Modele de inteligență artificială (deep learning) aplicate in analiza sintactico-semantică a limbii române vechi
DeLORo (Deep Learning for Old Romanian)
Cod proiect: PN-III-P2-2.1-PED-2019-3952

RAPORT ȘTIINȚIFIC ȘI TEHNIC
Etapa intermediară I
23 octombrie 2020 - 31 decembrie 2020

I. Rezumatul etapei

În prima etapă a proiectului au fost obținute următoarele rezultate:

- A fost identificată o listă de titluri de documente românești redactate în alfabet chirilic acoperind secolele XVI-XIX, adică perioada în care s-a utilizat acest alfabet pe teritoriul României de azi, aflate în format de imagini de pagini în posesia unor biblioteci din țară.
- A început procesul de dezvoltare a corpusului, prin depunerea temporară pe un server al IIT (ulterior transformat într-o platformă de lucru, după instalarea echipamentelor hard achiziționate în Etapa a II-a) a documentelor identificate.
- A fost elaborat standardul colecției de resurse (ROCC - Romanian Old Cyrillic Corpus), structura de metadate ale documentelor, structura imaginilor de pagini și ale obiectelor adnotate.
- A fost îmbunătățită interfața OOCIAT (*Online Old Cyrillic Image Annotation Tool*) de adnotare a resurselor (obiecte identificabile și conținutul lor lexical) și de completare metadate, proiectată înainte de începerea proiectului.
- S-a început procesul de adnotare a resurselor (obiecte și conținut lexical).
- A fost realizat site-ul proiectului și populat cu date.
- S-au elaborat mai multe lucrări științifice, rapoarte de doctorat cu legătură directă cu proiectul, și un proceedings de conferință.
- A fost organizat un eveniment științific: cea de-a 15-a conferință internațională din seria ConsILR.

Toate obiectivele acestei etape au fost realizate.

II. Descrierea științifică și tehnică

Descrierea care urmează respectă ordinea capitolelor din **Planul de realizare al proiectului**.

(punerea in evidenta a rezultatelor etapei si gradul de realizarea obiectivelor - se vor indica rezultatele si modul de diseminare a rezultatelor)

Act. 1.1 - Dezvoltarea corpusului - Tranșa I (partener: UAIC)
A. Identificarea resurselor (listă de titluri, deținători, drepturi de autor - posibil de lărgit ulterior) ce vor fi utilizate drept date de antrenare și validare pe parcursul primului an al proiectului (până la 22 octombrie 2021).

În această etapă au fost identificate o colecție de documente chirilice românești, tipărituri și semiunciale, din perioada secolelor XVI-XIX[1]. Principalii furnizori de resurse au fost: Biblioteca Academiei Române din București, Biblioteca Filialei Iași a Academiei Române și diverse teze de doctorat cu subiecte din paleolingvistică, dezvoltate în Școala Doctorală a Facultății de Litere din Universitatea "Alexandru Ioan Cuza" din Iași. Lista acestor resurse[2] conține titlurile, clasificate în tipărituri și semiunciale, numărul de pagini, anul producerii documentului, nivelul de zgomot vizibil în pagini (informație ce este marcată în metadate), dacă sunt ori nu transcrise, precum și alte informații.

B. Începerea procesului de dezvoltare a corpusului conform listei de resurse identificate (scanuri de pagini) prin depunerea lor în site-ul proiectului.

Deoarece datele implicate în DeLORo sunt de natură eterogenă și trebuie să corespundă diferitelor tipuri de activități (încărcare de imagini, CRUD[3] asupra adnotărilor efectuate manual, exportul de date către modelele de instruire DL etc.), am adoptat o abordare hibridă pentru stocarea de date și anume: păstrarea imaginilor și fișierelor text în sistemul de fișiere al serverului dedicat DeLORo și a restului datelor - într-o bază de date PostgreSQL. Ori de câte ori datele trebuie adaptate pentru un anumit caz de utilizare, ele sunt trecute printr-un pipeline de transformare care generează formatul necesar.

În DeLORo se disting trei tipuri principale de date: (i) date importate, (ii) date de bază și (iii) date exportate. Le descriem mai jos.

Date importate

Datele importate sunt cele care au fost create înaintea începerii proiectului și care, fie prin copierea ca atare în baza de date, fie prin procesare suplimentară, sunt transformate în date de bază. Ele constau din trei categorii principale: (i) imagini scanate ale paginilor, (ii) metadate și (iii) texte transcrise. Din aceste trei categorii, doar textele transcrise sunt copiate așa cum sunt în datele de bază ale proiectului, celorlalte două aplicându-li-se prelucrări suplimentare. Fiecare carte importată este fie un fișier PDF care conține imagini scanate ale paginilor, fie un director care conține imagini ale paginilor în format JPEG. În plus, pentru fiecare carte, un fișier XML suplimentar păstrează metadatele corespunzătoare (titlu, autori, anul de creație, informații despre publicare etc.). Canalul de import iterează fiecare director corespunzător unei anumite cărți, aplicând următoarele acțiuni: (a) replică directorul în datele de bază; (b) rulează instrumentul de import de metadate, care iterează peste fișierele de metadate și actualizează valorile metadatelor în datele de bază; (c) dacă directorul importat conține imagini, imaginile sunt convertite în format PNG, apoi imaginile sunt redenumite, pentru a asigura ordonarea lexicografică a paginilor, și copiate în datele de bază; (d) dacă directorul importat conține un singur fișier PDF, care concentrează toate imaginile paginilor, atunci imaginile sunt extrase în directorul de date de bază ca fișiere separate în format PNG, cu nume atribuite pentru a fi replicate lexicografic în ordinea paginilor din documentele originare.

1. Standardul intern al corpusului ROCC

[1] Pentru că proiectul DeLORo nu include o activitate de scanare a originalelor, ne-am concentrat atenția asupra unor surse aflate în posesia unor mari biblioteci, care dețin și copiile scanate ale paginilor.

[2] Aflată la adresa: https://docs.google.com/spreadsheets/d/1hv_jKwLzK8ZFiVsO8lA0L2I7nvXSAT7QOUdgxeJCIcs/edit?usp=sharing

[3] Acronimul rezumă cele 4 tipuri de operațiuni la baza de date: Create, Read, Update și Delete.

Ce urmează reprezintă propunerea de standard pentru colecțiile de date care vor popula platforma DeLORo. Acest standard este necesar pentru uniformizarea contribuțiilor membrilor consorțiului, precum și în vederea proiectării tehnologiilor de aliniere imagine_pagină – text, de identificare a obiectelor din imagine, de recunoaștere a caracterelor chirilice și de transcriere a lor în alfabetul latin. Standardul elaborat de noi se depărtează de standardul TEI-P5[4], pentru că nu am regăsit acolo multe elemente comune cu scopul nostru. Descrierea standardului urmărește convențiile XML, deși în implementare s-a adoptat un format care să faciliteze operațiile asupra bazei de date (SQL).

Corpusul ROCC este format dintr-o mulțime de documente (colecții de pagini), fiecare colecție de acest gen având în componența sa:

- o secțiune care descrie metadatele colecției: <pageCollectionMetadata/>;
- o secțiune dedicată imaginilor de pagină: <imagesOfPages/>;
- o secțiune dedicată segmentărilor făcute deopotrivă de adnotatori și de tehnologie asupra paginilor: <segmentationOfImages/>;
- o secțiune dedicată textelor în alfabet latin care decodifică înscrisul chirilic: <textsOfPages/>;
- o secțiune dedicată alinierilor între imagini și text; propunem două variante, una în care alinierile sunt în perechi slovă-la-literă, <charAlignmentsImage2Text/>, și una în care ele sunt prezentate ca perechi de seventențe de slove la secvențe de litere, <seqAlignmentsImage2Text/>.

În conformitate cu această structurare, o pagină din conținutul unui document se înregistrează în ROCC prin următoarele componente:

- metadatele atașate paginii (obligatoriu),
- imaginea propriu-zisă (obligatoriu),
- segmentările imaginii (opțional),
- textul transcris (opțional),
- alinierile imagine-text (opțional).

2. Metadatele atașate unei colecții de pagini: elementul <pageCollection>

La nivelul cel mai de sus al corpusului ROCC se plasează elementele <pageCollection>. O astfel de colecție de pagini caracterizează un document, adică o carte, un număr de ziar, în general o unitate editorială ce are asociat un cod de identificare de bibliotecă. La nivelul metadatelor, această descriere combină informații preluate din site-ul bibliotecii, cu altele validate și completate apoi în discuțiile purtate între membrii consorțiului DeLORo.

Este în interesul proiectului ca imaginile scanate ale paginilor chirilice să fie prezentate în corpus în pereche cu transcrierile lor textuale romane. Unitatea de bază în alinierea imagine-text este pagina de carte, dar subîntinderi ale ei vor fi de asemenea considerate: coloana, rândul, cuvântul, slova chirilică etc.

Descriem în cele ce urmează propunerea de standard a corpusului ROCC:

<ROCC>
<pageCollection>

***Notă**: Secțiunea* <pageCollectionMetadata/> *grupează descrieri generale despre colecție.*

---
[4] http://www.tei-c.org/

<pageCollectionMetadata>

@ROCC_Id (obligatoriu): identificatorul unic al documentului în corpusul ROCC;

@pageCollectionURL (obligatoriu): URL-ul din server al colecției de pagini (imagini); (exemplu: https://deloro.iit.academiaromana-is.ro/rocc/surse/sec-xix-1/tipar/crv-956a-beldiman-tragodia-lui-orest/)

@title (obligatoriu): titlul documentului (carte, ziar etc.);

@shortTitle (opțional): dacă documentul este cunoscut și sub un nume prescurtat (acest atribut corespunde câmpului UniTitle din codificarea Alef);

@language (obligatoriu): limba sau limbile folosite în document: “ro”, “sl” etc.;

@metadataCreator (obligatoriu): entitatea, una sau mai multe (persoane, instituții) care a editat metadatele colecției curente;

@distribution (obligatoriu): regimul de distribuție al acestei colecției ROCC, valorile fiind tipuri de licențe (ex. https://creativecommons.org/licenses/by-nc/3.0/).

<ROCC-code>

@difficultyLevel (obligatoriu): codifică nivelul global de dificultate al colecției de pagini, cu valorile: “1” = ușor, “2” = mediu, “3” = dificil.

***Notă**: Acest indicator trebuie inițial completat manual (printr-o inspecție globală făcută de curatorul metadatelor asupra documentului), iar ulterior actualizat automat (printr-o formulă de medie) calculată de interfața OOCIAT (the Online Old Cyrillic Image Annotation Tool) la fiecare adnotare a unei noi pagini;*

@writingType (obligatoriu): completat manual, codifică tipul de scris, cu valorile: “p” = tipăritură (print), “u” = manuscris uncial, “su” = semiuncial, “m” = manuscris cu ligaturi între litere.

@annotationLevel (obligatoriu): codifică nivelul de adnotare, cu valorile: “o” = original, neadnotat; “g” = (gold) adnotat (parțial sau în totalitate) de experți; “t” = (test) adnotat de mașină; “m” = (mixt) adnotat atât manual cât și de mașină (obiecte și transcriere). La includerea în ROCC, acest indicator are valoarea implicită “o”, apoi va fi modificat automat.

@century (obligatoriu), cu valorile: “XVI”, “XVII”, “XVIII” and “XIX”; acest atribut este completat automat prin interpretarea valorii din @**printingYear**.

@50years (obligatoriu): cu valorile: “1”, “2”, însemnând prima sau a doua jumătate a secolului; idem;

***Notă**: Valorile atributelor @century și @50years se generează automat din anul publicării.*

***Notă**: Perioadele sunt următoarele: XVI-1: 1500–1549; XVI-2: 1550–1599; XVII-1: 1600 - 1649; XVII-2: 1650 - 1699; XVIII-1: 1700 - 1749; XVIII-2: 1750 - 1799; XIX-1: 1800 - 1849; XIX-2: 1850 - 1899*

@zone (obligatoriu), cu valorile: “MD” = Moldova (pentru secolele XVI - XIX - 1812), “W” = Țara Românească, “T” = Transilvania, “MM” = Maramureș, “BT” = Banat, “BS” = Basarabia (pentru perioada 1812 - 1899), sau orice combinație a lor.

***Notă**: Acest indicator este preluat automat din câmpul @creationProvince.*

</ROCC-code>

***Notă**: Fiecare document, fiind o secvență de pagini, va fi luat în considerare drept o "colecție", ea având în constituția sa fie doar o secvență de imagini de pagini (când codificarea documentului ca membru al colecției ROCC este <ROCC-code … transcription="0"/>, fie o secvență de înregistrări <imagine>-<text>, unde <imagine> reprezintă reproducerea paginii originale în alfabet chirilic, iar <text> - transcrierea ei în alfabet latin (când, codificarea documentului ca membru al colecției ROCC este <ROCC-code … transcription="1"/>.*

<translation> (opțional)
@originalLanguage (obligatoriu): limba sursei directe a traducerii în română, nu limba originală a operei;
@originalAuthor (opțional):
@translator (opțional): traducător în română;
@secTranslator (opțional): alți traducători, dacă e cazul;
</translation>

***Notă**: Cel puțin informația din câmpul @translator se regăsește și în <creation> => <secAuthor> => @authority. Se va hotărî mai târziu dacă elementul <translation> trebuie să rămână.*

<scannedCopy>[5] dacă ROCC păstrează mai multe variante de copii scanate (sau fotografiate), fiecare este însoțită de metadate <pageCollection> proprii, cu toate că unele câmpuri vor deveni astfel redundante; un element <scanned-copy> va include atributele:
@library (obligatoriu): numele sau codul bibliotecii care deține originalul și de unde provine copia scanată; implicit BAR;
@libraryCode (obligatoriu): cota documentului în bibliotecă;
</scannedCopy>
<creation> (obligatoriu): conține informații relative la crearea documentului în limba română, cel care a constituit sursa scanului:
@creationYear (opțional): anul în care a fost concepută opera, dacă nu se cunoaște, se completează "unknown";
<author> (opțional): autorul sau traducătorul cu responsabilitate principală;
@content (opțional): valoarea transcrisă din fișă;
@surname (opțional): numele de familie al autorului;
@name (opțional): numele de botez al autorului/traducătorului, dacă e cunoscut; dacă nu, se completează "unknown";
<addToName> (opțional): adăugiri la nume;
<bio> (opțional): alte date bio.
</author>
<secAuthor> (opțional): alți autori/traducători, dacă sunt, în afara celui principal; câte o înregistrare pentru fiecare autor secundar;
@id (obligatoriu): pentru că pot exista mai mulți autori;
@surname (opțional): numele de familie al autorului;
@name (opțional): numele de botez al autorului/traducătorului, dacă e cunoscut; dacă nu, se completează "unknown";

[5] Am numit această secțiune "scannedCopy", deși nu excludem posibilitatea ca paginile unor documente să fie fotografiate ori sub formă de microfișe, nu numai scanate.

@authority (opțional): autoritatea autorului secundar (ex: copist, ilustrator, editor, traducător etc.)

<addToName> (opțional): adăugiri la nume; valoarea transcrisă din fișă apare drept conținut;

@id (obligatoriu): pentru că pot exista mai mulți autori;

</addToName>

<bio> (opțional): alte date bio.

@**id** (obligatoriu): pentru că pot exista mai mulți autori;

</bio>

</secAuthor>

<creationLocation> (opțional): locul unde a fost concepută opera, adică originalul sau traducerea, dacă e cunoscut, cu detaliile:

@creationProvince (opțional), cu una din valorile “Moldova”, “Țara Românească”, “Transilvania”, sau o combinație a lor, în ordine cronologică;

@creationTown (opțional): una sau mai multe valori, în ordine cronologică;

</creationLocation>

</creation>

<publishing>

@publishingYear (obligatoriu): anul în care a fost publicat documentul (anul producerii copiei sau al tipăririi);

@publisher: persoana care a editat/finanțat/comandat cartea;

@noOfPagesOrSheets: număr pagini ale documentului;

@pageOrSheet: “page”/”sheet”

@bookFormat: cu valorile “duo”, “quarto”, “octavo”, “folio”;

@noOfLinesPerPage: numărul de rânduri, dacă sunt precizate în metadatele cărții;

<dimensions>: dimensiunile paginii (în mm):

@contentWidth: oglinda paginii, lățime;

@contentHight: oglinda paginii, înălțime;

@**pageWidth**: dimensiunea exterioară a paginii, lățime;

@pageHeight: dimensiunea exterioară a paginii, înălțime:

</dimensions>

<publishingLocation> (opțional): locul unde a fost tipărit sau copiat documentul, dacă e cunoscut, cu detaliile:

@publishingProvince (opțional), cu una din valorile “Moldova”, “Țara Românească”, “Transilvania”, “Maramureș”, “Banat”, “Basarabia”;

@publishingTown (opțional);

<publishingHouse> (optional): editura/tipografia sau locul unde a fost copiat manuscrisul (de ex., o mănăstire);

@id (obligatoriu): când există mai multe edituri;

</publishingHouse>

<publishingCountry> (opțional);

@id (obligatoriu): când există mai multe locații;

</publishingCountry>

</publishingLocation>

</publishing>
<contentDescription> (obligatoriu): o descriere liberă a conținutului operei, cu atributele:
@style (opțional), cu valorile: ”Bis.” (=Bisericesc); ”Jur.” (=Juridic); ”Șt.” (=Științific); ”Publ.” (=Publicistic); ”Beletr.” (=Beletristic), cât și combinații.
<subject> (opțional), de exemplu: carte de cântece bisericești, manual etc.
</contentDescription>
<formatDescription> (opțional): o descriere liberă a formatului operei; poate apare de mai multe ori;
<content>
@id (obligatoriu);
</content>
</formatDescription>
</pageCollectionMetadata>

3. Organizarea colecției imaginilor de pagini adnotate

<imagesOfPages> (obligatoriu): conține secvența de imagini de pagini ale documentului, în care fiecare pagină are structura de mai jos:
<onePageImage> (câte o astfel de înregistrare pentru fiecare imagine de pagină din document), fiecare conținând două atribute și o structură de dificultate de procesare:
@pageID (obligatoriu): ID-ul unic al imaginii de pagină din cadrul colecției (ID automat generat de Platformă la încărcarea secvenței de imagini de pagini a documentului);
@pageName: numele fișierului care conține pagina din directorul documentului (exemplu: tragodia-lui-orest-2-pagina-39.png)
@pageImageFile: URL-ul imaginii scanate a paginii;
<difficultyCriteria> (obligatoriu): element completat manual de adnotator după inspecția vizuală a paginii, cu atributele:
@damaged (obligatoriu), cu valorile: “true”: pagina are porțiuni lipsă; ”false”: pagina este integră.
@opaqueSpots (obligatoriu), cu valorile: “true”: în pagină sunt vizibile pete care obturează scrisul (cel puțin un caracter nu poate fi citit); ”false”: nu e cazul.
@transparentPaper (obligatoriu), cu valorile: “true”: prin transparență apar vizibile caractere de pe verso; ”false”: nu e cazul.
@overlayPrint (obligatoriu pentru tipărituri), cu valorile: “true”: tipar suprapus, de exemplu roșul e deplasat față de negru, cu acoperire parțială; ”false”: nu e cazul.
@interlineWriting (obligatoriu), cu valorile: “true”: există scris de mână între rânduri, abrevieri și semne speciale; ”false”: nu e cazul.
@palimpsest (obligatoriu), cu valorile: “true”: hârtia a mai fost folosită o dată, apoi a fost ștearsă și s-a scris din nou pe ea; ”false”: nu e cazul.
@corrections (obligatoriu), cu valorile: “many”: sunt mai mult de 5 corecturi în pagină; ”few”: între 1 și 5 corecturi în pagină; “none”: nicio corectură în pagină.

@marginalWriting (obligatoriu), cu valorile: “many”: sunt mai mult de 5 note marginale; ”few”: între 1 și 5 note marginale; “none”: nu există nicio notă marginală în pagină.

</difficultyCriteria>

</onePageImage>

</imagesOfPages>

4. Organizarea colecției de obiecte adnotate în pagini

Notă*: Această secțiune descrie obiectele generate de interfața OOCIAT și/sau de mașină în urma operațiilor de segmentare a imaginilor de pagini. Așa cum se va vedea, diferența dintre marcarea manuală a unui obiect într-o pagină (făcută de unul sau chiar de mai mulți adnotatori) și cea efectuată automat este dată de valoarea câmpului @objectAnnotator, care poate fi un ID de adnotator uman sau “AUTOMATIC”. Existența acestor două tipuri de valori, una caracterizând segmentările făcute manual (gold), cealaltă - pe cele făcute automat (test), va face posibilă evaluarea, prin comparare, a tehnologiei de identificare de obiecte, la nivel de tip de obiect.*

<segmentationOfImages> (opțional, doar dacă asupra documentului s-au operat segmentări), cu structura:

<pageSegmentation> (se repetă pentru fiecare dintre paginile asupra cărora s-au efectuat operațiuni de segmentare manuală), cu structura:

@**pageID** (obligatoriu): ID-ul imaginii de pagină corespunzătoare din cadrul colecției, care se găsește aici: <pageCollection> ⇒ <imagesOfPages> ⇒ <onePageImage> ⇒ @pageID;

Notă*: În continuare este descrisă întreaga gamă de obiecte ce se pot afla pe o pagină.*

<graphicalObject:Frontispiece> (opțional): descrie un obiect desenat sau imprimat de tip frontispiciu, plasat pe pagina de gardă documentului (carte, ziar), cu atributele:

@objectId (oblig.): v. <pageCollection> ⇒ <imagesOfPages> ⇒ <onePageImage> ⇒ @pageID.

@objectAnnotator: user-name-ul adnotatorului obiectului, sau “AUTOMATIC” dacă obiectul e adnotat de mașină;

<objectCoordinates> (obligatoriu), cu atributele:

@leftUpHoriz (obligatoriu): coordonata orizontală a colțului stânga sus al obiectului;

@leftUpVert (obligatoriu): coordonata verticală a colțului stânga sus al obiectului;

@rightDownHoriz (obligatoriu): coordonata orizontală a colțului dreapta jos al obiectului;

@rightDownVert (obligatoriu): coordonata verticală a colțului dreapta jos al obiectului;

</objectCoordinates>

</graphicalObject:Frontispiece>

<object:Title> (opțional): descrie un obiect de tip titlu, scris mare, în susul unei pagini, cu atributele:

@objectId (obligatoriu): ca la **<graphicalObject:Frontispiece>**;
@objectAnnotator: (oblig.) v. **<graphicalObject:Frontispiece>**;
@objectContent (opțional): conține un șir de caractere cu transcrierea titlului în alfabet latin.
<objectCoordinates/> (oblig.) v. **<graphicalObject:Frontispiece>**
</object:Title>

<object:Column> (o pagină poate conține obligator una, maximum două elemente <objectColumn>): descrie un obiect de tip coloană, cu atributele:

@objectId (obligatoriu): ca la **<graphicalObject:Frontispiece>**;
@objectAnnotator: (oblig.) v. **<graphicalObject:Frontispiece>**;
@columnPosition (obligatoriu), cu valorile: "U" = coloană unică în pagină; "L" = coloana din stânga; "R" = coloana din dreapta;
<objectCoordinates/> (oblig.) v. **<graphicalObject:Frontispiece>**
</object:Column>

<object:Line> (opțional): descrie un obiect de tip linie, cu atributele:

@objectId (obligatoriu): ca la **<graphicalObject:Frontispiece>**;
@objectAnnotator: (oblig.) v. **<graphicalObject:Frontispiece>**;
@objectContent (opțional): conține un șir de caractere cu transcrierea conținutului obiectului în alfabet latin.
@inColumn (opțional): cu una din valorile: "header" = dacă linia e deasupra coloanelor, în capul paginii; "footer" = dacă linia e dedesubtul coloanelor, în josul paginii; "ordinary" = dacă linia face parte dintr-o coloană a paginii.

***Notă**: Nu se indică o poziționare a liniilor în cadrul unei coloane, pentru că acest lucru ar complica procesul de adnotare. Propunerea de mai sus ar nota doar cazurile liniilor care nu se încadrează în coloane (am putea renunța la valoarea implicită "ordinary").*

<objectCoordinates/> (oblig.) v. **<graphicalObject:Frontispiece>**
</object:Line>

***Notă**: Nu se indică o poziționare a liniilor una față de alta. O astfel de poziționare ar presupune o adnotare în secvență a lor, probabil de sus în jos, astfel încât linia de dedesubt să aibă o referință către cea de deasupra, adnotată imediat anterior, ceea ce constrânge inutil procesul de adnotare. Referințe de acest gen ar fi fost utile pentru reconstituirea secvenței de text în pagină, dar secvențierea textului poate fi rezolvată și dacă nu există o definire explicită a secvenței liniilor, prin considerații geometrice de poziționare a dreptunghiurilor ce le încadrează. Observația e valabilă pentru mai multe tipuri de obiecte a căror poziționare relativă e importantă în reconstituirea secvențialității textului. Lăsăm această operație de secvențiere pe seama unor algoritmi care, din interpretarea coordonatelor de poziționare a obiectelor, să fie capabili să recupereze secvențe textuale.*

<object:Character> (opțional): descrie un obiect de tip literă sau număr (slovă), cu atributele:

@objectId (obligatoriu): ca la **<graphicalObject:Frontispiece>**;

@objectAnnotator: (oblig.) v. **<graphicalObject:Frontispiece>**;
@objectAccuracy: în cazul în care obiectul are @objectAnnotator = "AUTOMATIC", acest atribut notează gradul de încredere (acuratețea) a recunoașterii lui, obținută în urma ultimei evaluări;
@objectContent (opțional): atunci când apare, poate avea valorile: unul sau două caractere: transcrierea literei chirilice în alfabet latin; "VOID" = dacă litera nu se transcrie.
<objectCoordinates/> (oblig.) v. **<graphicalObject:Frontispiece>**
</object:Character>
<object:Marginal> (opțional): descrie un obiect de tip zonă de text pe manșetă, cu atributele:
@objectId (obligatoriu): ca la **<graphicalObject:Frontispiece>**;
@objectAnnotator: (oblig.) v. **<graphicalObject:Frontispiece>**;
@objectContent (opțional): conține un șir de caractere cu transcrierea zonei de text marginal în alfabet latin.
<objectCoordinates/> (oblig.) v. **<graphicalObject:Frontispiece>**
</object:Marginal>
<object:OutOfLineCharacters> (opțional): descrie un obiect care conține o literă sau o secvență de litere plasată/e în afara unui rând (deasupra sau dedesubt), cu atributele:
@objectId (obligatoriu): ca la **<graphicalObject:Frontispiece>**;
@objectAnnotator: (oblig.) v. **<graphicalObject:Frontispiece>**;
@objectContent (opțional): transcrierea literei sau secvenței în alfabet latin.
<objectCoordinates/> (oblig.) v. **<graphicalObject:Frontispiece>**
</object:outOfLineCharacters>
<object:Modifier> (opțional): descrie un obiect de tip modificator plasat deasupra slovelor; acestea pot fi: accent ascuțit (ʹ), accent grav (ʻ), titlă (notată la valoare ca tilda: ~) și pot semnala abrevieri, schimbarea semnificației caracterului din literă în număr etc., cu atributele:
@objectId (obligatoriu): ca la **<graphicalObject:Frontispiece>**;
@objectAnnotator: (oblig.) v. **<graphicalObject:Frontispiece>**;
@objectContent (opț.): se notează, după caz, "ʹ", "ʻ" sau "~".
<objectCoordinates/> (oblig.) v. **<graphicalObject:Frontispiece>**
</object:Modifier>
<object:InitialLetter> (opțional): descrie un obiect de tip literă ornată la început de paragraf, cu atributele:
@objectId (obligatoriu): ca la **<graphicalObject:Frontispiece>**;
@objectAnnotator: (oblig.) v. **<graphicalObject:Frontispiece>**;
@objectContent (opț.): echivalentul slovei ornate în alfabet latin.
<objectCoordinates/> (oblig.) v. **<graphicalObject:Frontispiece>**
</object:InitialLetter>
<object:ReferenceMarkOnMargin> (opțional): descrie un obiect de tip trimitere (semn cu valoare de reper, vrahie etc.), plasat pe manșetă, cu atributele:
@objectId (obligatoriu): ca la **<graphicalObject:Frontispiece>**;
@objectAnnotator: (oblig.) v. **<graphicalObject:Frontispiece>**;
@objectContent (opțional): o codificare a semnului.

<objectCoordinates/> (oblig.) v. **<graphicalObject:Frontispiece>**
</object:ReferenceMarkOnMargin>
<object:ReferenceMarkAboveLine> (opțional): descrie un obiect de tip trimitere (semn cu valoare de reper), plasat deasupra unei linii, cu atributele:
@objectId (obligatoriu): ca la **<graphicalObject:Frontispiece>**;
@objectAnnotator: (oblig.) v. **<graphicalObject:Frontispiece>**;
@objectContent (opțional): o codificare a semnului.
<objectCoordinates/> (oblig.) v. **<graphicalObject:Frontispiece>**
</object:ReferenceMarkAboveLine>
<graphicalObject:Accolade> (opțional): descrie un obiect desenat de tip acoladă, plasat pe margine, cu atributele:
@objectId (obligatoriu): ca la **<graphicalObject:Frontispiece>**;
@objectAnnotator: (oblig.) v. **<graphicalObject:Frontispiece>**;
<objectCoordinates> (obligatoriu), cu atributele:
@leftUpHoriz (obligatoriu): coordonata orizontală a colțului stânga sus al chenarului care încadrează acolada;
@leftUpVert (obligatoriu): coordonata verticală a colțului stânga sus al chenarului care încadrează acolada;
@rightDownHoriz (obligatoriu): coordonata orizontală a colțului dreapta jos al chenarului care încadrează acolada;
@rightDownVert (obligatoriu): coordonata verticală a colțului dreapta jos al chenarului care încadrează acolada;
@direction (obligatoriu): descrie direcția deschiderii, cu valorile: "directionRight" = vârful acoladei este direcționat spre dreapta (de obicei, pentru acolade plasate pe manșeta stângă); "directionLeft" = vârful acoladei este direcționat spre stânga (de obicei, pentru acolade plasate pe manșeta dreaptă;
@horizCoordOfPeak (obligatoriu): coordonata orizontală a vârfului acoladei;
@vertCoordOfPeak (obligatoriu): coordonata verticală a vârfului acoladei;
</objectCoordinates>
</graphicalObject:Accolade>
<graphicalObject:Ornament> (opțional): descrie un obiect desenat sau imprimat de tip ornament, gravură, plasat fie pe margine, fie în partea de sus sau de jos a paginii, cu atributele:
@objectId (obligatoriu): ca la **<graphicalObject:Frontispiece>**;
@objectAnnotator: (oblig.) v. **<graphicalObject:Frontispiece>**;
<objectCoordinates/> (oblig.) v. **<graphicalObject:Frontispiece>**
</graphicalObject:Ornament>
<graphicalObject:Frame> (opțional): descrie un obiect desenat sau imprimat de tip chenar, ancadrament, de regulă plasat în jurul unei zone care conține text, cu atributele:
@objectId (obligatoriu): ca la **<graphicalObject:Frontispiece>**;
@objectAnnotator: (oblig.) v. **<graphicalObject:Frontispiece>**;
<objectCoordinates> (obligatoriu), cu atributele:

@extFrameLeftUpHoriz (obligatoriu): coordonata orizontală a colțului stânga sus al dreptunghiului exterior al chenarului;
@extFrameLeftUpVert (obligatoriu): coordonata verticală a colțului stânga sus al dreptunghiului exterior al chenarului;
@extFrameRightDownHoriz (obligatoriu): coordonata orizontală a colțului dreapta jos al dreptunghiului exterior al chenarului;
@extFrameRightDownVert (obligatoriu): coordonata verticală a colțului dreapta jos al dreptunghiului exterior al chenarului;
@intFrameLeftUpHoriz (obligatoriu): coordonata orizontală a colțului stânga sus al dreptunghiului interior al chenarului;
@intFrameLeftUpVert (obligatoriu): coordonata verticală a colțului stânga sus al dreptunghiului interior al chenarului;
@intFrameRightDownHoriz (obligatoriu): coordonata orizontală a colțului dreapta jos al dreptunghiului interior al chenarului;
@intFrameRightDownVert (obligatoriu): coordonata verticală a colțului dreapta jos al dreptunghiului interior al chenarului;
</objectCoordinates>
</graphicalObject:Frame>
</pageSegmentation>
</segmentationOfImages>
</pageCollection>
</ROCC>

C. Îmbunătățirea interfeței de adnotare a resurselor (obiecte identificabile și conținutul lor lexical) și de completare metadate

Interfața de editare interactivă online OOCIAT (*the Online Old Cyrillic Image Annotation Tool*), a cărui design a fost realizat înainte de începerea proiectului, a început să fie implementat în această etapă. Tehnologia lui de realizare este descrisă mai jos.

Aplicația OOCIAT este creată folosind tehnologii precum HTML, CSS, JavaScript, PHP și baze de date MySQL. Pentru modelarea mai ușoară a elementelor din pagină s-a folosit *framework*-ul *Bootstrap*, care pune la dispoziție o serie de clase CSS care ajută la alinierea și stilizarea elementelor paginii. Pentru implementarea *backend*-ului s-a folosit limbajul PHP. Acesta oferă o modalitate ușoară de comunicare atât cu elementele din *frontend* cât și cu baza de date. Avantajele folosirii acestui limbaj sunt reprezentate de multitudinea de funcții care fac ușoară comunicarea la nivel de *request* cu utilizatorul, cât și intuitivitatea folosirii acestuia.

Baza de date este de tip MySQL. Acesta oferă o sintaxă ușoară de interogare a tabelelor și este des utilizată în combinație cu PHP care oferă o integrare facilă cu MySQL.

Workspace-ul dedicat încărcării și adnotării paginilor a fost dezvoltat în care măsură cu ajutorul limbajului JavaScript. Posibilitatea de desenare a dreptunghiurilor, încărcarea dinamică a adnotărilor deja făcute, ștergerea și modificarea acestora, redimensionare paginilor etc. sunt realizate folosind tehnologii JavaScript.

D. Începerea procesului de adnotare a resurselor (obiecte și conținut lexical)

S-au organizat ședințe de antrenare a membrilor proiectului în utilizarea interfeței OOCIAT. Simultan, a fost redactat un manual de adnotare, care să fie folosit nu numai de experți lingviști, dar și de voluntari nespecialiști[6]. Diferite bug-uri în funcționarea interfeței au fost semnalate și discutate în consorțiul proiectului. Raportarea bug-urilor s-a făcut atât în prima etapă dar a continuat și în Etapa a II-a, prin înregistrarea lor într-un document plasat în Drive-ul Google[7].

Act. 1.2 - Diseminare și permanentizare (partener: UAIC)
A. Co-organizarea ediției a 15-a a "International Conference on Resources and Tools for Natural Language Processing (ConsILR-2020), derulată online între 14-16 decembrie 2020.

ConsILR-2020, cea de a 15-a ediție a seriei de conferințe începută în 2001, a fost organizată în colaborare de: două institute ale Academiei Române, Institutul de Cercetări în Inteligență Artificială "Mihai Drăgănescu" din București și Institutul de Informatică Teoretică din Filiala Iași, Facultatea de Informatică a Universității "Alexandru Ioan Cuza" din Iași, Asociația Română de Lingvistică Computațională, desfășurându-se sub auspiciile Academiei de Științe Tehnice din România, în perioada 14-16 decembrie 2020 online, în următorii parametri:

- 5 conferințe invitate și un număr de 19 lucrări;
- site web: https://profs.info.uaic.ro/~consilr/2020/
- volumul tipărit la Editura Universității "Alexandru Ioan cuza" din Iași[8] are referința:

V. Barbu Mititelu, E. Irimia, D. Tufiș, D. Cristea (2020). Proceedings of the 15th edition of the International Conference on Linguistic Resources and Tools for Natural Language Processing – ConsILR-2019, Ed. Universității "Alexandru Ioan Cuza" din Iași, ISSB: 1843-911X.

În a 3-a zi a Conferinței s-a desfășurat și cel de al 3-a workshop al proiectului ReTeRom. Implicarea membrilor proiectului DeLORo în organizarea ambelor evenimente a fost esențială.

B. Elaborarea primelor publicații ale proiectului.

Lista publicațiilor elaborate de membri ai proiectului care menționează tehnologiile DeLORo anterior începerii proiectului precum și în prima etapă este următoarea:

- Constantin Cristian Padurariu (2019). State-of-the-Art Approaches to Image to Text Conversion, first PhD report, "Alexandru Ioan Cuza" University of Iași, Faculty of Computer Science.
- Pădurariu C., Cristea, D. (2019). Solution for scanned documents segmentation and letter recognition, in Proceedings of the 14th edition of the International Conference on Linguistic Resources and Tools for Natural Language Processing – ConsILR-2019, Ed. Universității "Alexandru Ioan Cuza" din Iași, p. 127-137, ISSB: 1843-911X, link: https://profs.info.uaic.ro/~dcristea/papers/Solutions%20for%20scaned%20documents%20segmentation%20and%20letter%20recognition.pdf
- Constantin Cristian Padurariu (2020). From Scan to Text. A Solution for Deciphering Old Cyrillic Documents to Modern Latin Language, second PhD report, "Alexandru Ioan Cuza" University of Iași, Faculty of Computer Science.

---

[6] Manualul poate fi găsit la adresa:
[7] https://docs.google.com/spreadsheets/d/1_bcxaN1rsa0JfqQoYnw19hPGX5y0yat4V7rMN8WQsnE/edit#gid=0
[8] Poate fi găsit aici: https://profs.info.uaic.ro/~consilr/2021/wp-content/uploads/2021/03/volum-ConsILR-v-4-final-revizuit.pdf

- D. Cristea, C. Pădurariu, P. Rebeja, M. Onofrei (2020). From Scan to Text. Methodology, Solutions and Perspectives of Deciphering Old Cyrillic Romanian Documents into the Latin Script. In: Knowledge, Language, Models, Volume in Honour of Prof. Galia Angelova on Her 65th Birthday. Milena Slavcheva, editor. INCOMA Ltd. Shoumen, BULGARIA. ISBN 978-954-452-062-5, pp. 38-56, link: https://profs.info.uaic.ro/~dcristea/papers/Paper%20volume%20Bulgaria-Cristea_etAl.pdf

Lista comunicărilor care au menționat proiectul:

- D. Cristea (2020). Resources for Romanian NLP. Building and using them. Invited talk in the Romanian AI Days, 2 decembrie (online).

Act. 1.3 - Site web, implementare și integrare
A. Se va construi structura site-ului web al Proiectului.

Site-ul web poate fi accesat la adresa: http://deloro.iit.academiaromana-is.ro/ și are structura:

- Informații proiect
- Obiectiv
- Istoric
- Platforma tehnologică (OOCIAT)
- Echipa
    - Parteneri
    - Colaboratori
    - Membri
    - Parteneri externi
- Diseminare
    - Publicații
    - Rapoarte
- Contact

B. Se vor încărca în paginile site-ului informații despre proiect, iar în zona de resurse vor fi depuse primele achiziții din ROCC. Site-ul va fi ținut la zi pe toată durata de derulare a proiectului.

Paginile site-ului sunt complet încărcate și ținute la zi permanent.

III. Concluzii

Alte informații utile asupra proiectului pot fi găsite la următoarele adrese web:

- Pagina oficială a proiectului: http://deloro.iit.academiaromana-is.ro/
- Spațiu de lucru GitHub: https://github.com/deloro-project
- Scurt manual de utilizare pentru scriptul de import al datelor (în limba engleză): https://github.com/deloro-project/rocc-pipelines#import-data
- Schemele pentru validarea fișierelor XML ROCC: https://github.com/deloro-project/rocc-schema

Obiectivele Etapei I a proiectului DeLORo au fost realizate în totalitate.

Director Proiect,
Dan Cristea