu acesta în 3 moduri:
    - Utilizarea tuturor modulelor disponibile, prin introducerea unui text de mici dimensiuni și lansarea în execuție a operației de adnotare
    - Configurarea unui flux de adnotare pe baza operațiunilor disponibile și lansarea în execuție, serviciul urmând să se reconfigureze automat în cazul în care utilizatorul cere o operațiune dependentă de alte adnotări fără ca acestea să fie marcate în mod explicit
    - Includerea tuturor operațiunilor în fluxul de adnotare pe volume mari de date prin paralelizare.
- Oferirea rezultatelor adnotării în formate de fișiere standard, precum și vizualizarea grafică în cadrul platformei (în special pentru texte scurte)
- Integrarea funcționalităților relevante asociate corpusului CoRoLa în mod interactiv în cadrul elementelor de vizualizare ale platformei:
    - căutare în componenta scrisă a corpusului CoRoLa
    - căutare în componenta vorbită a corpusului CoRoLa
    - accesare reprezentări distribuite (”word embeddings”) derivate din corpusul CoRoLa
- Căutare în WordNet-ul românesc și exploatarea alinierii cu WordNet-ul în limba engleză prin oferirea posibilității unei căutări ”aliniate” oferind rezultate din cele două WordNet-uri
- Implementarea unui mecanism de management al corpusului textual:
    - Încărcare fișiere text: direct sau sub formă de arhive
    - Procesare paralelă cu utilizarea resurselor disponibile la nivelul unui server sau la nivelul unei rețele de calcul
    - Vizualizare rezultate procesare
    - Descărcare fișiere originale și prelucrate: atât direct, fișier cu fișier, cât și sub forma unei arhive

Pornind de la aceste cerințe din perspectiva utilizatorului, au fost formulate adițional o serie de cerințe tehnice în vederea implementării:

- Utilizarea unor tehnologii web moderne
- Posibilitatea configurării resurselor de calcul disponibile în vederea paralelizării fluxurilor de prelucrare, cu definirea adreselor (”endpoint”-urilor) de conectare la diferitele servicii de adnotare
- Configurarea adreselor asociate interfețelor oferite de corpusul CoRoLa și WordNet
- Separarea funcționalităților în două componente:
    - Fără autentificare: acces la adnotări pe texte scurte și interacțiune cu diferitele unelte disponibile în acest context

- Cu autentificare: necesită autentificare pe bază de nume de utilizator și parolă și permite acces la fluxul paralel de adnotare, importarea unui corpus de mari dimensiuni, exportarea rezultatelor.

Implementare prototip

Pornind de la cerințele funcționale prezentate anterior, s-a urmărit realizarea unui prototip de platformă care să respecte toate aceste cerințe. Fiind un prototip, implementarea s-a concentrat pe realizarea tuturor integrărilor presupunând prezența unui utilizator competent, familiarizat cu necesitățile proiectului RETEROM. Astfel, paginile din cadrul platformei conțin informațiile de bază care permite interacțiunea cu aceasta, fără a include mesaje adiționale care să permită ghidarea unui utilizator începător, complet nefamiliar cu problematica adnotării și analizei unui corpus textual. Astfel de mesaje adiționale ar putea fi introduse ulterior, în vederea transformării platformei într-una mai prietenoasă cu eventuali utilizatori începători. De asemenea, au fost tratate erorile de bază care determină blocarea lanțurilor de procesare, dar nu au fost tratate erori provenind din utilizarea necorespunzătoare a platformei.

Tehnologiile utilizate sunt: HTML5, JavaScript, CSS 3, PHP 7. Acestea reprezintă tehnologii moderne utilizate în platforme web avansate. Interfața este împărțită în 3 zone:

- Meniul general al platformei: se află în partea stângă și oferă posibilitatea accesului direct la diferitele elemente componente.
- Bara de titlu: se află în partea de sus, conține numele platformei și oferă posibilitatea conectării unui utilizator la platformă sau modificarea preferințelor personale în cazul în care este deja conectat, precum și deconectarea.
- Zona centrală: afișează conținutul din contextul curent în care se află utilizatorul. Permite și navigarea între elemente acolo unde astfel de legături sunt definite și posibile conform contextului curent.

Pentru a oferi o experiență cât mai confortabilă utilizatorilor, diferitele elemente ale platformei pot fi minimizate sau maximizate pentru ca informațiile de interes să poată ocupa o zonă cât mai mare. Astfel, meniul platformei poate fi minimizat rămânând vizibile doar pictogramele asociate elementelor de meniu, iar diferite zone de lucru (cum ar fi unele tabele cu multe coloane) pot fi maximizate.

Diagrama generală a platformei este prezentată în următoarea Figura 1. Aceasta respectă cerințele formulate anterior, privind separarea într-o parte publică și una privată. În cadrul părții private fiind implementate diferitele elemente corespunzătoare procesării unui corpus de mari dimensiuni.

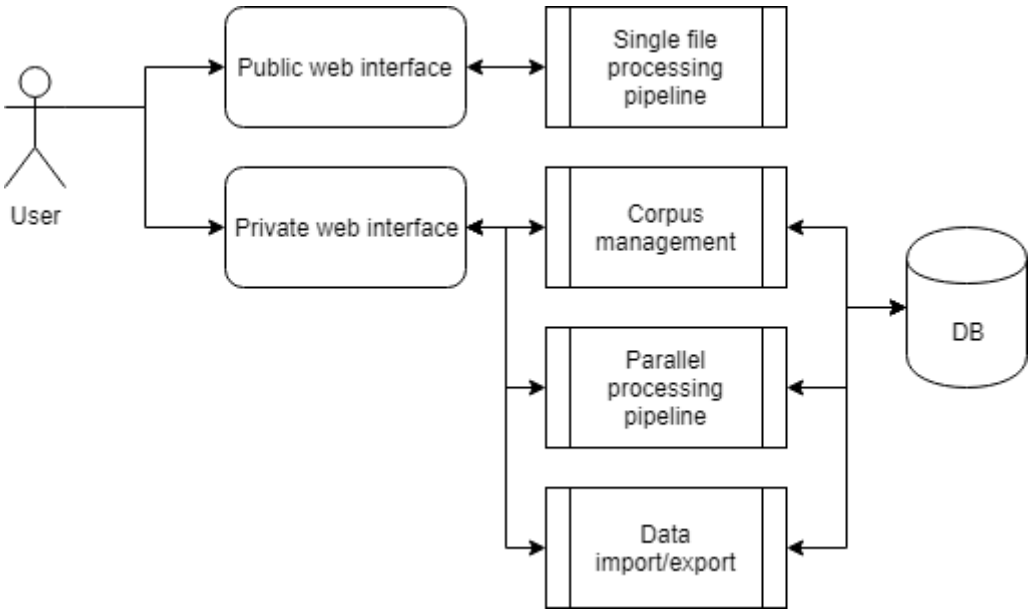

Figura 1. Diagrama generală a platformei

Cum se poate vedea din Figura 1, partea publică nu conține elemente de tip bază de date care să permită stocarea informațiilor procesate prin platformă. Acest caz corespunde cerințelor referitoare la texte de mici dimensiuni care sunt procesate fără utilizarea unui cont de utilizator.

Integrarea serviciului web TEPROLIN este realizată în cadrul a două pagini ale platformei realizându-se procesarea completă sau pe baza unui flux definit de utilizator.

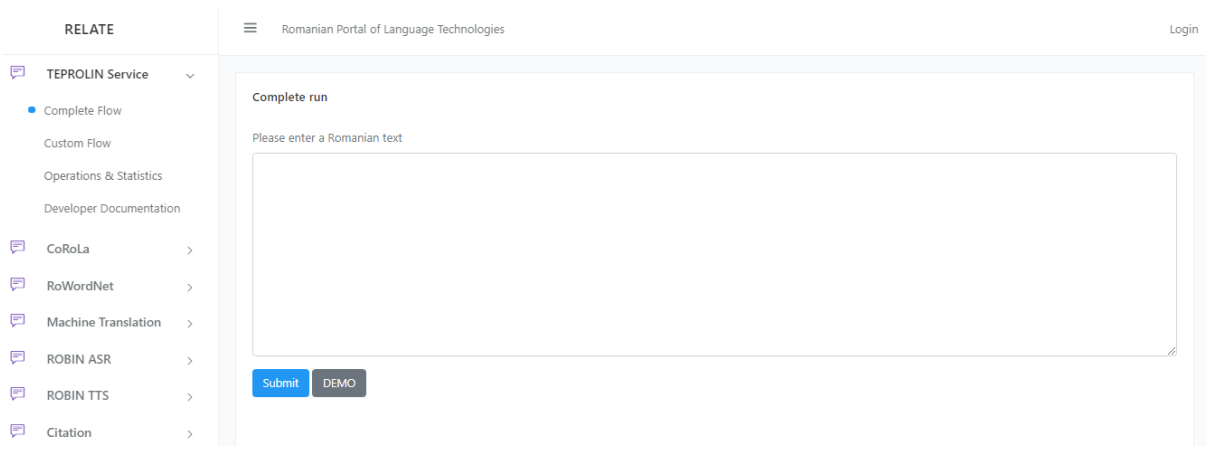

Figura 2. Integrarea serviciului web TEPROLIN într-un flux complet de adnotare

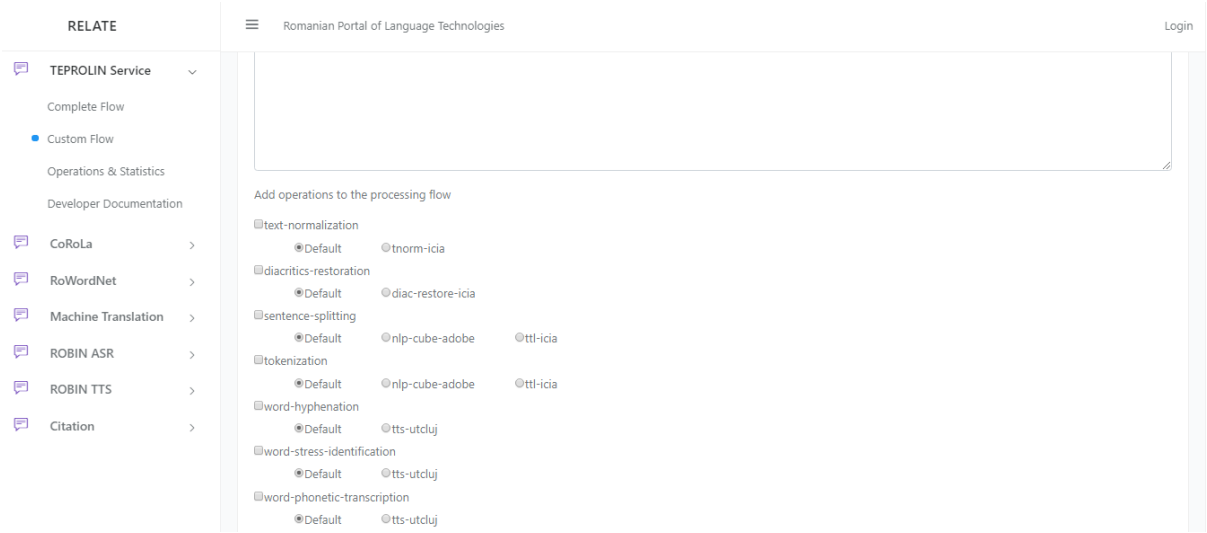

Figura 3. Integrarea serviciului web TEPROLIN cu posibilitate definirii unui flux de prelucrare

În Figura 3 este prezentată pagina din cadrul platformei care permite definirea unui flux de prelucrare format din diferitele module ale serviciului web TEPROLIN. Acolo unde sunt disponibile mai multe module cu funcționalități similare, utilizatorul poate selecta modulul dorit. De exemplu, segmentarea unui text la nivel de tokeni se poate realiza cu unul din modulele: ”nlp-cube” sau ”ttl”. De asemenea, utilizatorul poate selecta doar operațiunile de care are nevoie.

Operațiunile sunt prezentate într-o ordine logică, de la cele mai simple către cele mai complexe. În general o operațiune de nivel superior va necesita operațiunile mai simple prezentate în interfață înaintea ei. Dacă utilizatorul uită să bifeze operațiunile mai simple, serviciul web TEPROLIN va rula automat toate operațiunile de care are nevoie pentru a satisface cea mai complexă operațiune selectată.

Având în vedere cerința de a prezenta rezultatele în formate de date standard, platforma convertește automat rezultatele adnotării în format tabulat CoNLL-U sau CoNLL-U Plus. Formatul CoNLL-U[2] specifică următoarele 10 coloane:

1. ID: Numărul tokenului în propoziția curentă.
2. FORM: Tokenul așa cum a rezultat în urma operației de segmentare.
3. LEMMA: Lemma aferentă tokenului identificat.
4. UPOS: Partea de vorbire în format ”universal part-of-speech”[3].
5. XPOS: Tag-ul aferent părții de vorbire așa cum este returnat de modulul de adnotare.
6. FEATS: Caracteristici morfologice sau ”_” dacă această informație nu este disponibilă.
7. HEAD: Id-ul tokenului de care se leagă tokenul curent într-un graf al dependințelor, sau zero.
8. DEPREL: Relația de dependință.
9. DEPS: La ora actuală nu este completată această informație, fiind mereu ”_”.
10. MISC: Alte adnotări.

[2] https://universaldependencies.org/format.html
[3] https://universaldependencies.org/u/pos/index.html

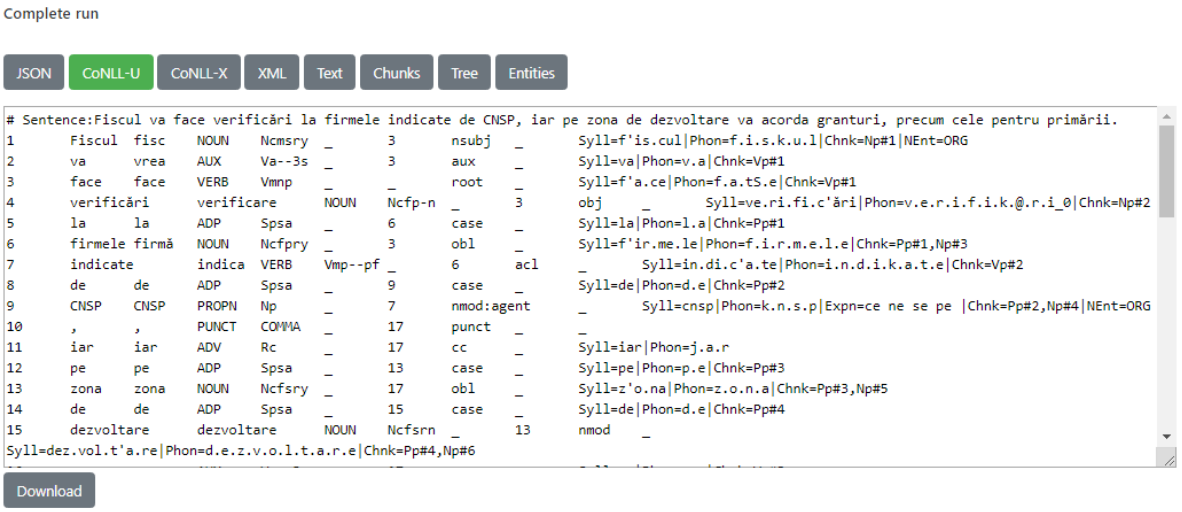

Figura 4. Vizualizare CoNLL-U

Pe lângă vizualizarea în format CoNLL-U prezentată în Figura 4, mai sunt disponibile vizualizări în format JSON (Figura 5), CoNLL-X (Figura 6) și XML (Figura 7). Fișierul cu adnotări vizualizat în fiecare din formate, poate fi descărcat cu ajutorul butonului ”Download” pentru a putea fi utilizat în prelucrări ulterioare.

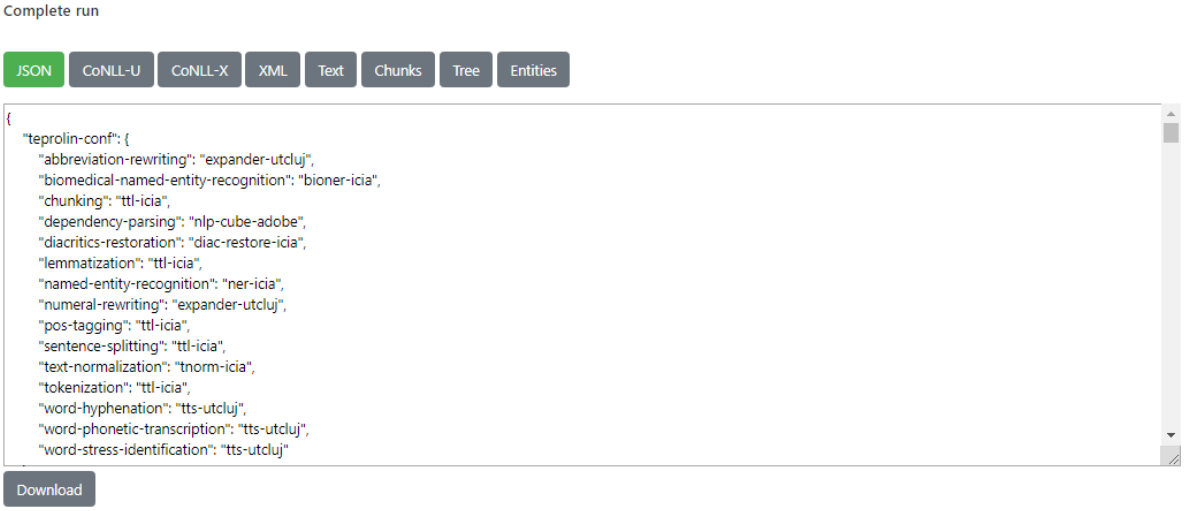

Figura 5. Vizualizare JSON

Figura 6. Vizualizare CoNLL-X

Figura 7. Vizualizare XML

În cazul utilizării unei operații care modifică textul, cum ar fi restaurarea de diacritice, este posibilă vizualizare textului transformat într-un element de vizualizare specific (Figura 8). De asemenea, noul text poate fi descărcat sub forma unui fișier.

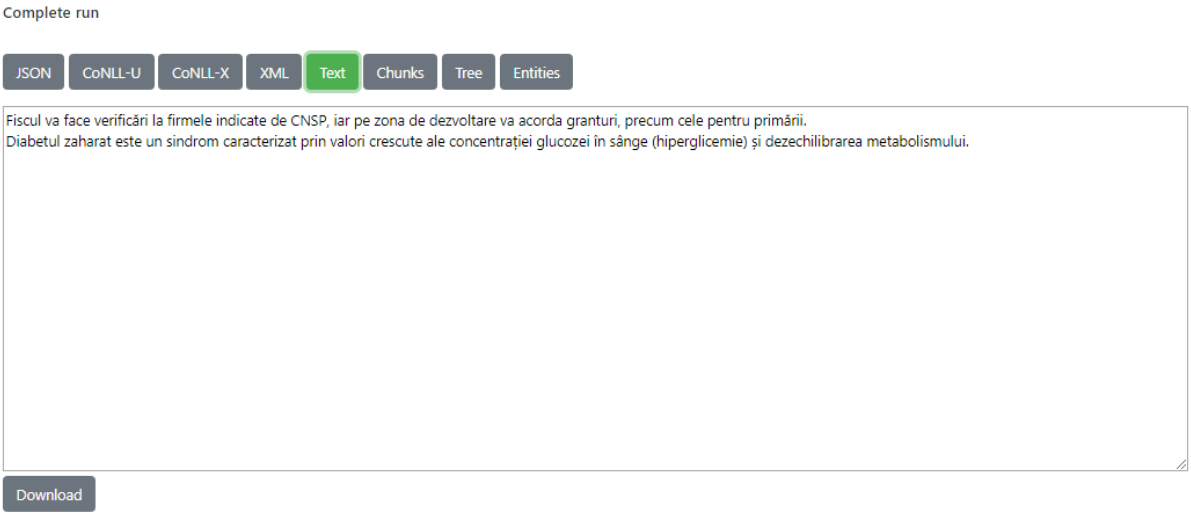

Figura 8. Vizualizare text, posibil modificat în urma unei operațiuni cum ar fi restaurare diacritice

Dacă au fost identificate entități cu nume în text, în urma execuției unei operații de identificare a entităților, aceste pot fi vizualizate într-un element specific, prezentat în Figura 9. La baza acestei vizualizări se află ”BRAT Rapid Annotation Tool” (Stenetorp et al., 2012).

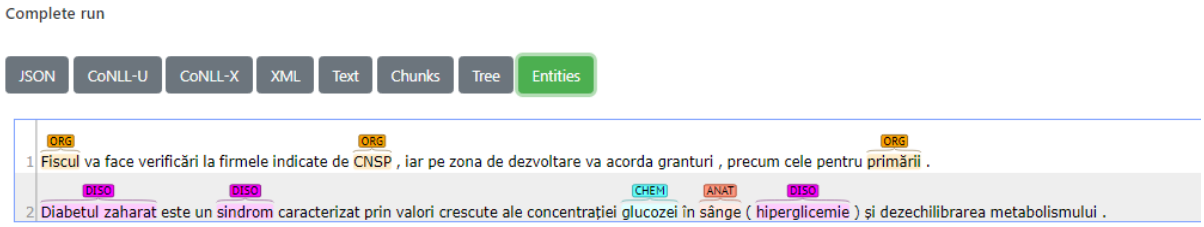

Figura 9. Vizualizare entități în text

O reprezentare grafică a fiecărei propoziții pe baza dependențelor identificate între cuvinte, se realizează utilizând un element de tip graf (Figura 10). În acest caz, fiecare token este un nod al grafului, iar muchiile sunt reprezentate de relațiile identificate între aceștia. În această vizualizare, în partea dreaptă este inclusă o listă de informații asociate nodului selectat, precum și legături către alte module din platformă care vor fi interogate pornind de la informația contextuală selectată.

Figura 10. Reprezentare grafică sub formă de graf a unei propoziții

În Figura 11 sunt prezentate opțiunile disponibile în secțiunea din partea dreaptă asociată reprezentării grafice. Acestea sunt:

- Căutare cuvânt în interfața specifică componentei scrise a corpusului CoRoLa : Korap
- Căutare lemma în WordNet
- Căutare pronunție în interfața specifică componentei vorbite a corpusului CoRoLa
- Sintetizare vorbire pornind de la un cuvânt
- Afișare cuvinte similare din punct de vedere al contextului de utilizare pe baza reprezentărilor vectoriale (”word embeddings”) antrenate utilizând corpusul CoRoLa.

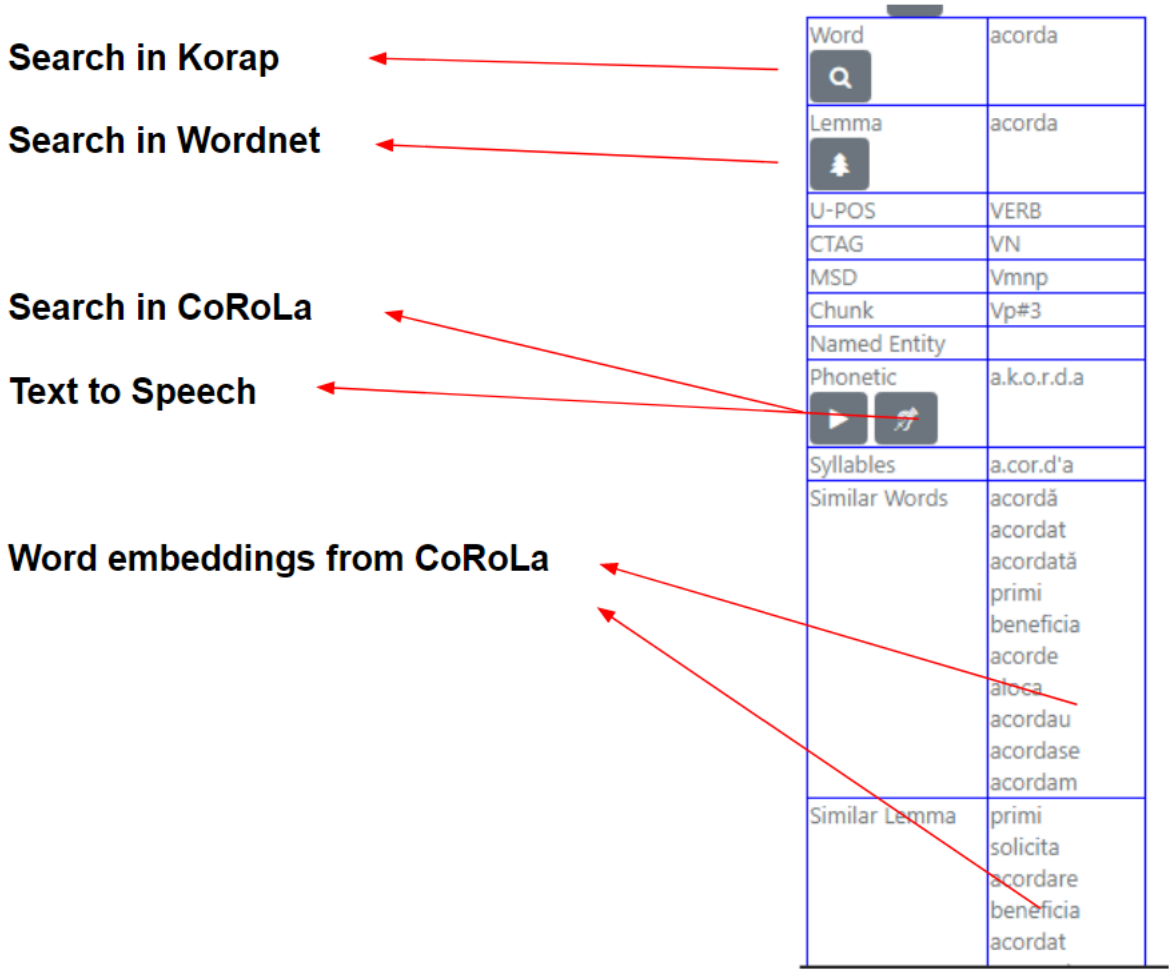

Figura 11. Legături către alte module integrate în platformă

Interfețele de căutare pot fi accesate și direct pe baza meniului principal al platformei disponibil în partea stângă în toate paginile. Cu toate acestea, dacă sunt accesate pe baza contextului de lucru, este posibilă revenirea utilizând butonul ”Back”, care permite utilizatorului să efectueze investigații suplimentare în contextul de lucru curent. Un astfel de exemplu este prezentat în Figura 12, pentru interfata de căutare în componenta scrisă din CoRoLa: Korap.

Figura 12. Interfata Korap integrată în platformă

Partea internă a platformei, dedicată procesării unui corpus de mari dimensiuni, utilizează un mecanism pe bază de ”task”-uri pentru a realiza prelucrările. Această structură este prezentată în Figura 13.

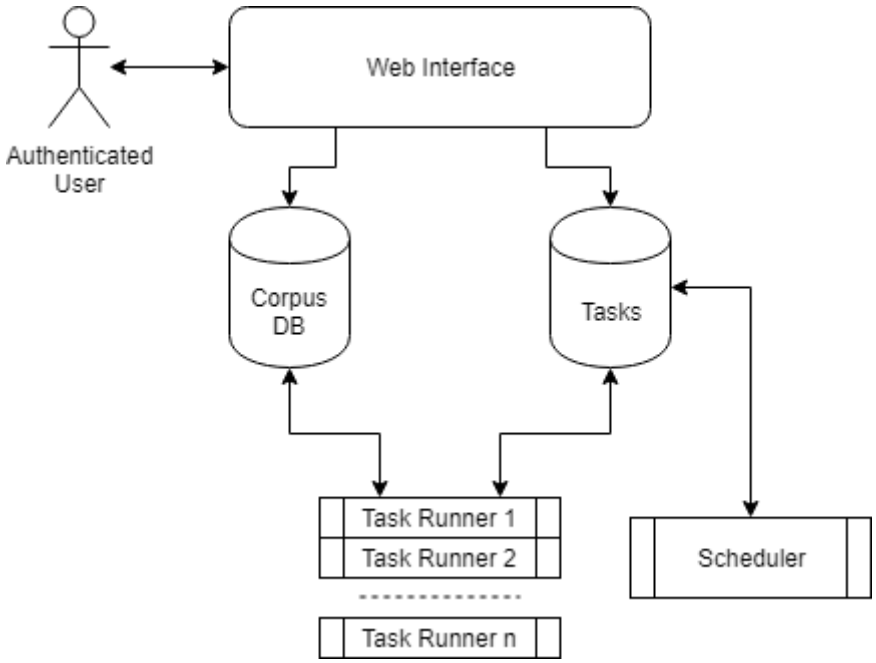

Figura 13. Mecanismul de paralelizare la nivelul platformei

Prin intermediul interfeței web, utilizatorul începe prin încărcarea fișierelor aferente unui corpus (Figura 14). Ulterior, are acces la interfața de management a task-urilor (Figura 15) prin care poate lansa în execuție un task de adnotare. Lansarea în execuție este gestionată de un proces de tip ”Scheduler” care permite alocarea diferitelor fișiere către diferite procese de execuție (”Task Runner”) disponibile pe serverul curent sau în rețea, în funcție de resursele configurate la nivelul platformei.

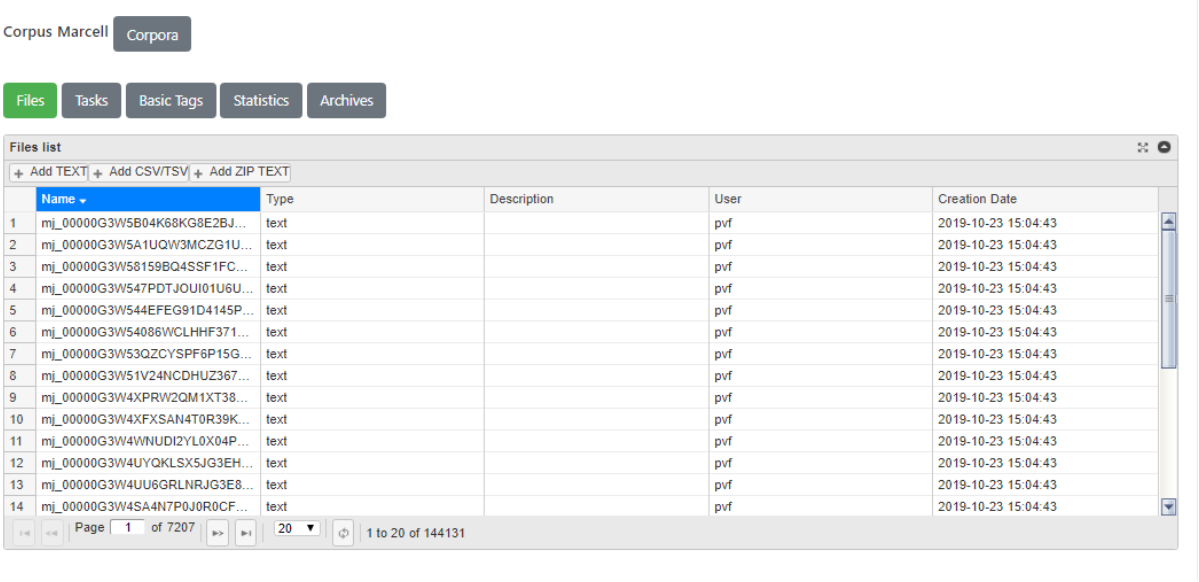

Figura 14. Încărcare fișiere în cadrul unui corpus

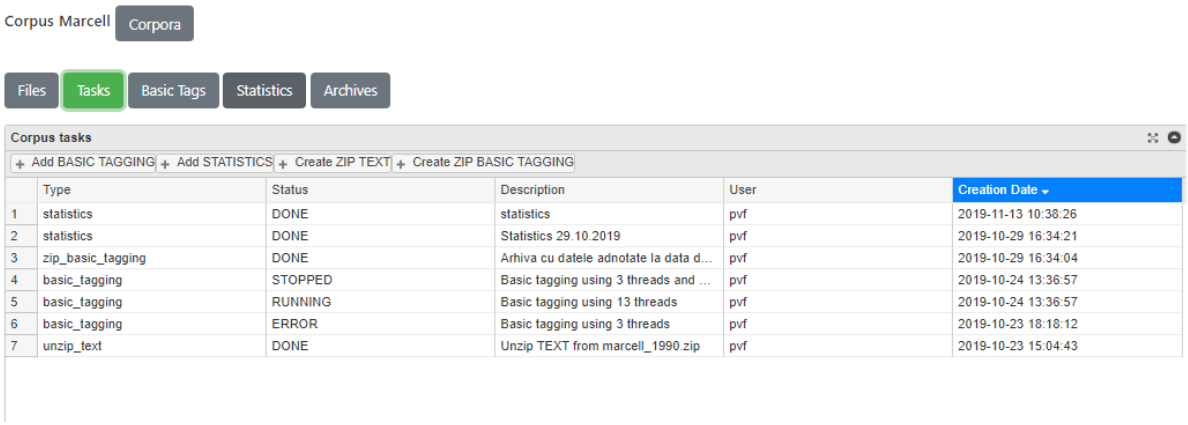

Figura 15. Interfața de management a taskurilor

Pe timpul execuției unui task, utilizatorul poate urmări starea acestuia, iar apoi poate accesa fișierele adnotate pe măsură ce acestea devin disponibile. Un task poate avea următoarele stări:

- NEW: task-ul a fost creat și așteaptă să fie preluat de procesul de alocare (”scheduling”)
- SCHEDULING: task-ul este în curs de alocare
- SCHEDULED: task-ul a fost alocat către procesele de execuție disponibile
- RUNNING: task-ul este în curs de execuție de către unul sau mai multe procese
- DONE: task-ul a fost finalizat cu succes
- ERROR: a intervenit o eroare in timpul execuției care a provocat oprirea permanentă a task-ului.

Pentru a permite lucrul cu volume mari de date, toate operațiunile în platformă, inclusiv dezarhivarea fișierelor la upload, arhivarea pentru download, crearea de statistici, sunt executate pe sistemul de task-uri.

Pe timpul execuției unui task, sau la finalul acestuia, rezultatele adnotării pot fi vizualizare sub forma unui tabel (Figura 16). De asemenea, este posibilă accesarea fiecărui fișier adnotat în mod individual pentru a vizualiza coloanele în formatul CoNLL-U Plus (Figura 17).

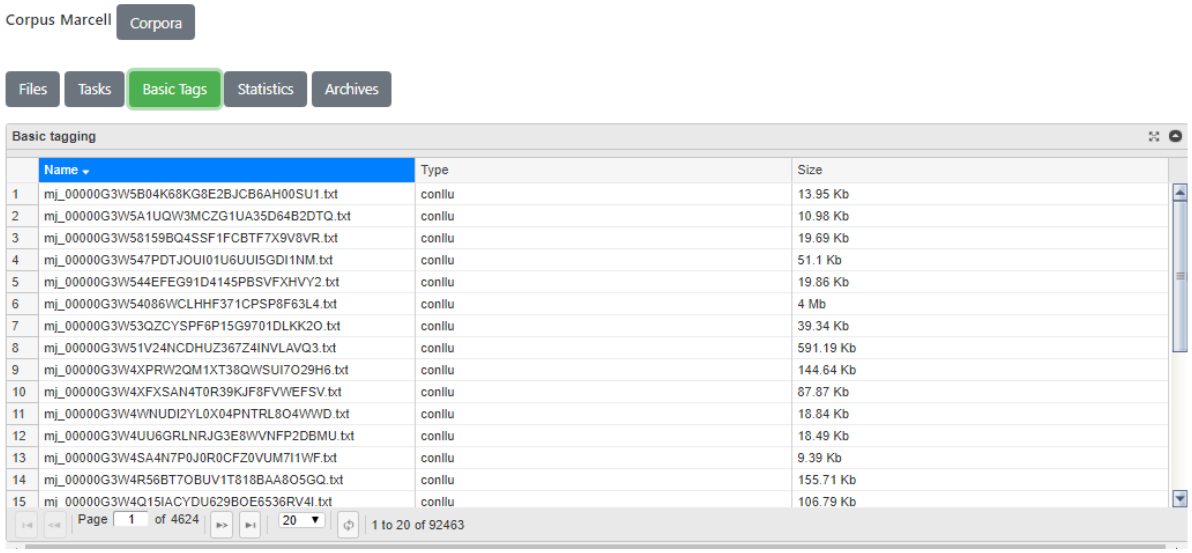

Figura 16. Lista de fișiere adnotate

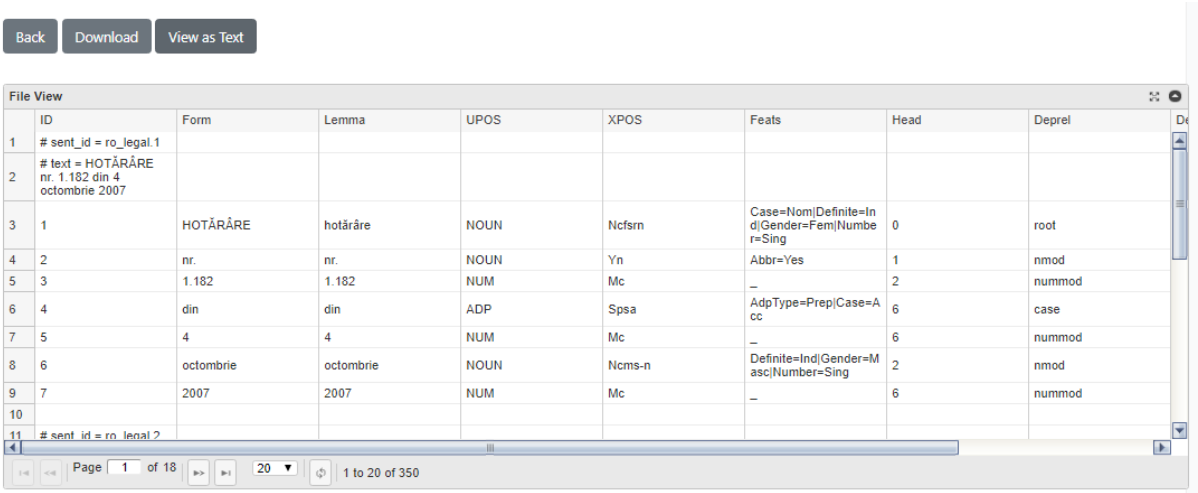

Figura 17. Vizualizare fișier adnotat în format CoNLL-U Plus

Vizualizarea unui fișier adnotat se poate realiza atât în format tabel cât și în format text. În cel de-al doilea caz, s-a avut în vedere păstrarea formatării cu tab-uri specifică formatului CoNLL-U. Totodată, în timpul vizualizării este posibilă și descărcarea fișierului în formatul CoNLL-U.

Pentru descărcarea unui volum mare de date, platforma oferă posibilitatea realizării unei arhive, fie cu datele neadnotate fie cu cele rezultate în urma procesului de adnotare. Arhivele generate pot fi apoi vizualizare și descărcate așa cum se vede in Figura 18.

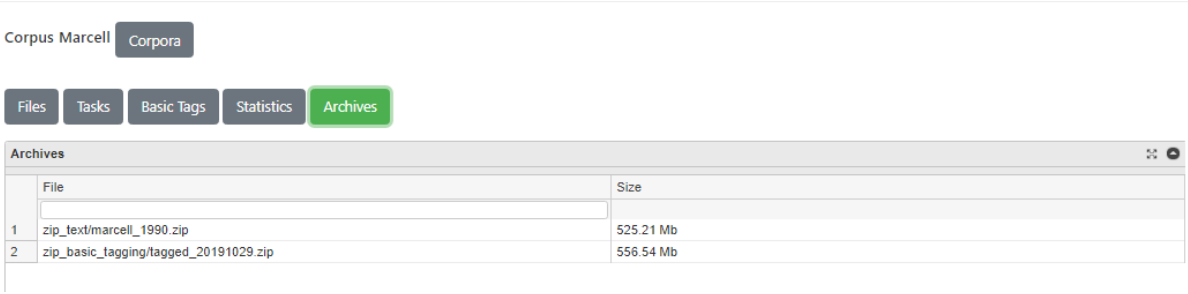

Figura 18. Vizualizare și descărcare arhive

Testare, evaluare și validare

În urma implementării, platforma a fost supusă unor teste în vederea validării îndeplinirii cerințelor funcționale precum și a evaluării performanțelor. Astfel, au fost încărcate diferite corpusuri cu dimensiuni de la 1 fișier până la un corpus cu peste 100.000 de fișiere. S-a confirmat astfel posibilitatea manipulării unor volume mari de date, ceea ce reprezenta o cerință fundamentală.

Fiecare cerință funcțională, așa cum a fost formulată în capitolul ”Cerințe funcționale”, a fost identificată în cadrul uneia sau mai multor pagini ale platformei.

Au fost create un număr de 7 conturi pentru membrii echipei de proiect implicați în lucrul cu corpus textual. Aceștia au accesat diferite elemente ale platformei, confirmând astfel ușurința în utilizare și conformitatea cu cerințele funcționale prezentate anterior.

De asemenea, a fost realizată o prezentare a platformei în cadrul Workshop-ului RETEROM organizat la Cluj pe 18 Noiembrie 2019. În cadrul acestei prezentări au fost expuse funcționalitățile disponibile către toți membri proiectului, fiind adunate observații în vederea unor îmbunătățiri ulterioare. Acestea nu au fost orientate spre probleme de bază, confirmând încă o dată variantele alese pentru implementare.

În vederea evaluării performanțelor sistemului de paralelizare, au fost efectuate teste cu un număr diferit de procese de execuție, rezultatele fiind raportate în Tabelul 1.

Procese	**Fișiere**	**Timp Scheduling (s)**	**Timp Execuție (s)**	**Timp total**	**Timp mediu / fișier (s)**
Public 1 proces	1	-	16,83	16,83	16,83
1 proces	10	0,01	177,69	178,7	17,87
2 procese	10	0,02	90,32	90,34	9,03
3 procese	10	0,02	69,2	69,24	6,92

Tabelul 1. Evaluarea performanțelor sistemului de paralelizare

În cazul utilizării a 10 fișiere, acestea au fost copii ale primului fișier, asigurând astfel o testare uniformă, independentă de conținutul propriu-zis al fișierului. De asemenea, procesele prezentate în Tabelul 1 au fost executate pe același sistem. Prezența proceselor pe același a sistem a fost aleasă pentru a garanta independența rezultatelor de aspecte precum viteza de transfer în rețea. Se observă o reducere proporțională a timpului de execuție cu mărirea numărului de procese. În cazul a două procese, timpul necesar este aproximativ jumătate, așa cum era de așteptat.

Separat, a fost realizat un test utilizând două servere fizice cu un total de 13 procese distribuite între acestea, care au adnotat timp de aproximativ 30 de zile peste 100.000 de fișiere. În acest caz platforma s-a comportat corespunzător. Pe tot parcursul testului, taskul era prezentat in stare ”RUNNING” iar fișierele puteau fi vizualizate pe măsură ce erau adnotate.

Diseminare

Platforma a fost prezentată în cadrul Workshop-ului RETEROM organizat la Cluj pe 18 Noiembrie 2019.

O descriere a procesului de integrare a diferitelor unelte în cadrul platformei a fost realizată în cadrul conferinței internaționale CONSILR – 14th International Conference on Linguistic Resources

and Tools for Natural Language Processing care s-a desfășurat la Cluj în perioada 18-20 Noiembrie 2019, fiind publicată lucrarea:

- 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

Bibliografie

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. Barbu Mititelu, D. Tufiș, E. Irimia (2018) The Reference Corpus of Contemporary Romanian Language (CoRoLa). In *Proceedings of the 11th Language Resources and Evaluation Conference – LREC'18*, Miyazaki, Japan, European Language Resources Association (ELRA).

Păiș, V., Tufiș, D. (2018) Computing distributed representations of words using the COROLA corpus. In *Proceedings of the Romanian Academy*, Series A, Volume 19, Number 2/2018, pp. 403–409.

Tufiș, D., and Mititelu, V.B. (2014) The Lexical Ontology for Romanian, pages 491–504. Springer.

P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, J. Tsujii, “Brat: a Web-based Tool for NLP-Assisted Text Annotation”, in Proceedings of the Demonstrations Session at EACL 2012.