

Contribuții la informatizarea cercetării filologice românești: MLD. Biblia 1688 și eDTLR

Gabriela HAJA, Elena DĂNILĂ,
Marius-Radu CLIM, Vlad PATRAȘ

1. Începuturile informatizării cercetării filologice la Institutul „A. Philippide”

În ultima vreme, în filologia românească, s-a început elaborarea unor programe ample de realizare a unor baze de date și a unor instrumente care să servească la modernizarea metodologiei de lucru. Acest lucru va permite sincronizarea cercetării lexicografice românești cu cercetări de același tip din întreaga lume.

Colaborarea cu specialiști de la Institutul de Informatică Teoretică al Academiei Române, începând cu anii '90, și de la Facultatea de Informatică, începând cu 2001. Proiectele de acest tip sunt facilitate și de specializarea unor cercetători tineri în domeniul lingvisticii computaționale, începând cu anul 2001 (anul creării unui astfel de masterat la Facultatea de Informatică din Iași). Tendința generală din ultima vreme vizează realizarea unor proiecte de cercetare interdisciplinară, de nivel național sau european, cu participarea unor specialiști din mai multe domenii: lingvistică, informatică, istorie literară, filosofie ș.a.

La Institutul de Filologie Română „A. Philippide”, aflat sub egida Filialei din Iași Academiei Române, astfel de demersuri au fost începute în aproape toate departamentele în cadrul cărora se desfășoară proiecte fundamentale ale Academiei. Astfel, prima lucrare din planul Academiei realizată cu mijloace informatizate a fost *Noul Atlas lingvistic pe regiuni. Moldova și Bucovina*. S-au inițiat, ulterior, în cadrul Institutului, demersurile de informatizare a lucrării fundamentale a lexicografiei românești – *Dicționarul limbii române*.

Până în prezent au fost finalizate sau sunt în curs de elaborare, în cadrul Departamentului de Lexicologie – Lexicografie, următoarele proiecte:

a) Dicționarul limbii române (DLR) în format electronic. Studii privind achiziționarea (cod CNCSIS 1815), grant finanțat de Ministerul Educației și Cercetării (MEC) prin Consiliul Național al Cercetării Științifice din Învățământul Superior (CNCSIS), desfășurat în perioada 2003–2005 la Institutul de Filologie Română „A. Philippide”. Prin acest proiect s-a verificat și demonstrat posibilitatea transformării *Dicționarului limbii române* din text tipărit în text electronic adnotat¹, prelucrat cu

¹ Textul electronic adnotat este un text analizat și marcat din punct de vedere formal astfel încât să poată fi consultat, corectat, modificat etc. de către specialiștii lexicografii, cu ajutorul calculatorului. Există posibilitatea extragerii din formatul complet a unei forme destinate numai consultării, care să se

ajutorul unui program specific, DLReX – un instrument de achiziționare, prelucrare și consultare a DLR, bazat pe o euristică prin care sunt recunoscute diferitele câmpuri formale ale textului unui articol, putându-se identifica automat textul definițiilor, al citatelor și acela al siglelor.

b) *Resurse lingvistice în format electronic: Monumenta linguae Dacoromanorum. Biblia 1688. Regum I, Regum II – Ediție critică și corpus adnotat. (MLD. Biblia 1688)* (cod CNCISIS 1454), desfășurat în perioada 2006–2007. Prin acest proiect a fost găsită o posibilă metodă de achiziționare în format electronic a unor cărți vechi din Bibliografia DLR, cu aplicație asupra a două cărți din *Biblia* tipărită la București în anul 1688, *A împărățiilor cea dentâiu*, *A împărățiilor a doua*, precum și crearea unor instrumente de indexare și adnotare automată, la nivel de cuvânt, a textelor românești vechi.

c) *DLRI. Bază lexicală informatizată. Derivate.* (cod CNCISIS nr. 1609), început în 2007 și care va fi finalizat în octombrie 2008. Prin acest proiect se propune realizarea unui eșantion lexicografic format din derivatele pe terenul limbii române cu sufixul *-ime* – de origine latină, și cele cu *-iște* – de origine veche slavă, din seria veche DA și din seria nouă a dicționarului DLR, precum și unificarea tehnico-lexicografică a articolelor DA – DLR. Practic, proiectul își propune realizarea unui eșantion bine delimitat etimologico-semantic din viitorul eDTLR.

În paralel au fost demarate lucrări de informatizare și în cadrul Departamentului de Istorie literară de la același Institut (vezi *Baza de date informatizată a Dicționarului General al Literaturii Române (DGLR)*).

În principiu, aceste demersuri de informatizare a cercetării filologice românești au în vedere, pe de o parte, realizarea de *instrumente* informatizate specifice cercetării filologice: *programe* de prelucrare / analiză automată a textului scris/vorbit, *interfețe* de lucru on-line, necesare valorificării resurselor create sau existente, iar, pe de altă parte, crearea de *resurse* lingvistice digitizate: *dicționare* informatizate, *corpusuri* de texte scrise / vorbite.

2. MLD. Biblia 1688

În comunicarea de față detaliem, pentru început, unul dintre granturile care au premers, parțial, parteneriatului din cadrul proiectului complex *eDTLR. Dicționarul Tezaur al Limbii Române în format electronic*, finanțat de CNMP, coordonat de Facultatea de Informatică din Iași, și anume *Resurse lingvistice în format electronic: Monumenta linguae Dacoromanorum. Biblia 1688. Regum I, Regum II – Ediție critică și corpus adnotat (MLD. Biblia 1688)*.

În cadrul proiectului *MLD. Biblia 1688* s-a realizat un *corpus* de mici dimensiuni de texte românești din secolul al XVII-lea, analizate morfologic, printr-un program de împărțire / analiză automată (corectat de lingviști), ceea ce a dus, pe de o parte, la *generarea automată a unui indice de cuvinte și forme* (cuprinzând numele comune din toate cele trei variante de traducere: Biblia 1688, Ms. 45, Ms. 4389) și, pe de altă parte, la *dezvoltarea unui program de analiză*

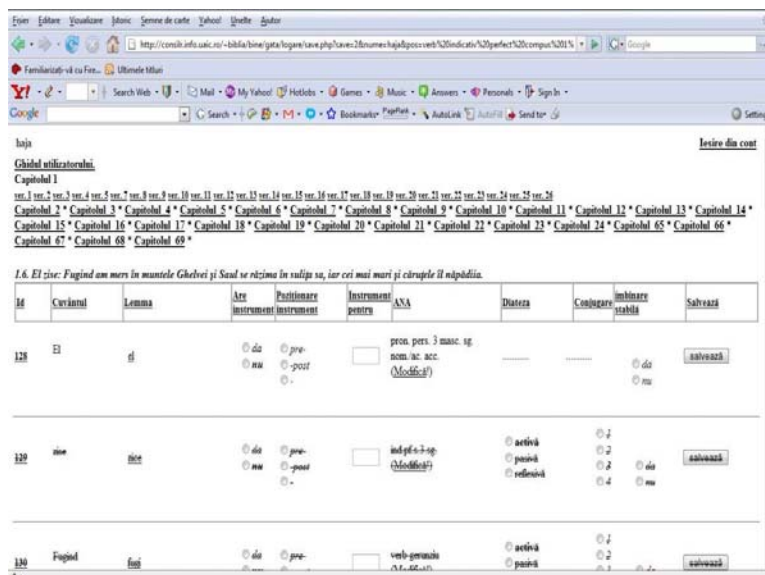
adreseze unui public mai larg decât cel al specialiștilor propriu-ziși. Pentru detalii, vezi și Haja, Dănilă et alii, 2005.

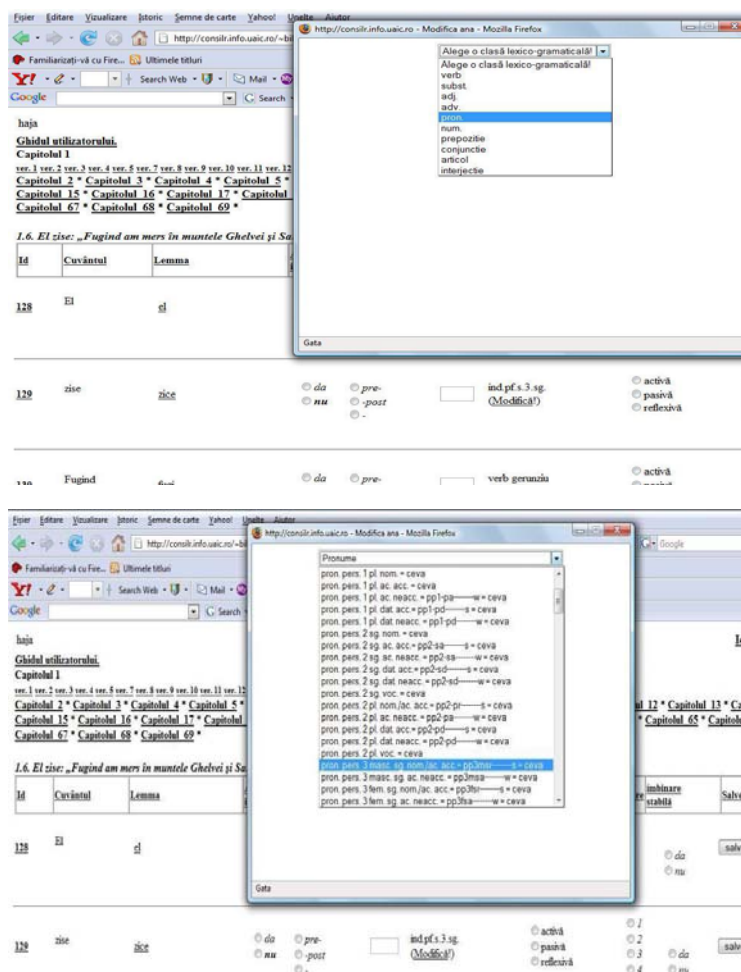
automată, adecvat limbii române vechi și la dezvoltarea unei modalități de lucru la distanță prin intermediul unei interfețe speciale, adaptabile, perfectibile, pentru validarea / corectarea prelucrării automate.

În continuare, prezentăm interfața de validare / corectare a analizei morfologice automate a textului Bibliei (sec. al XVII-lea). Prima captură de ecran reprezintă pagina web special creată pentru acest demers, fiecare dintre cercetătorii implicați având un cont propriu.



Următoarele capturi de ecran prezintă modalitățile de folosire ale acestei interfețe de corectură, cu posibilitățile de optare pentru o anumită categorie gramaticală, pentru diferitele elemente de nuanțare a analizei lexico-gramaticale.





Rezultatele proiectului s-au materializat într-un volum, al șaptelea al cunoscutei serii *Monumenta linguae Dacoromanorum* de la Iași, în forma consacrată și în format electronic. Gradul de recunoaștere a categoriei morfologice al cuvintelor din fondul principal lexical, comun celor două vârste ale limbii române, a fost de peste 95%, fără a lua în calcul situațiile de omonimie. În privința cuvintelor ori a formelor arhaice, procentul a fost destul de mic, după prima prelucrare automată. La a doua analiză automată, efectuată după ce s-a realizat corectura parțială a celei dintâi, rezultatele s-au ameliorat simțitor. Programul a recunoscut cu ușurință numele proprii, într-o primă fază fără a mai adăuga informații de natură morfologică. După prima corectură a specialiștilor lingviști, analiza morfologică a numelor proprii s-a putut realiza, cu o rată mai mare de corectitudine. În ediția de față, nu au fost indexate și numele proprii ori toponimele din textul *Bibliei de la 1688*. Aceasta, deoarece există, din păcate, încă destule lacune în studiile de toponomastică veche, pe de o parte, iar, pe de altă parte, studierea atentă a acestor cuvinte din prima traducere integrală a *Bibliei* în limba română trebuie să constituie

un demers științific de sine stătător, laborios și de lungă durată, cu rezultate pe care le estimăm a fi dintre cele mai interesante.

Pe de altă parte, programul de parsare, adnotare, indexare pus la punct de cel mai tânăr membru al echipei de cercetare, Sebastian Vlad Patraș, masterand în anul I, în domeniul lingvisticii computaționale, la Facultatea de Informatică din Iași, va putea fi antrenat și perfecționat pe măsură ce următoarele volume ale monumentalei ediții vor fi pregătite pentru tipar. Mai mult, acest instrument va putea fi utilizat pentru multe dintre scrierile românești din secolul al XVII-lea, dar și manuscrise și cărți din secolul următor, pentru care prima ediție integrală a Bibliei a constituit un model de limbă românească.

3. eDTLR

eDTLR este un proiect complex, cu finanțare națională, care pune în practică, pentru prima oară, o cooperare deschisă, la cel mai înalt nivel, între specialiști din domenii până de curând aparent incompatibile în lumea academică românească. Scopul principal al acestui demers este construirea formei digitizate a *Dicționarului limbii române* paralel cu încheierea activității de redactare și publicare a celei dintâi ediții a fundamentalei lucrări a Academiei Române. Coordonatorul lucrării este Facultatea de Informatică, Universitatea „Alexandru Ioan Cuza” din Iași. Director: dr. Dan Cristea.

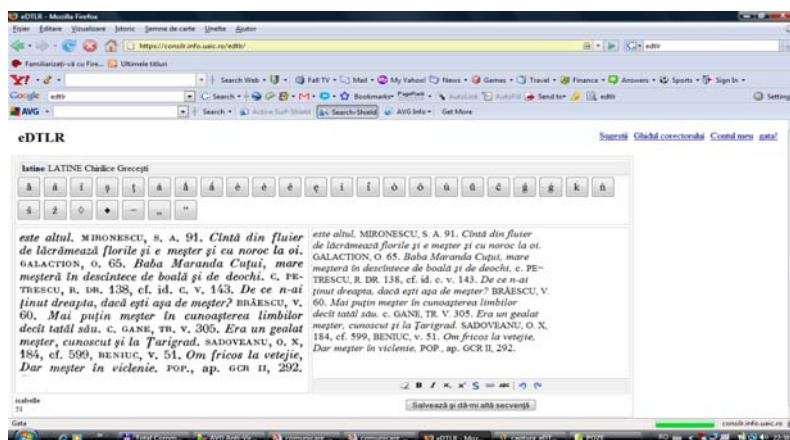
În echipa de cercetare sunt implicați ca parteneri:

- Institutul de Lingvistică „Iorgu Iordan – Al. Rosetti”, Academia Română, București. Responsabil de proiect acad. Marius Sala (prin dr. Monica Busuioc);
- Institutul de Filologie Română „A. Philippide”, Academia Română, Iași. Responsabil de proiect dr. Gabriela Haja;
- Institutul de Lingvistică și Istorie Literară „Sextil Pușcariu”, Academia Română, Cluj-Napoca. Responsabil de proiect dr. Rodica Marian;
- Institutul de Cercetări pentru Inteligență Artificială, Academia Română, București. Responsabil de proiect acad. Dan Tufiș;
- Institutul de Informatică Teoretică, Academia Română, Iași. Responsabil de proiect acad. Horia Neculai Teodorescu;
- Facultatea de Litere, Universitatea „Alexandru Ioan Cuza” din Iași. Responsabil de proiect dr. Eugen Munteanu.

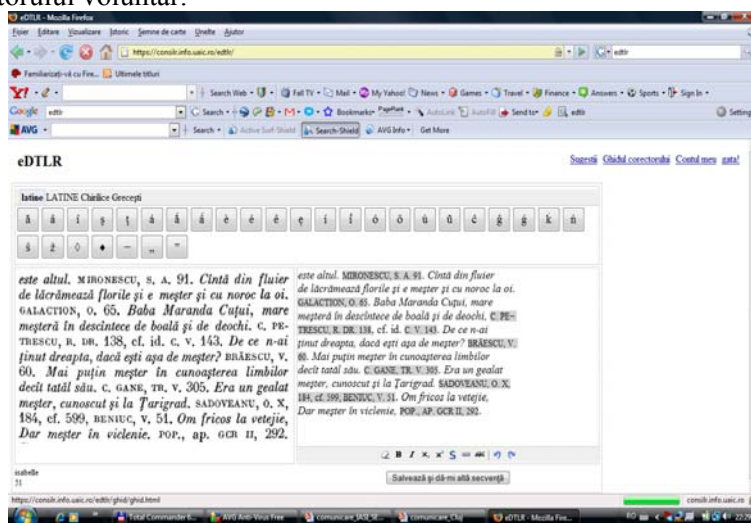
Dintre obiectivele eDTLR până în acest moment au fost realizate: 1. scanarea și convertirea în format editabil a tuturor volumelor DA și DLR publicate până în prezent; 2. realizarea unei interfețe de corectură on-line și corectarea a cca 30% din numărul total de pagini, cu participarea colaboratorilor voluntari, activitate ce are un caracter accentuat inovativ în cercetarea de nivel academic; 3. realizarea unei interfețe de corectură on-line, accesibilă lingviștilor lexicografi implicați în proiect, care vor corecta integral textul scanat și convertit al Dicționarului; 4. realizarea unui program de parsare a unui articol DLR și de generare automată a ierarhiei sale semantice; 5. realizarea unui corpus scanat și convertit într-un format accesibil pentru programul de indexare în surse a citatelor din Dicționar, format din texte de referință ale DA și DLR (potrivit legislației în vigoare).

Specialiștilor lingviști le revine sarcina corecturii atente a textului obținut în urma recunoașterii optice a caracterelor din imaginile scanate ale Dicționarului.

Corectura se face în trei etape: o primă etapă cu ajutorul corectorilor voluntari, nespecialiști și specialiști, care vor face această operație on-line, pe un site special realizat de informaticienii implicați, urmată de o a doua etapă care presupune revizia corecturii făcute de voluntari de către specialiștii lexicografi. În literatura de specialitate, pentru activitățile colaborative, asemănătoare celor ce se preconizează a fi desfășurate în proiect pentru corectarea primară a formatului electronic al Dicționarului, există deja un nume: „crowdsourcing” sau „digital sharecropping”². În captura de ecran următoare se poate vedea interfața de corectură on-line pentru voluntari.



Se observă în această captură de ecran corecturile făcute pe secvența propusă corectorului voluntar.



² Vezi și Cristea, Răschip *et alii* 2007: 195-206.

Cea de a treia etapă va consta în verificarea și corectarea arborilor semantici generați de programul de parsare pentru fiecare articol DA și DLR, pe baza unei gramatici de programare specifice, bazată pe formalizarea normelor de redactare ale Dicționarului.

Contribuții ale proiectului eDLTR.

eDLTR va deschide noi modalități de lucru/studiu/cercetare în lexicografia românească, incluzând latura ei computațională; va oferi singura cale modernă de completare și aducere la zi a marelui dicționar, ceea ce va duce, în viitor, în uniformizarea celor două serii ale Dicționarului, DA și DLR; va oferi posibilitatea de consultare interactivă a Dicționarului de către orice cunoscător al limbii române din arealul de limbă română ori din afara lui. În contextul eforturilor actuale de promovare a multilingvismului în Europa unită, eDLTR va contribui semnificativ la promovarea limbii române.

Noutatea acestui proiect constă în faptul că eDLTR va fi primul dicționar de o asemenea anvergură dedicat limbii române, plasat pe suport electronic. Pentru prima dată, cercetătorii vor putea regăsi citatele direct în sursele bibliografice. În momentul lansării, eDLTR va fi cel mai mare dicționar în format electronic din lume în ceea ce privește numărul de exemple care să susțină sensurile cuvintelor (estimăm că cele aproximativ 3.000 de volume în care vor fi indexate exemplele Dicționarului conțin aproximativ un miliard de cuvinte), reprezentând un instrument important pentru programele de dezambiguizare semantică și traducere automată.

4. Concluzii.

Inițiativa locală în ceea ce privește informatizarea cercetării românești, indiferent de domeniu, este un fenomen pozitiv și constructiv, dar este necesară o *corelare a rezultatelor locale* printr-un sistem eficient de publicare / comunicare a rezultatelor cercetării la toate nivelurile.

Domeniile de interes național, precum este cel al conservării și promovării *moștenirii culturale* – limba fiind una dintre cele mai importante forme ale acestei moșteniri, trebuie *susținute prin proiecte de mare anvergură și prin strategii naționale specifice* (după modelul marilor culturi).

Rezultatele cercetării sunt benefice întregii comunități științifice din domeniul filologiei române, din țară și din străinătate, constituind o cale de atingere a unor standarde de performanță competitive măcar la nivel european, precum și o mai bună cunoaștere a limbii și culturii noastre în lumea modernă.

Bibliografie

Cristea, Răschip *et alii*, 2007: Dan Cristea, Marius Răschip, Corina Forăscu, Gabriela Haja, Cristina Florescu, Bogdan Aldea, Elena Dănilă, *The Digital Form of the Thesaurus Dictionary of the Romanian Language*, în *Advances in Spoken Language Technology* (edit. Corneliu Burileanu, Horia-Nicolai Teodorescu), București, Editura Academiei Române, p. 195–206.

- Dănilă 2007: Dănilă, Elena, *Tradiție și inovație în cercetarea lexicografică românească în Evoluția și funcționarea limbii – perspective normative în noul context european*, Suceav, Editura Universității Suceava, p. 210–215.
- Haja, Dănilă *et alii* 2005: Haja, Gabriela, Dănilă, Elena, Forăscu, Corina, Aldea, Bogdan-Mihai, *Dicționarul limbii române (DLR) în format electronic. Studii privind achiziționarea*, Iași, Editura Alfa, publicat și electronic pe www.consilr.info.uaic.ro.
- Haja, Forăscu *et alii* 2006: Haja, G., Forăscu, C., Aldea, B. M., Dănilă, E., *The dictionary of Romanian Language: steps toward the electronic version*. in *Proceedings of EURALEX 2006*, Torino, Italy, september 2006.
- Haja 2007: Haja, Gabriela, *Resurse electronice pentru cercetarea lexicografică românească, în Limba română azi*, Iași, Editura Universității „Alexandru Ioan Cuza”, p. 129-134.
- Tușiș, Diaconu 1995: Tușiș, D., Diaconu, L., Barbu, A.M., Diaconu, C., *The Mac-ELU implementation of derivative morphology for Romanian*, Research Report, I.C. I, iunie 1995.

Contributions à l’informatisation de la recherche philologique roumaine: MLD. Biblia 1688 et eDTLR

On a commencé l’élaboration des programmes amples de réalisation de certaines bases de données et de certains instruments qui puissent soutenir la modernisation de la méthodologie de travail dans la philologie roumaine. L’innovation dans la recherche lexicographique roumaine est représentée justement par l’informatisation de ce type de démarche scientifique, ce que permettra la synchronisation de la recherche lexicographique roumaine avec la recherche similaire de tout le monde.

eDTLR représente une démarche essentielle de la linguistique roumaine et de la culture roumaine au cadre de l’ouverture et de la valorisation de celle-ci au contexte de la globalisation contemporaine.

Iași, România