GABRIELA HAJA, ELENA ISABELLE TAMBA

MODELE DE INTELIGENȚĂ ARTIFICIALĂ (DEEP LEARNING) APLICATE ÎN ANALIZA SINTACTICO-SEMANTICĂ A LIMBII ROMÂNE VECHE (DeLORo). PERSPECTIVA FILOLOGICĂ

Ce este DeLORo?

DeLORo¹ este un proiect experimental demonstrativ, care are drept scop realizarea unei tehnologii de recunoaștere și transliterare a textelor românești, manuscrise (semiunciale) și tipărite (cu alfabet chirilic sau de tranziție). Spre deosebire de proiecte similare – de exemplu Abbyy FineReader sau Transkribus – care au în vedere numai recunoașterea optică și reproducerea fidelă a caracterelor dintr-o imagine, proiectul de față are ca principal obiectiv transpunerea textului chirilic scanat în text editabil transliterat în alfabet latin. O altă deosebire față de proiecte precum cele menționate este aceea că rezultatele obținute în DeLORo vor fi puse gratuit la dispoziția specialiștilor (filologi, istorici etc.), dar și a publicului larg, interesat de accesul mai facil la documente românești vechi.

Echipa DeLORo

Partenerii implicați în proiect sunt: Academia Română – Filiala Iași, prin Institutul de Informatică Teoretică și Institutul de Filologie Română „A. Philippide” și Universitatea din București prin Facultatea de Matematică și Informatică. Colaboratorii externi ai proiectului sunt: Biblioteca Academiei Române din București, Universitatea „Alexandru Ioan Cuza” din Iași, prin Facultatea de Litere și Facultatea de Informatică, Biblioteca Centrală Universitară „Mihai Eminescu” din Iași, Biblioteca Academiei Române – Filiala Iași, dexonline.ro.

Echipa proiectului DeLORo este una mixtă, formată din specialiști în lingvistică computațională, prelucrarea limbajului natural etc. și în filologie, după cum urmează: dr. Dan Cristea m.c. al A.R. (director de proiect), dr. Gabriela Haja, dr. Elena Isabelle Tamba, dr. Roxana Vieru, dr. Mihaela Onofrei (experți filologi), dr. Daniela Gîfu, dr. Paula Crucianu, CS Cecilia Bolea (participante la adnotare), drd. Cristian Pădurariu, drd. Petru Rebeja, drd. Andrei Scutelnicu (informaticieni) – de la Iași, expert în texte românești vechi Gabriela Dumitrescu, precum și dr. Marius Popescu, dr. Radu Tudor Ionescu, dr. Mihaela Gaman, ing. Radu Borza – de la București.

¹ Această lucrare prezintă aspecte filologice din proiectul *Artificial Intelligence Models (Deep Learning#) Applied in the Analysis of Old Romanian Language (DeLORo – Deep Learning for Old Romanian)*, proiect UEFISCDI – PN-III-P2-2.1-PED-2019-3952, no. 400PED; director: prof. univ. dr. Dan Cristea m. c. A. R.) (2020–2022).

Obiectivele DeLORo

Scopul principal al acestui proiect este construirea unei tehnologii care să permită transpunerea automată a tipăriturilor și a manuscriselor semiunciale românești vechi din alfabet chirilic în alfabet latin. Un alt obiectiv important este oferirea, cu acces liber, a acestei tehnologii unor categorii variate de utilizatori. Drept urmare, rezultatele proiectului vor aduce în atenția publicului larg aceste tipuri de texte.

Pe lângă obiectivul principal privind tehnologia de transliterare a textelor chirilice, s-au conturat și obiective specifice precum identificarea automată a unor modele de limbă română (lexic, morfologie), pentru fiecare jumătate de secol din perioada avută în vedere (sec. al XVI-lea – prima jumătate a secolului al XIX-lea).

Activitatea filologilor din proiect este determinată de atingerea acestor obiective. De aceea, alcătuirea corpusului de texte scanate, pentru care există sau nu ediții critice, maniera de adnotare manuală a eșantioanelor de texte (transliterare vs. transcriere interpretativă) au în vedere, pe de o parte, facilitarea antrenării programului de recunoaștere a slovelor și de redare a lor cu litere latine și, pe de altă parte, crearea premiselor ca materialul furnizat de filologi (prin adnotare și transliterare/transcriere) să asigure rezultate valide în generarea modelelor de limbă. Aceste modele de limbă sunt necesare pentru ameliorarea rezultatelor obținute prin tehnologia de recunoaștere și transliterare automată.

Etape de lucru planificate²

1. Achiziția de date
 - 1.1. Achiziția de corpus – în funcție de calitatea documentului, tipul de scriere (tipăritură sau uncial).
 - 1.2. Achiziția de metadate privitoare la documente (conform Catalogului BAR București).
 - 1.3. Construcția de modele diacronice – modele de limbă – lexic.
2. Crearea tehnologiei
 - 2.1. Identificarea obiectelor adnotate: obiecte grafice (frontispiciu și ornament); obiecte primare (titlu, coloană, rând, caracter); obiecte secundare (modificator – accent, titla, text marginal, text interliniar, majuscula).
 - 2.2. Dezvoltarea modelului de recunoaștere a slovelor
 - 2.3. Modele de clusterizare lexicală (metodă specifică *machine learning* pentru identificarea și gruparea datelor similare în structuri mai mari, astfel încât acestea să fie mai ușor de gestionat).
3. Antrenament și evaluare
 - 3.1. Evaluarea și antrenamentul modelelor pentru identificarea de obiecte și recunoașterea caracterelor
 - 3.2 Analiza lexicală și semantică a limbii române în evoluția ei – etapa I.

² Informațiile privitoare la etapele de lucru sunt preluate din descrierea proiectului din aplicația depusă.

Echipa de filologi este implicată în activitatea 1.1. *Achiziția de corpus*, prin stabilirea listei de texte care alcătuiesc corpusul de lucru, și în activitatea 2.1. *Identificarea obiectelor adnotate*, prin adnotarea manuală a unor eşantioane de câte zece pagini din fiecare document. Dacă timpul o va permite în această etapă, rezultatele obținute în cadrul activităților 1.3. *Construcția de modele diacronice de limbă* și 3. *Antrenament și evaluare* vor fi validate, parțial, de filologi.

Colecția de texte din DeLORo

Cuprinde un număr de 60 de documente scanate, selectate după mai multe criterii: criteriul temporal, criteriul spațial, tipul textului (tipăritură/manuscris), stilul funcțional, limba folosită, dificultatea textului și, implicit, a adnotării.

Așadar, avem în vedere diferite tipuri de texte, grupate conform criteriilor de mai sus astfel:

- după conținut: religioase și laice (text științific, text literar, text publicistic),
- după formă: manuscrise semiunciale și tipărituri (cu alfabet chirilic și de tranziție),
- după repartiția geografică: pe cât posibil, din toate regiunile istorice,
- după dificultatea adnotării (calitatea originalului; tipul scrierii; tipul de alfabet etc.).

Potrivit planului inițial, au fost selectate din texte din perioada menționată, grupate pe jumătăți de secol, tipărituri și manuscrise, în toate stilurile funcționale existente, preponderent în limba română (incluzând însă și textele bilingve păstrate din secolul al XVI-lea), cu diverse grade de dificultate.

Sursa principală a colecției de texte, în format .pdf, o reprezintă Biblioteca Academiei din București, unde acestea au fost digitalizate. Colecția de imagini va fi folosită numai în scop de cercetare și nu își schimbă proprietarul de drept (B.A.R. București, dar și alte biblioteci care au permis accesul online la colecțiile de carte veche, precum B.C.U. Iași, B.C.U. Cluj).

În cazul în care unele titluri nu fuseseră incluse pe lista cărților digitalizate, acestea au fost recuperate prin bunăvoința doamnei Gabriela Dumitrescu, șef Serviciu Manuscrise, Carte Rară. Grație colaborării cu B.A.R. București, echipa de informaticieni din proiect a putut avea acces la baza de metadate biblioteconomice de unde au fost extrase metadatele care interesează în stabilirea modelelor de limbă.

În urma adnotării manuale și automate a acestor documente, vor rezulta corpuri de limbă română, ce vor putea fi grupate după mai multe criterii, în funcție de metadatele aferente fiecărui document, dar și de caracteristicile lexico-gramaticale.

Calitatea filologică a acestor corpuri depinde de următorii factori:

- existența unei ediții critice în format editabil, fără erori formale (neocerizate),
- necesitatea alcătuirii unui număr minim necesar de slove adnotate, suficient pentru antrenarea programului de recunoaștere și transliterare automată (ceea ce justifică opțiunea filologilor de a transcrie cu minime intervenții textele pentru care nu avem la dispoziție o ediție critică în format editabil).

(Redăm pe paginile imediat următoare fragmente din tabelul alcătuit de filologi la începutul proiectului, iar apoi diferite tipuri de texte pe care le avem în vedere în proiect, în funcție și de gradul de dificultate al adnotării.)

An	Autor	Titlu	Loc	Tipar /ms.	Stil	Limba	Difi- cul- tate rela- tivă	Cota BAR	Transcris (editat critic/ transcris)	Nr. de p.	Meta- date XML BAR
1500/ 1520	(?)	<i>Psaltirea Hurmuzaki</i>	Maramureș (?)	uncial	beletristic (text biblic)	română	3	MS. rom. 3077	DA	249	DA
1559/ 1564	Coresi	<i>Catehism</i>	Brașov (Trans S = Munt)	tipări- tură	bisericesc	română	1	MS. ROM .5032	DA	21	NU
1563	Coresi	<i>Apostol</i>	Brașov (Trans S = Munt)	tipări- tură	bisericesc (text biblic)	română	1	CRV 12	DA	618	DA

An	Autor	Titlu	Loc	Tipar/ ms.	Stil	Limba	Dificultate relativă	Cota BAR	Transcris (editat critic/ transcris)	Nr. de p.	Meta- date XML BAR
cca 1658– 1661	Kigalas (autorul originalului în ngr.)	<i>Hronograf den începutul lumii Ms. 3517</i>	Mold.	ms. uncial	științific, beletristic traducere	română	2	MS. ROM 3517	DA	608	NU
1700	Radovici, Gheorghe (tipograf, pref.) Filotei Svetagoreț ul (trad.)	<i>Învățăături creștinești foarte de folos, acuma întâi scoase de pre limba grecească, pre limbă rumânească că...</i>	Snagov Munt.	Tipar negru; 21 de rânduri pe pagină	bisericesc	română	1	CRV 118 De făcut		225	NU

An	Autor	Titlu	Loc	Tipar/ ms.	Stil	Limba	Dificultate relativă	Cota BAR	Transcris (editat critic/ transcris)	Nr. de p.	Meta- date XML BAR
1799	Iorgovici, Paul	<i>Observații de limba românească. Gramatica românească</i>	Buda	Tipar negru; 20–23 de rânduri pe pagină	știin- țific	română	2	CRV 621	DA	116	DA
1806	Micu- Klein, Samuil	<i>Istoria, lucrurile și întâmplările românilor pre scurt așezată și din mulți vechi și noi scriitori culeasă și scrisă</i>	Buda	Tipar negru în pagină cu chenar, 28 de rânduri pe pagină	știin- țific	română	1	CRV 695		92	NU

Manuscris semiuncial (exemplu extras din *Codicele Voronețean*, 1563–1583)

Manuscris semiuncial slavo-român, cu dificultate medie (exemplu din *Floarea darurilor*, Putna, 1592/1604)

Manuscris semiuncial, foarte dificil de adnotat/procesat (exemple extrase din *Codicele Bratul*, 1559–1560) – cu pete (opace), cu rânduri oblice sau curbate, cu slove indescifrabile etc.

Tipăritură – cu dificultate medie de adnotare, generată de pete, adnotări manuscrite marginale, sublinieri etc. (exemplu din Coresi, *Apostol românesc*, Braşov, 1563).

Tipăritură dificil de adnotat, din cauza rândurilor tipărite oblic (exemplu din *Bucvar*, Iaşi, 1755)

Tipăritură dificil de adnotat, din cauza tiparului suprapus (exemplu din *Catavasiariu tălmăcit pre limba rumânească cu multe trebuincioase cântări ce să cântă preste tot anul, care s-au tipărit acum...*, Bucureşti, 1742)

Tipăritură în alfabet de tranziție (exemplu din Ioan Văcărescu, *Poezii alese*, [Bucureşti], 1830)

Tipuri de texte care nu se adnotează în DeLORo

Texte cu pagini parțial distruse (pete opace, pagini cu grad mare de transparență, palimpsest, pagini îndoite/rupte) – exemplu din *Codicele Voronețean* (ultima filă deteriorată #86).

Manuscris cu ligaturi (exemplu din Ms. 4389).

Este cazul textelor din perioada secolelor al XVI-lea – al XVII-lea, pentru care beneficiem de rezultatele obținute în cadrul proiectului ieșean Text Vechi (<https://textvechi.ro>). Imaginile din această perioadă vor fi adnotate, atribuindu-li-se transcrierile interpretative avute la dispoziție.

Pentru textele din secolele ulterioare situația este diferită: mare parte dintre documentele selectate fie nu s-au bucurat de (re)editare critică, fie au fost editate în perioade pentru care nu avem la dispoziție transcrierile într-un format electronic editabil, fără erori. În cazul acestora, am decis să adnotăm imaginile cu intervenții minime în transcriere. Mai precis, am preferat transcrierii interpretative transliterarea, astfel încât să punem la dispoziția program de recunoaștere și transliterare automată un model care să implice cât mai puține situații de ambiguitate. Am încercat, așadar, să redăm cât mai fidel textul original, fără să coborâm slovele suprascrise sau fără să completăm cuvintele abreviate. Tot pentru eliminarea ambiguităților și facilitarea recunoașterii și echivalării slovelor, am optat pentru ortografia cu *î*. A fost nevoie să luăm decizii în chestiuni pe care filologia tradițională le rezolvă tacit: marcarea pauzei între cuvinte, respectarea ortografiei actuale în cazul ortogramelor, marcarea la final de rând a despărțirii în silabe ș.a. În aceste cazuri am optat pentru transcrierea cu pauze între cuvinte, redând ortogramele conform normelor actuale, cu marcarea la final de rând a despărțirii în silabe ș.a.

Pentru alcătuirea corpusurilor textuale și a modelelor de limbă s-a făcut apel și la rezultatele obținute în alte două proiecte de cercetare: *Monumenta linguae Dacoromanorum. Biblia 1688 (MLD)* și *SOLIROM* (<https://solirom.ro/>). Astfel, baza de date MLD, realizată de cercetătorii ieșeni în proiectul de reeditare critică a *Bibliei de la București* paralel cu alte două traduceri din același secol, ms. 45 și ms. 4389, cuprinde adnotarea morfologică integrală a textului Bibliei, cu stabilirea lemelor de către editori (începând cu volumul tipărit în anul 2008 până la finalizarea ediției în 2015 – <https://biblia1688.solirom.ro/7/> și http://consilr.info.uaic.ro/~mld/monumenta/MLD_VIII/start.html). Colecția de forme flexionate (400.000) corelate cu lemele corespunzătoare din baza MLD constituie un corpus de antrenare foarte util pentru etapele de lucru din cadrul activităților 2. *Crearea tehnologiei* și 3. *Antrenament și evaluare*.

Așadar, baza de date finală va conține o colecție de imagini (adnotate cu date lingvistice și grafice și cu metadate), care va constitui o resursă pentru cercetări viitoare, care vor putea duce la optimizarea rezultatelor preconizate în proiectul DeLORo.

Interfața de adnotare OOCIAT³ (Online Old Cyrillic Image Annotation Tool)

Interfața OOCIAT permite filologilor adnotarea imaginii și transcrierea textului aferent. Este un instrument de lucru ce facilitează accesul de la distanță al utilizatorilor (adnotatorilor) la baza de date DeLORo, care cuprinde colecția de documente în format imagine însoțită de metadate și de datele rezultate din adnotarea manuală a acestora.

³ Această interfață a fost proiectată și realizată de drd. Cristian Pădurariu.

Interfața este organizată pe mai multe câmpuri. Primul câmp este destinat selecției documentului și a paginii ce urmează să fie adnotată. Astfel, pe prima linie apare titlul documentului și numărul imaginii aflate în lucru, iar al doilea rând permite trecerea de la o imagine la alta din aceeași lucrare și accesul la colecția de imagini din lucrarea selectată din lista documentelor.

Al doilea câmp al interfeței permite notarea de informații privitoare la starea paginilor originale și, implicit, a imaginilor, informații relevante pentru calitatea adnotării automate: sunt avute în vedere deopotrivă elementele privitoare la starea fizică a paginii respective (porțiuni lipsă, pete opace, transparența hârtiei, palimpsest) și cele referitoare la textul propriu-zis (calitatea tiparului, prezența slovelor suprascrise, prezența corecturilor și/sau a notelor marginale). În acest fel, se completează colecția de metadata a documentelor, cu informații privitoare la gradul de dificultate a adnotării pentru fiecare pagină/imagine a documentului respectiv.

Al treilea câmp este destinat adnotării conținutului fiecărei pagini/imagini. Se operează două tipuri de adnotări:

1. adnotarea elementelor nonlingvistice (grafice): frontispicii, ornamente;
2. adnotarea textului propriu-zis, la mai multe niveluri: coloană, rând, majusculă (ornată), cuvânt, slovă/caracter.

Clasificarea obiectelor (numele generic dat elementelor adnotate prezente pe o pagină) s-a făcut în funcție de principiile după care va funcționa programul DeLORO, care va fi antrenat să recunoască în primul rând obiectele primare pentru ca, ulterior, pentru fiecare dintre elementele din coloana a doua (v. *infra*) să opereze procese secundare de recunoaștere și transcriere. Obiectele grafice sunt adnotate pentru ca programul să le poată identifica, fără a mai face efortul de transliterare.

Obiecte primare	Obiecte secundare	Obiecte grafice
<ul style="list-style-type: none"> • <titlu> • <coloană> => L, R • <rând> • <caracter> 	<ul style="list-style-type: none"> • <modificator⁴> • <text marginal> • <text interliniar⁵> • <majusculă> • <trimitere pe manșetă⁶> • <trimitere interliniară⁷> 	<ul style="list-style-type: none"> • <frontispiciu> • <ornament>

Practic, operația de adnotare de către filologi constă în trasarea unui chenar în jurul obiectului adnotat, stabilirea tipului de obiect, transliterarea/transcrierea conținutului în cazul obiectelor primare și secundare. În acest fel, fiecărui obiect adnotat îi sunt atribuite coordonatele spațiale în cadrul imaginii/paginii și reprezentarea în scris. Pentru obiectele grafice, semnificația lor este dată de tipul selectat.

(Mai jos, începând cu pagina următoare, prezentăm capturi de ecran cu exemple de pagini adnotate.)

Corpus de antrenare pentru DeLORo

Pentru realizarea unui program automat de procesare a limbajului natural este nevoie de o etapă de antrenare a programului respectiv care se bazează pe un „corpus de aur” realizat manual de specialiști lingviști, într-un limbaj accesibil mașinii (cel mai adesea, în limbaj .xml), în care informația lingvistică este etichetată/adnotată cu informații suplimentare, de diverse categorii și niveluri. În etapa de antrenare, este important ca specialiștii lingviști să corecteze erorile generate de sistemul de operare, pe măsură ce sunt procesate noi date de intrare.

Pentru tehnologia DeLORo s-a decis alcătuirea unui *gold corpus* care să cuprindă un număr de minimum 120 pagini adnotate pentru fiecare jumătate de secol. De asemenea, s-a decis ca dintre acestea, 60 % să fie reprezentate pagini cu dificultate minimă în adnotare, 40 % – pagini cu dificultate medie, 20 % – pagini cu dificultate mare.

În realizarea acestui corpus am întâlnit următoarele situații problematice:

– pentru prima jumătate a secolului al XVI-lea ni s-au păstrat foarte puține documente; pentru această perioadă, vor fi adnotate documentele existente;

⁴ Modificatorii sunt semnele care aduc o informație suplimentară privitoare la obiectele primare – accentul și titla.

⁵ Textul interliniar este reprezentat de orice slovă/grup de slove care este suprascris.

⁶ Trimiterea pe manșetă reprezintă semnul/semnele (de tip vrahie, de exemplu) reluat care anunță o notație textuală marginală.

⁷ Trimiterea interliniară reprezintă semnul/semnele (de tip vrahie, de exemplu) care anunță în text, în poziție suprascrisă, o notație textuală marginală.

Tipăritură în alfabet chirilic

Învățătură preoșilor pe scurt. De șapte taini ale
besearicii, cu dreapte dovediri den Pravila
Sfinților Părinți..., Buzău, 1702 (alfabet
chirilic)

Titlu: Taina sfințului Bote
Rând: te să fie atâta apă cită poate fi adundare de 3
Rând: ori. Precum învăța sfinții Părinți. La pra: Cap
Rând: feate. Întru carele ne botează, și o crediță.
Rând: trandafir, sau seara, sau altă apă fiată cu al
Rând: Pentru aceia s Botează unuși iate. Precum
Rând: scaldătoarea sfințului Botez, trebuia să fie
Rând: trupului: carea trebuia să se sfințească cu mlt
Rând: sfințită cu mltve, și îndată să se botează
Rând: dize sfința scriptură: Un Dmns, o crediță
Rând: țească pre ea, curățindu o cu spălare de apă în cu
Rând: vele de la Botez: apoi cu aceia să se botează
Rând: Acea să se știe, că această taină o dată luidu
Rând: Pruncuș, cu cuvintele ce le vom arăta. Și ia
Rând: cu apă firească: adevă să nu fie apă meșugită
Rând: sau amestecată cu altăceva, precum iaste apă di
Rând: reace să poată încălzi cită poate răbda firea
Rând: Pruncuș: lără la vreme de nevoie poate fi și ne
Rând: un Botez. Însă să se știe că spălarea acesta în
Rând: te dresuri: ci singură apă. Și de iaste prea Rând: vint.
Rând: să nu să dă a doară: Că un Dmndău iate în trei
Modifier:~ Modifier:~ Modifier:~ Modifier:~
Modifier:~ Modifier:~ Modifier:~ Modifier:~
Modifier:~ Modifier:~ Modifier:~ Modifier:~
Modifier:~ Modifier:~ Modifier:~ Modifier:~
Modifier:~ Modifier:~ Modifier:~ Modifier:~
Text marginal: 9 Text marginal: efe
Text marginal: 141 Sau ș alite Text marginal: st
Text marginal: 25 Text interliniar: Text interliniar
Text interliniar: Text interliniar: Text interliniar
Text interliniar: Text interliniar: Text interliniar
Text interliniar: Text interliniar: Text interliniar
Text interliniar: Text interliniar: Text interliniar
Text interliniar: Text interliniar: Text interliniar
Text interliniar: Text interliniar: Text interliniar

Catavasiariu tălmăcit pre limba rumă-
nească cu multe trebuincioase cântări ce
să cântă preste tot anul, care s-au tipărit
acum..., București, 1742

Ornament: null Ornament: null Titlu: Nașterii.
Rând: Venit-ai intrupindu-te din cea ne-
Rând: din muntele cel cu umbra deasă.
Rând: fecioară ai odrăslit, celș lăuda,
Rând: de trup și Dmnezu: mărire pu-
Rând: nainte o au vestit, învrednicin- Rând: terii tale, Dmne.
Rând: ispitită de bărbat, celș fără
Rând: cioarei, au egiț cuvintul spre
Rând: durărilor, pre ingerul statului
Rând: tău celui mare, dăruindu-ne pace
Rând: Dmnezu fiindș păcii, Tată in-
Rând: du-să negrait ai vedea inchipuirea.
Rând: I-a trimis noao. Dec povățuți fil
Rând: înnoirea noroadelo.
Rând: Că pruncș tinărș din muntele fe- Rând: de
Rând: d la lumina cunoștitei de Dmnezu, Literă: 2 Literă: t
Literă: o Literă: r Literă: n Literă: 2 Literă: i Literă: n
Modifier:~ Modifier:~ Modifier:~ Modifier:~

Tipăritură în alfabet de tranziție

Ioan Văcărescu, *Poezii alese*, [București],
1830

Rând: Stam gindind: așa natura Rând: Planeti, Răsărit, Apus,
Rând: 11 Rând: Cita slava se cuvine
Rând: D-amărunt privind fapturn
Rând: Mă duceam și mă culcam,
Rând: Și-nsumi mulțumit cu mine Rând: La unul ca el i-o dam
Rând: Cin-ar fi-mvirtind de sus!
Rând: Strigind intreb "? Cine este;
Rând: Dormind asfel într-o noapte Rând: ? Cin-alcea s-a băgat;
Rând: Somnu-nțiu, ca vreun ceas,
Rând: ? Cin-alci fără de veste Rând: Simț că de copil e glas;
Rând: Mă deștept înțeleg șoapte Rând: Prea ciudat un copilaș,
Rând: Se roagă să i dau lăcaș. Rând: Fricos cere ertăciune,
Rând: Mă uit, ? ce să văz, minune! Rând: A-drăznit de a intrat."

- prezența textelor bilingve (slavo-române, româno-slave) sau multilingve (vezi Codex Sturdzanus); în aceste documente are este adnotat doar textul românesc;
- scrierea semiuncială este specifică secolelor al XVI-lea și al XVII-lea; începând cu secolul al XVIII-lea este utilizată scrierea cursivă, pe care specialiștii informaticieni nu o au în vedere în această etapă; pentru această perioadă vom adnota numai tipărituri;
- documentele din secolul al XIX-lea prezintă dificultăți generate de utilizarea alfabetelor de tranziție.

Concluzii

Contribuția filologilor în cadrul proiectului este importantă pentru că acuratețea rezultatelor obținute prin tehnologia DeLORo depinde într-o bună măsură de volumul și de calitatea activității de adnotare manuală a corpusului pentru antrenare. Procesul de adnotare este însă unul cronofag.

Din perspectiva noastră, faptul că tehnologia construită în cadrul proiectului va fi pusă la dispoziția cercetătorilor poate fi salutar, în condițiile în care servicii similare, dezvoltate deja, sunt costisitoare. În faza actuală, preconizăm că rezultatele transcrierii automate nu vor fi impecabile, dar vor facilita realizarea viitoarelor ediții de texte românești vechi, iar, în etapa ulterioară, programul va putea fi îmbunătățit pe măsură ce va fi utilizat și corectat. Chiar în aceste condiții, prin DeLORo, se creează premisele unei digitalizări de nivel superior a colecțiilor de documente vechi românești.

De asemenea, considerăm importantă oferirea acestei tehnologii publicului larg, căruia i se va deschide, astfel, interesul pentru documente, ne(re)editate încă, scrise în limba română până în a doua jumătate a secolului al XIX-lea.

ARTIFICIAL INTELLIGENCE MODELS (DEEP LEARNING) APPLIED
IN THE ANALYSIS OF OLD ROMANIAN LANGUAGE (DELORO).
PHILOLOGICAL POINT OF VIEW
(Abstract)

This paper presents the project *Artificial Intelligence Models (Deep Learning) Applied in the Analysis of Old Romanian Language (DeLORO)* from the perspective of the philologists involved. The main goal of the project is to build a technology that would allow the automatic transposition of Old Romanian semi-uncials prints and manuscripts from the Cyrillic alphabet or the transition to the Latin alphabet. In addition to this objective regarding the technology of transliteration of Cyrillic texts, specific objectives were outlined such as the automatic identification of Romanian language models (vocabulary, morphology) for each half century of the considered period (the XVIth century until 1862). The philologist's activity in the project is determined by the achievement of these objectives. Therefore, the composition of the corpus of scanned texts (with or without published critical editions), the manner of manual annotation of text samples (transliteration vs. transcription) are intended to facilitate the training of the word recognition technology and transpose them in Latin alphabet and to create the premises for the material provided by philologists (by annotation and transliteration/transcription) to ensure valid results for the automatic generation of language models. From our perspective, the fact that the technology created within the project will be made available to researchers can be useful, given that similar services, already developed, are expensive. In the current phase, we expect that the results of the automatic transcription will not be impeccable, but will facilitate future editions of Old Romanian texts, and in the later stage, the program can be improved, by being used and revised. Even in these conditions, by realizing this technology, the premises of a higher-level digitalization of the collections of Old Romanian documents are created. It's also important to offer this technology to the public, interested in documents written in Old Romanian.

Cuvinte-cheie: *deep learning, gold corpus, filologie, română veche, manuscrise, tipărituri, digitalizare.*

Keywords: *deep learning, gold corpus, philology, Old Romanian texts, manuscripts, prints, digitalisation.*

*Institutul de Filologie Română „A. Philippide”
al Academiei Române
Iași, str. Codrescu, 2
gabihaja@gmail.com
isabelle.tamba@gmail.com*