Contribuţii la informatizarea cercetării filologice româneşti: MLD. Biblia 1688 şi eDTLR

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

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 şi a unor instrumente care să servească la modernizarea metodologiei de lucru. Acest lucru va permite sincronizarea cercetării lexicografice româneşti cu cercetări de acelaşi tip din întreaga lume.

Colaborarea cu specialişti de la Institutul de Informatică Teoretică al Academiei Române, începând cu anii ’90, şi de la Facultatea de Informatică, începând cu 2001. Proiectele de acest tip sunt facilitate şi de specializarea unor cercetători tineri în domeniul lingvisticii computaţionale, începând cu anul 2001 (anul creării unui astfel de masterat la Facultatea de Informatică din Iaşi). Tendinţa generală din ultima vreme vizează realizarea unor proiecte de cercetare interdisciplinară, de nivel naţional sau european, cu participarea unor specialişti din mai multe domenii: lingvistică, informatică, istorie literară, filosofie ş.a.

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

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

a) *Dicţionarul limbii române (DLR) în format electronic. Studii privind achiziţionarea* (cod CNCSIS 1815), grant finanţat de Ministerul Educaţiei şi Cercetării (MEC) prin Consiliul Naţional al Cercetării Ştiinţifice din Învăţământul Superior (CNCSIS), desfăşurat în perioada 2003–2005 la Institutul de Filologie Română „A. Philippide”. Prin acest proiect s-a verificat şi demonstrat posibilitatea transformării *Dicţionarului limbii române* din text tipărit în text electronic adnotat[1], prelucrat cu

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

BDD-V366 © 2009 Editura Alfa
Provided by Diacronia.ro for IP 216.73.217.42 (2026-11-15 19:44:51 UTC)

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

b) *Resurse lingvistice în format electronic: Monumenta linguae Dacoromanorum. Biblia 1688. Regum I, Regum II – Ediţie critică şi corpus adnotat. (MLD. Biblia 1688)* (cod CNCSIS 1454), desfăşurat în perioada 2006–2007. Prin acest proiect a fost găsită o posibilă metodă de achiziţionare în format electronic a unor cărţi vechi din Bibliografia DLR, cu aplicaţie asupra a două cărţi din *Biblia* tipărită la Bucureşti în anul 1688, *A împărăţiilor cea dentâiu, A împărăţiilor a doua,* precum şi crearea unor instrumente de indexare şi adnotare automată, la nivel de cuvânt, a textelor româneşti vechi.

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

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

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

2. MLD. Biblia 1688

În comunicarea de faţă detaliem, pentru început, unul dintre granturile care au premers, parţial, parteneriatului din cadrul proiectului complex *eDTLR. Dicţionarul Tezaur al Limbii Române în format electronic*, finanţat de CNMP, coordonat de Facultatea de Informatică din Iaşi, şi anume *Resurse lingvistice în format electronic: Monumenta linguae Dacoromanorum. Biblia 1688. Regum I, Regum II – Ediţie critică şi corpus adnotat (MLD. Biblia 1688)*.

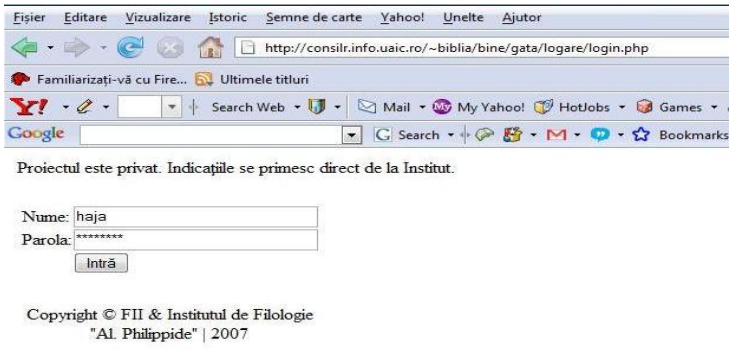
În cadrul proiectului *MLD. Biblia 1688* s-a realizat un *corpus* de mici dimensiuni de texte româneşti din secolul al XVII-lea, analizate morfologic, printr-un program de împărţire / analiză automată (corectat de lingvişti), ceea ce a dus, pe de o parte, la *generarea automată a unui indice de cuvinte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

adreseze unui public mai larg decât cel al specialiştilor propriu-zişi. Pentru detalii, vezi şi Haja, Dănilă *et alii*, 2005.

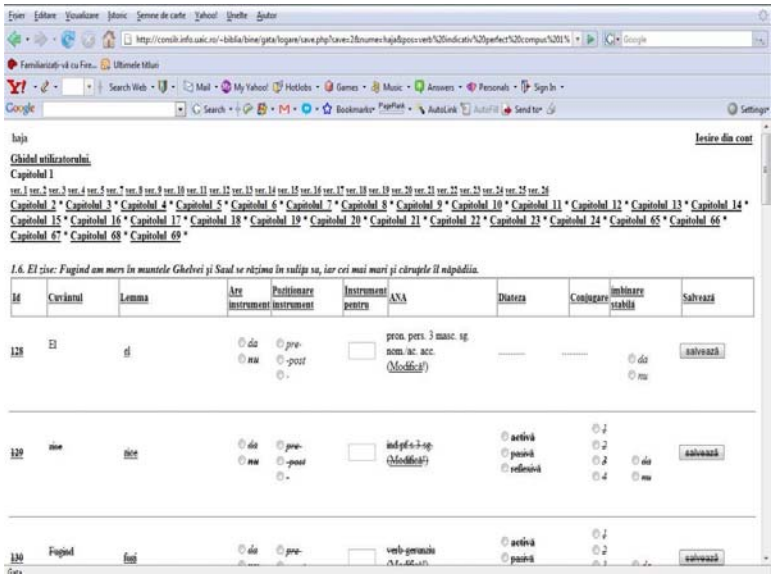
BDD-V366 © 2009 Editura Alfa
Provided by Diacronia.ro for IP 216.73.217.42 (2026-11-15 19:44:51 UTC)

automată, adecvat limbii române vechi şi la *dezvoltarea unei modalităţi de lucru la distanţă* prin intermediul unei interfeţe speciale, adaptabile, perfectibile, pentru validarea / corectarea prelucrării automate.

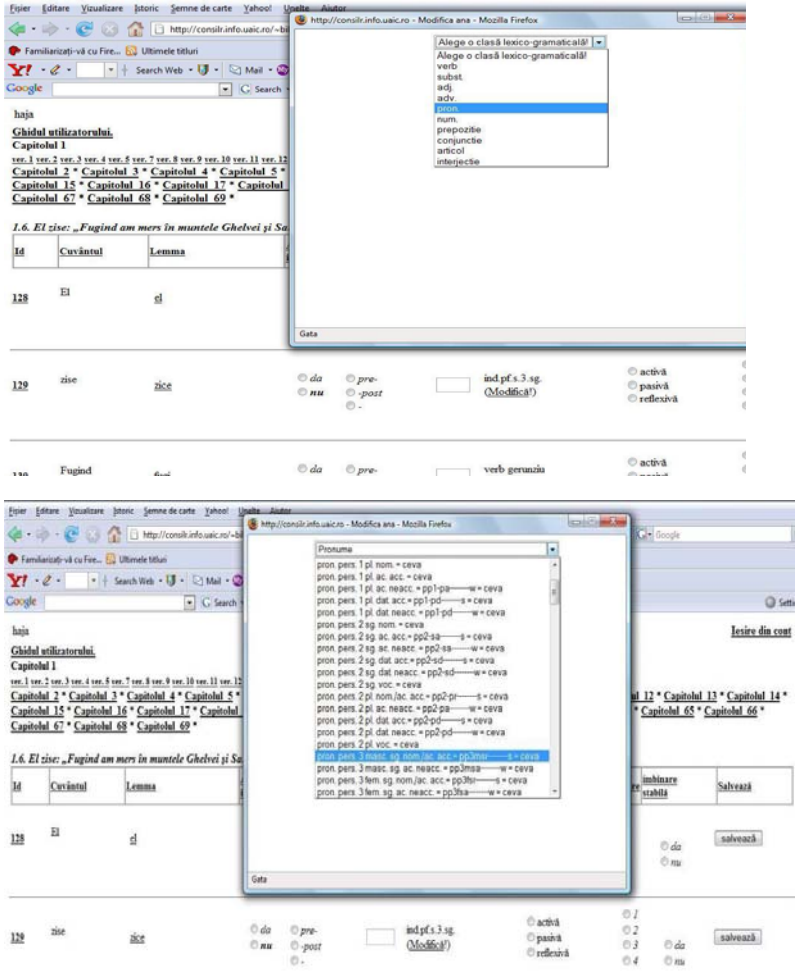
În continuare, prezentăm interfaţa de validare / corectare a analizei morfologice automate a textului Bibliei (sec. al XVII-lea). Prima captură de ecran reprezintă pagina web special creată pentru acest demers, fiecare dintre cercetătorii implicaţi având un cont propriu.



Următoarele capturi de ecran prezintă modalităţile de folosire ale acestei interfeţe de corectură, cu posibilităţile de optare pentru o anumită categorie gramaticală, pentru diferitele elemente de nuanţare a analizei lexico-gramaticale.



BDD-V366 © 2009 Editura Alfa
Provided by Diacronia.ro for IP 216.73.217.42 (2026-11-15 19:44:51 UTC)



Rezultatele proiectului s-au materializat într-un volum, al şaptelea al cunoscutei serii *Monumenta linguae Dacoromanorum* de la Iaşi, în forma consacrată şi în format electronic. Gradul de recunoaştere 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 situaţiile de omonimie. În privinţa 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 parţială a celei dintâi, rezultatele s-au ameliorat simţitor. Programul a recunoscut cu uşurinţă numele proprii, într-o primă fază fără a mai adăuga informaţii de natură morfologică. După prima corectură a specialiştilor lingvişti, analiza morfologică a numelor proprii s-a putut realiza, cu o rată mai mare de corectitudine. În ediţia de faţă, nu au fost indexate şi 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

BDD-V366 © 2009 Editura Alfa
Provided by Diacronia.ro for IP 216.73.217.42 (2026-11-15 19:44:51 UTC)

un demers ştiinţific de sine stătător, laborios şi 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 Patraş, masterand în anul I, în domeniul lingvisticii computaţionale, la Facultatea de Informatică din Iaşi, va putea fi antrenat şi perfecţionat pe măsură ce următoarele volume ale monumentalei ediţii vor fi pregătite pentru tipar. Mai mult, acest instrument va putea fi utilizat pentru multe dintre scrierile româneşti din secolul al XVII-lea, dar şi manuscrise şi cărţi din secolul următor, pentru care prima ediţie integrală a Bibliei a constituit un model de limbă românească.

3. eDTLR

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

În echipa de cercetare sunt implicaţi ca parteneri:

– Institutul de Lingvistică „Iorgu Iordan – Al. Rosetti”, Academia Română, Bucureşti. Responsabil de proiect acad. Marius Sala (prin dr. Monica Busuioc);

– Institutul de Filologie Română „A. Philippide”, Academia Română, Iaşi. Responsabil de proiect dr. Gabriela Haja;

– Institutul de Lingvistică şi Istorie Literară „Sextil Puşcariu”, Academia Română, Cluj-Napoca. Responsabil de proiect dr. Rodica Marian;

– Institutul de Cercetări pentru Inteligenţă Artificială, Academia Română, Bucureşti. Responsabil de proiect acad. Dan Tufiş;

– Institutul de Informatică Teoretică, Academia Română, Iaşi. Responsabil de proiect acad. Horia Neculai Teodorescu;

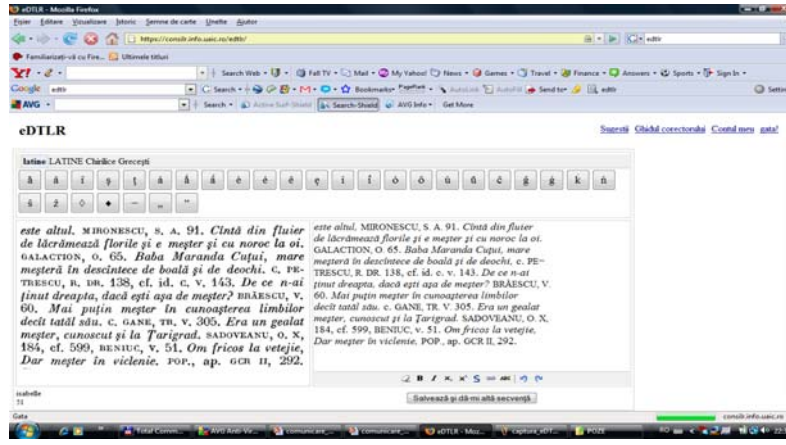
– Facultatea de Litere, Universitatea „Alexandru Ioan Cuza” din Iaşi. Responsabil de proiect dr. Eugen Munteanu.

Dintre obiectivele eDTLR până în acest moment au fost realizate: 1. scanarea şi convertirea în format editabil a tuturor volumelor DA şi DLR publicate până în prezent; 2. realizarea unei interfeţe de corectură on-line şi 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 interfeţe de corectură on-line, accesibilă lingviştilor lexicografi implicaţi în proiect, care vor corecta integral textul scanat şi convertit al Dicţionarului; 4. realizarea unui program de parsare a unui articol DLR şi de generare automată a ierarhiei sale semantice; 5. realizarea unui corpus scanat şi convertit într-un format accesibil pentru programul de indexare în surse a citatelor din Dicţionar, format din texte de referinţă ale DA şi DLR (potrivit legislaţiei în vigoare).

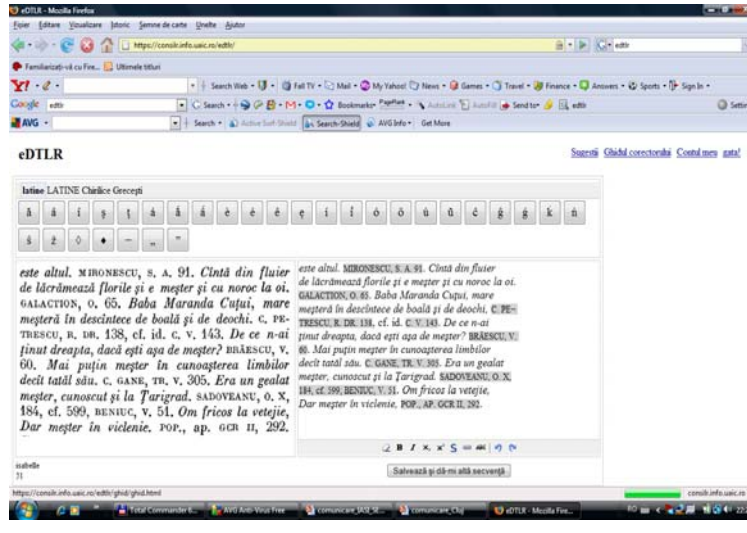
BDD-V366 © 2009 Editura Alfa
Provided by Diacronia.ro for IP 216.73.217.42 (2026-11-15 19:44:51 UTC)

Specialiştilor lingvişti le revine sarcina corecturii atente a textului obţinut în urma recunoaşterii optice a caracterelor din imaginile scanate ale Dicţionarului.

Corectura se face în trei etape: o primă etapă cu ajutorul corectorilor voluntari, nespecialişti şi specialişti, care vor face această operaţie on-line, pe un site special realizat de informaticienii implicaţi, urmată de o a doua etapă care presupune revizia corecturii făcute de voluntari de către specialiştii lexicografi. În literatura de specialitate, pentru activităţile colaborative, asemănătoare celor ce se preconizează a fi desfăşurate în proiect pentru corectarea primară a formatului electronic al Dicţionarului, există deja un nume: „crowdsourcing" sau „digital sharecropping"[2]. În captura de ecran următoare se poate vedea interfaţa de corectură on-line pentru voluntari.



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



[2] Vezi şi Cristea, Răschip *et alii* 2007: 195-206.

BDD-V366 © 2009 Editura Alfa
Provided by Diacronia.ro for IP 216.73.217.42 (2026-11-15 19:44:51 UTC)

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

Contribuţii ale proiectului eDLTR.

eDTLR va deschide noi modalităţi de lucru/studiu/cercetare în lexicografia românească, incluzând latura ei computaţională; va oferi singura cale modernă de completare şi aducere la zi a marelui dicţionar, ceea ce va duce, în viitor, în uniformizarea celor două serii ale Dicţionarului, DA şi DLR; va oferi posibilitatea de consultare interactivă a Dicţionarului 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, eDTLR va contribui semnificativ la promovarea limbii române.

Noutatea acestui proiect constă în faptul că eDTLR va fi primul dicţionar 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, eDTLR va fi cel mai mare dicţionar în format electronic din lume în ceea ce priveşte numărul de exemple care să susţină sensurile cuvintelor (estimăm că cele aproximativ 3.000 de volume în care vor fi indexate exemplele Dicţionarului conţin aproximativ un miliard de cuvinte), reprezentând un instrument important pentru programele de dezambiguizare semantică şi traducere automată.

4. Concluzii.

Iniţiativa locală în ceea ce priveşte informatizarea cercetării româneşti, indiferent de domeniu, este un fenomen pozitiv şi 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 naţional, precum este cel al conservării şi promovării *moşteniri culturale* – limba fiind una dintre cele mai importante forme ale acestei moşteniri, trebuie *susţinute prin proiecte de mare anvergură şi prin strategii naţionale specifice* (după modelul marilor culturi).

Rezultatele cercetării sunt benefice întregii comunităţi ştiinţifice din domeniul filologiei române, din ţară şi din străinătate, constituind o cale de atingere a unor standarde de performanţă competitive măcar la nivel european, precum şi o mai bună cunoaştere a limbii şi 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), Bucureşti, Editura Academiei Române, p. 195–206.

BDD-V366 © 2009 Editura Alfa
Provided by Diacronia.ro for IP 216.73.217.42 (2026-11-15 19:44:51 UTC)

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*, Iaşi, Editura Alfa, publicat şi 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*, Iaşi, Editura Universității „Alexandru Ioan Cuza", p. 129-134.

Tufiş, Diaconu 1995: Tufiş, 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.

Iaşi, România

BDD-V366 © 2009 Editura Alfa
Provided by Diacronia.ro for IP 216.73.217.42 (2026-11-15 19:44:51 UTC)