STATISTICI PARŢIALE LA ÎNCHEIEREA PROIECTULUI eDTLR – DICŢIONARUL TEZAUR AL LIMBII ROMÂNE ÎN FORMAT ELECTRONIC

DAN CRISTEA
Universitatea „Alexandru Ioan Cuza" din Iaşi
GABRIELA HAJA
Academia Română, Filiala din Iaşi
ALEX MORUZ
Academia Română, Filiala din Iaşi
MARIUS RĂSCHIP
Universitatea „Alexandru Ioan Cuza" din Iaşi
MĂDĂLIN PĂTRAŞCU
Universitatea „Alexandru Ioan Cuza" din Iaşi

1. Dicţionarul Academiei

Pe parcursul a mai bine de un secol, a fost elaborată sub egida Academiei Române[1] opera lexicografică cea mai importantă pentru cultura noastră, *Dicţionarul limbii române*, comparabil, prin dimensiuni şi perspectiva abordării lexicografice, cu lucrări similare din lexicografia europeană. O prezentare statistică a unor date generale puse paralel cu cele referitoare la alte două mari dicţionare este relevantă: *Dicţionarul limbii române* a fost redactat şi editat în două etape (seria cunoscută sub sigla DA, în perioada 1906–1944; seria nouă, DLR, în perioada 1965–2010), în 37 de volume şi cuprinde cca 175.000 de cuvinte şi variante, cu peste 1.300.000 de citate. Elaborarea variantei electronice s-a efectuat în intervalul 2007–2010. *Deutsches Wörterbuch* „der Grimm" (DWB), realizat şi publicat în perioada 1838–1961, în 32 de volume, cuprinde 350.000 cuvinte şi variante. Elaborarea variantei electronice s-a derulat între 1997 şi 2004. *Trésor de la Langue Française* (TLFi), cuprinzând lexicul scris al secolelor XIX şi XX, elaborat în perioada 1971–1994 (prima ediţie tipărită), în 16 volume, cuprinde 100.000 cuvinte, 270.000 definiţii, 430.000 exemple, conceput *ab initio* în variantă informatizată.

1.1 Dicţionarul limbii române (DA)

Dicţionarul limbii române, coordonat de Sextil Puşcariu între 1906–1944, a rămas cunoscut sub numele de *Dicţionarul Academiei* (DA). Elaborarea lucrării a fost încredinţată, după două încercări nefinalizate, unui colectiv mai larg, alcătuindu-se două echipe, alături de autorul principal, Sextil Puşcariu. În timp, însă, a rămas câte un lexicograf în fiecare echipă: C. Lacea şi, respectiv, Th. Capidan.

[1] La 1 aprilie 1866, se organizează „Societatea Literară Română", care, peste un an, devine „Societatea Academică Română" (Pascu 1991: 54–55). *Ortografia*, *Gramatica* şi *Dicţionarul* limbii române au constituit de la bun început repere ale existenţei Societăţii Academice Române, ceea ce a făcut ca Secţia filologică-literară să ocupe un loc privilegiat, atât prin numărul de membri, cât şi prin problemele dezbătute şi rezolvate (Pascu 1991: 105). Societatea Academică Română a fost declarată „Institut Naţional", sub numele de *Academia Română*, la 29 martie 1879.

Folosind fişierul preluat de la A. Philippide, precum şi mai multe zeci de mii de fişe extrase de Puşcariu şi echipa lui, s-a trecut la redactarea propriu-zisă, urmând ca îmbogăţirea materialului lexical să se facă paralel cu operaţia de redactare. Vreme de 38 de ani, din 1906 până în 1944, s-a realizat aproape jumătate din lucrare (3071 de pagini tipărite însumând literele *A-C, F-K* şi o bună parte din litera *L*, *l – lojniţă*, iar începutul literei *D* (*d – de*) a rămas în fază de tipar). Pentru întâia oară în istoria lexicografiei româneşti redactarea propriu-zisă este dublată de o preocupare lexicologică permanentă; de aceea, *Introducerea* semnată de Sextil Puşcariu la primul tom din DA constituie una dintre cele mai largi expuneri de motive din istoria lexicografiei române (Seche 1969: 37). Sextil Puşcariu prezintă principiile de alcătuire a listei de cuvinte, pornind de la dicţionarele anterioare precum şi de la extrasele din literatură, aparţinând unor texte cât mai variate, din epoci diverse. În privinţa criteriului de introducere a cuvintelor în lista dicţionarului, DA, ca dicţionar istoric şi general, include atât cuvinte vechi, populare şi regionale, cât şi împrumuturi recente care „exprimă o idee sau nuanţa unei idei pentru care limba noastră nu are un termen neechivoc" (DA 1913: XV). Numărul cuvintelor-titlu este mare: dintr-o statistică întreprinsă de S. Puşcariu rezultă că fasciculele cuprinzând patru litere, medii ca întindere (*A, B, F* şi *G*), însumează 15.444 de intrări. În DA elementele componente ale fiecărui articol au un loc stabil: după cuvântul-titlu şi indicarea categoriei lui morfologice urmează, în ordine, precizarea domeniului căruia îi aparţine cuvântul, traducerea sensurilor termenului românesc în limba franceză, indicaţii asupra situaţiei sale istorice sau asupra răspândirii diatopice ori indicaţii stilistice, definiţia, izvoarele ei, unităţile frazeologice, formele gramaticale, variantele lexicale, etimologia. De la litera *B* se schimbă modul general de organizare a materialului, adoptându-se metoda „cuiburilor" lexicale, potrivit căreia derivatele sunt subsumate cuvântului-titlu, figurând în acelaşi articol cu acesta, ceea ce determină simplificarea excesivă a analizei semantice a termenilor subsumaţi. Sensurile au fost despărţite în numeroase ramificaţii, cu cele mai diverse sisteme de izolare (cifre romane şi arabe, litere mari şi mici, bare duble şi simple etc.), împrumutate din lexicografia franceză. Este o premieră, în istoria lexicografiei româneşti, organizarea semantică a materialului, în mod sistematic, pe două linii dominante: prin *coordonare* şi prin *subordonare*. Unul dintre dezideratele coordonatorului, acela de a realiza, prin modul de explicare a termenilor, o operă lexicografică lingvistică şi nu una enciclopedică, a fost în bună măsură atins. Lista bibliografică a dicţionarului a fost mult îmbogăţită, ajungându-se de la cele 199 de izvoare folosite de A. Philippide, la circa 553 de titluri în 1926. DA se deosebeşte de versiunile anterioare şi prin faptul că numărul citatelor creşte în cadrul unui articol, potrivit unor necesităţi diferenţiate, până la câteva zeci (faţă de cele trei folosite de Philippide), astfel încât fiecare segment dintr-o epocă istorică este reprezentat prin texte (Seche 1969: 36-39). Autorul a acordat, însă, o mai mică atenţie neologismelor în ansamblul lucrării, potrivit concepţiei sale privitoare la neologizarea exagerată a limbii. În dicţionarul condus de Puşcariu este continuată linia lui Philippide cu privire la unităţile frazeologice, în sensul integrării lor în filiaţia semantică a articolului, subordonându-le direct, pe fiecare, sensurilor din care se dezvoltă.

1.2 Dicţionarul limbii române (DLR). Serie nouă

Lucrările la *Dicţionarul limbii române,* întrerupte brusc de război şi, mai cu seamă, de schimbarea regimului politic, au fost reluate, într-o nouă formulă, începând din 1949, după reorganizarea Academiei Române. S-a considerat util ca noua versiune a dicţionarului academic, fără a fi propriu-zis o continuare a celui anterior, să înceapă cu litera *M*, urmând ca, ulterior, după terminarea porţiunii până la *Z*, să fie elaborată, pe aceleaşi baze, întreaga porţiune de la *A* la *L*; obiectivul a fost atins prin volumele ce cuprind literele *D*, *E*, *K*, *L*, *Q*. Noua serie a dicţionarului se deosebeşte mult de cea imediat anterioară prin condiţiile în care este elaborat: trei echipe largi, la Bucureşti, la Cluj şi la Iaşi, care au lucrat paralel la redactarea Dicţionarului. La Bucureşti s-a efectuat, sub conducerea lui Iorgu Iordan, Alexandru Graur şi Ion Coteanu, revizia finală a materialului. Ulterior, revizia finală a fost făcută în fiecare dintre cele trei centre de cercetare, coordonarea lucrării fiind asumată, din 1997, de Marius Sala şi Gh. Mihăilă. O altă deosebire faţă de DA o constituie materialul lexical, astfel că, numai la litera *M*, care începe noua

versiune, se foloseşte o bibliografie cu 1627 de titluri. DLR înregistrează toate cuvintele atestate în limba literară generală şi în limbajul literaturii artistice, în vorbirea populară şi regională, în textele vechi; îşi propune (ca şi DA) să fie mai puţin deschis la cuvintele aparţinând terminologiei tehnico-ştiinţifice. În noua versiune, se aplică un criteriu riguros de selectare a neologismelor: au dreptul să figureze în dicţionar acei termeni ai limbajelor tehnico-ştiinţifice care au pătruns sau care manifestă tendinţa evidentă de a pătrunde în limba literară generală, în limbajul literar artistic şi în cel popular (DLR 1965: VI); hotărâtoare este aşadar atestarea circulaţiei unui cuvânt în cel puţin două dintre stilurile funcţionale ale limbii române. În DLR, sistemul de concentrare a termenilor în „cuiburi" lexicale a fost abandonat, fiecare lexem fiind înregistrat ca articol independent. Cele două opere lexicografice se deosebesc între ele şi prin modul de folosire a izvoarelor. Diferenţa cea mai evidentă rezultă din volumul diferit de extrase de care a dispus fiecare dintre lucrări; în plus, dimensiunea bibliografiei de referinţă a Dicţionarului a sporit de la un tom la altul, încât, dacă în 1965 erau citate 1627 de titluri, la finele primei ediţii numărul acestora ajungea la 2587. Fişarea s-a făcut continuu, pentru fiecare literă în lucru, în fiecare dintre cele trei centre. În DLR, ordinea tuturor izvoarelor ilustrative este strict cronologică, încadrându-se astfel în caracterul de bază al lucrării, acela de dicţionar istoric. În DLR alineatul final al articolului, cuprinzând etimologia, este semnificativ mai redus decât în versiunea anterioară, pentru că numeroase explicaţii de amănunt au fost lăsate pe seama *Dicţionarului etimologic al limbii române*, în curs de elaborare. Însă chiar DLR suportă adăugiri şi completări, de la un volum la altul, datorită faptului că multe neologisme, pătrunse în limbă de la începutul secolului al XX-lea, regionalisme incluse în *Atlasele lingvistice* sau în culegeri de texte recente, precum şi termeni vechi extraşi din documente aparţinând secolelor al XVI-lea – al XVIII-lea editate în ultimele decenii (Sala, Mihăilă 2000: VI) nu au fost cuprinse în primele volume ale noii serii şi lipsesc, parţial, şi din celelalte tomuri.

Redactarea intrărilor s-a făcut până în anii '90 în manieră tradiţională, textul scris de mână era dactilografiat şi corectat înainte de a fi trimis la editură, unde se culegea din nou, pentru a fi corectat şi în şpalturi. Abia din anul 1995 tehnoredactarea intrărilor a început să se facă pe calculator, utilizând programe de editare nespecializate (Microsoft Word). Textul acestor litere (*D, E, K, L, Q, V, W, X, Y, Z*) a fost trimis la editură pe suport electronic, acolo având loc (potrivit tradiţiei) ultimele corecturi, formatul final electronic fiind prelucrat printr-un program de editare (PageMaker).

Ca punct de plecare în achiziţionarea formatului electronic al marelui Dicţionar, s-a dispus aşadar de două tipuri de surse primare: un număr de volume, publicate înainte de 1997 (Haja et al. 2005: 23), care nu existau decât în forma tipărită şi restul de volume, pentru care se dispunea de formatul electronic ante-tipar, adică cel păstrat la sediile celor trei institute de lingvistică, în momentul predării materialului spre editare. Formatul electronic al variantelor finale, tipărite de Editura Academiei, nu a putut fi recuperat de la păstrătoarea acestuia.

2. Prezentarea proiectului eDTLR – Dicţionarul tezaur al limbii române în format electronic

În primăvara anului 2007, răspunzând apelului Ministerul Educaţiei, Cercetării şi Tineretului de formare de Parteneriate în Domenii Prioritare, se constituia un consorţiu din 7 parteneri aparţinând Universităţii „Alexandru Ioan Cuza" din Iaşi şi Academiei Române. Partenerii din cadrul Universităţii ieşene au fost Facultatea de Informatică şi Facultatea de Litere, iar din cadrul Academiei Române, Institutul de Lingvistică „Iorgu Iordan – Al. Rosetti", Bucureşti, Institutul de Filologie Română „A. Philippide", Iaşi, Institutul de Lingvistică şi Istorie Literară „Sextil Puşcariu", Cluj-Napoca, Institutul de Cercetări pentru Inteligenţă Artificială, Bucureşti şi Institutul de Informatică Teoretică, Iaşi. Aceştia au câştigat finanţarea unui proiect care s-a numit ***eDTLR – Dicţionarul tezaur al limbii române în format electronic.*** La începerea proiectului, în septembrie 2007, se mai lucra încă la finalizarea volumelor din marea serie a Dicţionarului (literele *D, E, K, L, Q*). În proiect s-a pornit însă de la premisa că activitatea de elaborare a Dicţionarului avea să se încheie înainte de terminarea proiectului eDTLR, ceea ce s-a şi împlinit (ultimul volum al seriei a fost tipărit în aprilie 2010).

Pe parcursul a 3 ani, proiectul eDTLR ar fi trebuit să materializeze următoarele obiective principale: **1.** realizarea unei variante electronice a marelui *Dicţionar Tezaur*; **2.** realizarea unei baze de date care să cuprindă izvoarele textuale ale Dicţionarului; **3.** realizarea legăturii dintre intrările Dicţionarului şi paginile de unde au fost excerptate exemplele; **4.** construirea unei interfeţe care să permită consultarea interactivă a Dicţionarului şi regăsirea exemplelor în contextele lor originare.

În ciuda mai multor dificultăţi, dintre care cele mai mari au fost generate de reducerea finanţării şi de procesul de adaptare a lingviştilor lexicografi la nou createle tehnologii şi metode de lucru, precum şi de găsirea unui limbaj comun de comunicare între specialiştii celor două domenii – informatică şi lingvistică, la terminarea lui, proiectul raporta următoarele realizări:

– un mediu complex de prelucrare a formatelor electronice ale Dicţionarului, care include: interfeţe de editare online a formei obţinute în urma OCR[2]-izării; programe de curăţare şi de conversie a formatului HTML, obţinut după editarea online, într-o variantă simplificată XML[3]; două variante (pentru formatul DA şi formatul DLR) ale unui program de recunoaştere a câmpurilor din intrările de dicţionar (parser) şi etichetarea lor corespunzătoare în XML utilizând convenţiile TEI-P5[4]; un lanţ de post-procesare care realizează o curăţare a codurilor XML rezultate în urma parsării; programe de actualizare a bazei de date a intrărilor de dicţionar şi de acces la baza de date; o interfaţă care dă acces utilizatorilor la eDTLR, permiţând căutări condiţionate şi generarea de statistici; programe de verificare a integrităţii surselor scanate; programe de segmentare a imaginilor paginilor scanate pentru recuperarea coordonatelor spaţiale ale fiecărui cuvânt în pagină; programe de regăsire a citatelor în formatele txt imperfecte ale surselor (*approximate string matching*) şi de indexare a citatelor în imaginile paginilor scanate;

– o bază de date a Dicţionarului, conţinând Dicţionarul, integral, în următoarele formate: **a.** pentru partea veche (secţiunile pentru care nu s-a dispus decât de suportul pe hârtie): imaginile paginilor scanate, formatul rtf[5] rezultat în urma OCR-izării, formatul HTML[6] rezultat în urma corectării, formatul XML TEI-P5 rezultat în urma parsării; **b.** pentru partea nouă (secţiunile pentru care s-a dispus de formatul electronic): formatul rtf originar al pre-tipăriturilor, formatul XML TEI-P5 rezultat în urma parsării; o bază de date a surselor bibliografice, conţinând lista tuturor surselor cu numele acestora, autorii, siglele şi cronologia lor;

– o bază de date a surselor bibliografice ale Dicţionarului, conţinând, pentru fiecare volum: forma imagine a volumului scanat şi forma rtf obţinută în urma OCR-izării lui (aşadar texte cu imperfecţiuni).

eDTLR a fost realizat în mai multe etape, rezumate mai jos în această secţiune.

2.1 Procesări preliminare aplicate volumelor existente doar pe hârtie

---

[2] OCR = *Optical Character Recognition*, operaţiunea de recunoaştere a codului caracterelor scrise de mână ori tipărite, dintr-o pagină scanată. În funcţie de calitatea hârtiei, de alfabet şi de limbă, recunoaşterea poate genera mai multe ori mai puţine erori.

[3] XML = *eXtensible Markup Language*, un limbaj de adnotare, cu aplicare la formate text, imagine, sunet şi multimedia.

[4] TEI – *Text Encoding Initiative*, un standard de adnotate a documentelor, cuprinzând un capitol special, relativ la dicţionare online. P5 este versiunea ultimă, cuprinzând şi o descriere a convenţiilor de marcare a intrărilor de dicţionar.

[5] Un format care cuprinde adnotări de formatare, utilizat curent de editoarele de texte, precum *Microsoft Word*, de exemplu.

[6] HTML – *HyperText Markup Language*, un format de adnotare utilizat cu precădere în realizarea paginilor Web.

Proiectul a început prin scanarea volumelor din seria veche a Dicționarului, care nu existau decât în formă tipărită. Fiecare imagine astfel obținută includea două pagini A4, care au fost separate, segmentate pe coloane, curățate de margini negre, de adnotări manuscrise marginale, de pete etc., reduse apoi de la 400 la 300 dpi, înainte de a fi OCR-izate. Dificultăți suplimentare la transpunerea în text a dicționarului au fost adăugate de calitatea inegală a tipăriturilor și de existența mai multor alfabete utilizate în redarea etimoanelor: latin, grecesc, chirilic, turcesc, polonez, rusesc etc.

Simultan, după o scurtă perioadă în care s-a actualizat fișierul surselor bibliografice, s-a început lucrul la scanarea surselor bibliografice, acest proces derulându-se apoi pe toată durata proiectului.

2.2 Faze de corectare

Întrucât operația de OCR-izare introduce erori, volumele Dicționarului care au suferit acest tip de prelucrare au necesitat corecții. În proiect s-a hotărât ca aceste volume să fie corectate în două etape, mai întâi de către novici, printr-un proces colaborativ (Cristea et al. 2008), iar ulterior de către experți lexicografi – parteneri în proiect. A fost creat un portal dedicat corecturii online[7] și, pe diverse căi, s-a adresat chemarea de a contribui la această acțiune[8].

Întrucât Academia Română nu putea permite, fapt lesne de înțeles, ca versiuni ale Dicționarului care ar fi putut fi extrase de pe site-ul proiectului să circule neautorizat, s-a adoptat soluția de a împărți fiecare pagină a Dicționarului în 12 segmente, doar un singur segment fiind accesibil pentru corectură la un moment dat, iar extragerea lor în secvență făcându-se aleatoriu. Prin această strategie, orice tentativă de a recompune chiar și o singură pagină a fost descurajată.

S-a dovedit că prima corectură a fost destul de inegală, o parte dintre experți raportând la corectura a doua un număr exagerat de erori găsite pe anumite porțiuni, când așteptările ar fi fost ca doar puține erori să mai fi rămas în urma corectorilor benevoli. Cu toate acestea, considerăm că antamarea acestui efort colaborativ a fost o alegere corectă a proiectului, pentru că, fără el, corectura a doua ar fi fost mult mai greoaie. La asamblarea finală a lucrării, s-a constatat totuși, în câteva locuri, lipsa unor secțiuni. Acestea nu puteau fi datorate decât unor omisiuni în scanările originale, care nu fuseseră verificate pagină cu pagină, așa cum s-a procedat ulterior cu sursele bibliografice. În consecință, în ultimele două luni ale derulării proiectului aproximativ 200 de pagini au fost introduse în baza de date a Dicționarului, direct în format HTML, de un număr mic de cercetători care s-au oferit să refacă aceste lipsuri.

Au participat la corectarea celor 11410 pagini, prelucrate după cum am descris mai sus, peste 900 voluntari, dintre care: 392 de curioși și 508, cu următoarea distribuție, după numărul de secvențe corectate: 127 au corectat între 10 și 50 de secvențe, 90 au corectat 50 – 100 de secvențe, 102 au corectat 100 – 200 de secvențe, 102 au corectat 200 – 500 de secvențe, 40 au corectat 500 –1000 de secvențe și 47 au corectat peste 1000 de secvențe.

2.3 Parsarea intrărilor

După aceste două faze de corectură, textul electronic al volumelor vechi ar fi trebuit să fie identic cu cel tipărit. Ultima fază în procesare, parsarea, urma să expliciteze semnificația câmpurilor intrărilor, operație extrem de dificilă pentru că presupune interpretarea marcajelor de formatare și de demarcare a câmpurilor în context. Respectând practica internațională, structura intrărilor a fost evidențiată în notații ale limbajului de marcare XML, cu aplicarea standardului utilizat în descrierea dicționarelor, TEI-P5. Parsarea a fost pregătită prin desemnarea categoriilor de marcaje utilizate în Dicționar pentru a pune în evidență diferite tipuri de câmpuri ale intrărilor. Aceste marcaje, clasificate pe niveluri, ajută parserul să descopere mai întâi un arbore general al sensurilor, pentru ca mai apoi fiecare nod al arborelui să fie detaliat suplimentar (Curteanu et al. 2008), rezultând o recunoaştere fidelă a

[7] https://consilr.info.uaic.ro/edtlr/

[8] Aproximativ 900 de profesori, doctoranzi, studenți de la universitățile din Iași, Suceava, Bacău, Baia Mare, Galați și altele, precum și din Republica Moldova au răspuns apelului nostru.

schemei semantice a fiecărui articol din Dicţionar, fiecare sens şi subsens fiind însoţit de şirul citatelor ilustrative cu precizarea izvoarelor.

Această tehnică are instanţieri diferite pentru volumele din seria DA şi cele din seria DRL. Mai mult decât atât, pentru că volumele seriei DLR provin, la rândul lor, din două surse diferite (cele obţinute prin OCR-izare din scanări şi corectate, şi cele recuperate din formate electronice originare), cele două serii de volume au necesitat prelucrări specifice de pregătire a parsării. Prelucrarea volumelor editate cu ajutorul calculatorului (seria nouă DLR, tipărită după 1997) a fost terminată înaintea volumelor editate cu mijloace vechi (identificând aproximativ seria veche DA). Diferenţele de structură dintre DA şi DLR ne-au obligat să elaborăm două versiuni diferite de parsere, unele dintre informaţiile codificate în varianta DA trebuind să rămână nespecificate în descrierea TEI finală.

Parsarea reprezintă cel mai complicat proces din întregul lanţ de prelucrare al Dicţionarului. În mod normal, programele de parsare sunt realizate cu ajutorul unor gramatici speciale, dependente de context, extrem de dificil de elaborat. De aceea, am optat pentru o soluţie care se bazează pe recunoaşterea şi ierarhizarea marcajelor delimitatoare ale câmpurilor din intrare, care a permis reducerea semnificativă a dimensiunii gramaticii şi elaborarea parserului într-un timp mult mai scurt decât cel necesar aplicării metodelor clasice. Cu toate acestea, acum, în momentul încheierii oficiale a proiectului, parserul nu e considerat încă finisat suficient. Erorile de structură găsite de noi au, în esenţă, trei surse: greşeli de dactilografiere a textului, încălcarea normelor de redactare la realizarea unui articol şi situaţii rare, încă netratate de către parser. Parsarea scoate la iveală imediat încălcarea normelor. Acestea produc fie structuri paradoxale, fie erori de parsare. Parserul a fost proiectat să aibă un comportament extrem de robust, o eroare de parsare afectând, în general, doar local structura. În astfel de situaţii, parserul îşi revine din starea de eroare, ca după o sincopă, primul indice semnificativ găsit din acel loc înainte fiind folosit pentru a repune parserul pe nivelul corect al arborelui de structură, procesul putând fi astfel continuat.

Iniţial, proiectul prevedea şi o fază de corectare a structurii (care nu a mai fost finanţată), în care erorile de structură ar fi trebuit observate şi remediate de către experţii lexicografi, capabili să interpreteze formatul XML şi să opereze modificări asupra lui. Soluţia pe care o gândim acum are în vedere o altă strategie, care ar presupune, din nou, o activitate benevolă a unei mase mari de utilizatori cunoscători, care se va limita, de astă dată, strict la *semnalarea* erorilor. Interfaţa de consultare a Dicţionarului incorporează deja o funcţionalitate de semnalare a erorilor: utilizatorul selectează o porţiune a textului pe care o consideră eronată şi, dacă doreşte, completează şi un comentariu. Un mesaj e-mail pleacă apoi spre server, care depozitează astfel de semnale ce urmează a fi tratate ulterior de experţi informaticieni şi lexicografi. Operaţiile de corectare nu vizează o corectare de suprafaţă a structurii generate greşit, ci, în funcţie de caz, fie o intervenţie asupra textului pentru corectarea unei erori evidente de dactilografiere, fie una asupra parserului însuşi. Reluarea parsării pentru intrarea respectivă va produce structura corectă.

2.4 Utilizarea surselor bibliografice ale Dicţionarului

Două au fost motivele care ne-au determinat să scanăm sursele bibliografice ale marelui Dicţionar. În primul rând, pentru că am dorit să legăm citatele din Dicţionar de imaginile paginilor de unde acestea au fost extrase. Utilizatorul interesat de ocurenţa cuvântului este acum în măsură să consulte contextul fiecărui citat ilustrativ, care, în funcţie de situaţia drepturilor de autor a sursei respective, poate fi egal ca întindere cu citatul însuşi ori nelimitat. În al doilea rând, pentru că sursele se constituie astfel într-o bibliotecă online care ar putea fi deschisă spre consultare interactivă lexicografilor, pentru căutarea de contexte ale cuvintelor. Biblioteca incorporează în prezent fişiere cu imaginile paginilor volumelor şi cu textele rezultate în urma OCR-izării. Variantele de text nu sunt corectate, pentru că nu ne-am propus acest lucru. Nici nu era posibil această operaţie, deoarece, baza textuală conţine 2.600 de volume (cărţi, reviste, dicţionare etc.) din cele aproximativ 4.000, câte are *Bibliografia* DLR. În prezent, sunt prelucrate complet şi accesibile doar 842 de volume, dar se lucrează în continuare la postarea în Biblioteca eDTLR a celor rămase. Numărul total de pagini scanate şi OCR-izate este de 972.000.

Pentru a contracara lipsa de acurateţe a textelor izvoarelor, obţinute după OCR-izare, s-au dezvoltat algoritmi de căutare aproximativă, care fac posibilă regăsirea citatelor, inclusiv în cazurile în care cuvântul căutat a fost OCR-izat imperfect. Utilizatorul are sub ochi doar pagina scanată, deşi interfaţa ascunde sub ea formatul text. Toate operaţiile de selecţie şi de copiere realizate de utilizatorul lexicograf se realizează asupra textului (aşa cum precizam, imperfect), însă confruntarea cu imaginea originalului aflată pe ecran lasă posibilitatea corectării eventualelor erori.

Dintre cele 2.600 de volume scanate, 2.148 fac parte din Biblioteca DLR a Institutului de Filologie Română „A. Philippide", 308 volume au fost scanate în cadrul Bibliotecii Centrale Universitare din Iaşi, unde imaginile şi textele rezultate sunt depozitate şi accesibile online, iar 144 au fost scanate în cadrul Institutului de Lingvistică „Iorgu Iordan – Al. Rosetti" din Bucureşti, care le-a pus la dispoziţia proiectului.

3. Analize statistice ale materialului din DLR

Consultarea Dicţionarului şi obţinerea statisticilor de mai jos este posibilă graţie interfeţei create în cadrul proiectului şi care oferă o serie de opţiuni pe care le prezentăm aici pe scurt. Aşa cum rezultă şi din nume, interfaţa este un instrument electronic ce permite legătura dintre utilizator şi imensa bază de date şi programe care este eDTLR. În prezent, sunt accesibile doar volume corespunzătoare seriei noi a Dicţionarului. Celelalte tomuri, inclusiv cele din seria DA există în baza eDTLR, asupra lor derulându-se procesul de prelucrare (parsare) a articolelor, acestea urmând a fi puse, treptat, la dispoziţia celor ce vor dori să le consulte.

Pe lângă vizualizarea informaţiei din *Dicţionar*, interfaţa permite o serie de statistici, operate pe întregul material lexicografic. În chip firesc, pe măsură ce se adaugă datele din volumele în curs de prelucrare, rezultatele analizelor statistice se modifică. Este permisă, în mod firesc, operaţia de consultare la nivel de literă ori de cuvânt.

Criteriile de consultare au fost definite în funcţie de normele de structurare lexicografică a Dicţionarului, aşa cum au fost ele integrate în parser. Pentru consultarea după criterii cronologice a fost necesară o prelucrare suplimentară a bazei de date eDLTR, în sensul adnotării fiecărui citat cu indicaţia anului, aşa cum a putut fi el extras automat, după mai multe elemente auxiliare, conţinute în *Siglarul* DLR, completat şi transpus în format XML. Potrivit *Siglarului* DLR, fiecare autor sau lucrare (dicţionar, periodic etc.) din Bibliografia DLR are atribuit un indice numeric ce precizează rangul său cronologic. În unele situaţii, nu este precizat acest indice cronologic, deoarece lucrarea respectivă este o colecţie de documente dintr-o perioadă istorică extinsă ori o publicaţie periodică; pentru aceste cazuri, s-a luat în considerare anul precizat între paranteze rotunde la finele citatului, înainte de siglă (ex. FURNICĂ, I. C.) sau anul din siglă (V. ROM. iunie 1970). În alte situaţii, indicelui numeric atribuit unui autor sau unei lucrări nu îi corespunde un an (ex. NEGRUZZI, A., NEGRUZZI, A. P.) şi atunci a trebuit să se găsească o rezolvare de compromis, prin echivalarea indicelui (124, în cazul nostru) cu un interval de ani. O a treia situaţie problematică, este aceea a citatelor populare din colecţii mixte, care cuprind texte literare şi populare, dintr-o perioadă istorică veche (ex. ŞIO). Citatelor populare extrase din aceste culegeri nu li s-a putut atribui niciun an, rămânând ca soluţie realizarea unui algoritm de deducere a cronologiei în funcţie de cronologia vecinătăţilor citatului.

Toate criteriile de consultare a Dicţionarului pot fi utilizate singular ori în combinaţii. Aceste formule de căutare pot fi salvate de utilizator, pentru căutări ulterioare.

Pentru a sugera posibilităţile de documentare pe care formatul electronic al Dicţionarului le oferă, am operat o serie de interogări după diverse criterii asupra a trei litere: E (5.885 de intrări redactate la Iaşi), L (6.039 de intrări redactate la Iaşi şi Cluj) şi S (14.347 de intrări redactate la Bucureşti).

Statisticile care urmează sunt realizate la nivel de intrare, urmărindu-se situaţia cuvintelor în contexte, lucru care poate antrena concluzii de ordin semantic şi funcţional.

Criterii			E		L		S	
Criterii morfologice	substantive	masculine	616	10,48%	401	6,64%	2140	14,92%
		feminine	2269	38,56%	1278	21,16%	5209	36,31%
		neutre	1093	18,57%	592	9,80%	2395	16,69%
	verbe		544	9,24%	482	7,98%	1731	12,07%
	adjective		1151	19,56%	621	10,28%	2495	17,39%
	adverbe		82	1,39%	80	1,32%	117	0,82%
Criteriu etimologic	franceză		2439	41,44%	1008	16,69%	1388	9,67%
	latină		612	10,40%	344	5,70%	453	3,16%
	albaneză		1	0,02%	5	0,08%	9	0,06%
	slava veche		6	0,10%	15	0,25%	71	0,49%
	maghiară		24	0,41%	100	1,66%	63	0,44%
	turcă		49	0,83%	44	0,73%	160	1,12%
Criteriul circulației	învechit		518	8,80%	802	13,28%	1870	13,03%
frecvență	rar		279	4,74%	304	5,03%	1046	7,29%
diacronie și frecvență	învechit+rar		33	0,56%	74	1,23%	249	1,74%
spațiu	regional		1	0,02%	655	10,85%	1367	9,53%
	și Moldova		1	0,02%	107	1,77%	236	1,64%
	numai Moldova		0	0	44	0,73%	105	0,73%
	numai Transilvania		9	0,15%	86	1,42%	202	1,41%
Îmbinări stabile	intrări sub care apar expresii		40	0,68%	177	2,93%	336	2,34%
Compuse	intrări sub care apar cuvinte compuse		0	0	42	0,70%	62	0,43%
După autori	CORESI		18	0,31%	120	1,99%	267	1,86%
	EMINESCU		421	7,15%	356	5,90%	427	2,98%
	PREDA		194	3,30%	201	3,33%	498	3,47%
După cronologie	înainte de 1600		2	0,05%	7	0,12%	13	0,09%
	între 1700 și 1800		4	0,07%	8	0,13%	22	0,15%
	între 1800 și 1900		70	1,19%	139	2,30%	352	2,45%
	după 1900		451	7,66%	421	6,97%	979	6,82%

Procentele au fost calculate raportând numărul de intrări rezultat în urma căutării la numărul total de intrări pe fiecare literă. Criteriul morfologic poate fi exploatat mai fin, dar am optat aici pentru căutarea după doar patru părți de vorbire (substantiv, verb, adjectiv, adverb) distingând, pentru substantive, categoria gramaticală a genului. Suma rezultată şi lista cuvintelor cuprind acele intrări sub care se află, cel puţin o dată, una dintre aceste folosiri (încât, de exemplu, un substantiv care are forme de masculin şi de feminin va apărea în două liste). Prima observaţie evidentă, în urma comparării ponderii părţilor de vorbire analizate în cele trei litere, este aceea că în lista de cuvinte domină numeric şi proporţional substantivele, iar, dintre acestea, substantivele feminine sunt cele mai numeroase. De asemenea, este interesant de analizat faptul că o literă de dimensiuni reduse spre medii, precum litera *E* depăşeşte, în privinţa ponderii substantivelor, o literă de mari dimensiuni cum este *S*.

La criteriul etimologie valorile raportate cuprind cuvinte cu etimon cert, unic, sau cu etimologie multiplă; sunt luate în considerare şi etimoanele complexe, conform cărora un cuvânt se formează pe teren românesc după un model străin. Interfaţa nu face, deocamdată, distincţie între cuvintele de origine

latină împrumutate şi cele moştenite. Rezultatele obţinute în urma căutărilor după criteriul etimologic demonstrează importanţa influenţei limbii franceze asupra lexicului românesc, începând cu secolul al XIX-lea, când intră în circulaţie majoritatea neologismelor cu această origine. Fireşte, şi în cazul căutărilor după criteriul etimonului, rezultatele obţinute trebuie înţelese astfel: litera *E* cuprinde 2439 de cuvinte care provin numai sau şi din limba franceză. Semnificativ este faptul că aproape jumătate dintre cuvintele literei *E* sunt (şi) de origine franceză (41,44%), în vreme ce pentru celelalte litere se schimbă raportul. Diferenţele au explicaţii lingvistice, dar şi conjuncturale (anul / intervalul redactării şi al publicării fiecărei litere). Aceste rezultate pot fi coroborate cu cele obţinute în urma căutării după vechimea primei atestări, cu cele obţinute în urma căutării regionalismelor ori a sensurilor regionale şi cu cele privitoare la numărul de cuvinte care au dezvoltat expresii. Remarcăm că îmbogăţirea lexicului limbii române a crescut semnificativ după 1900, dar această concluzie este determinată şi de faptul că numărul izvoarelor textuale editate după 1900, incluse în Bibliografia DLR, este mare, iar fişierul extras pentru această perioadă este foarte bogat. Din analiza datelor numerice şi a proporţiilor rezultate se impune, cu evidenţă, caracterul neologic al literei E, prin comparaţie cu celelalte două, mai echilibrate din punctul de vedere al distribuţiei lexicului, după vechime, spaţiu, uz etc.

Fie şi parţiale, aceste rezultate şi analiza lor sumară oferă specialiştilor interesaţi de studiul lexicului limbii române, aşa cum este el analizat şi prezentat în Dicţionar, câteva repere pentru modalitatea de consultare şi de valorificare a bazei eDTLR.

4. Concluzii

Așa cum era de așteptat, semnificația ieșită din comun pentru cultura românească a marelui Dicționar tezaur al limbii române și apariția formatului lui electronic, face acest produs extrem de dorit publicului vorbitor de limba română (incluzând aici toate categoriile de utilizatori, de la oameni obișnuiți care doresc să se informeze asupra sensului unui cuvânt, până la cercetători ai limbii). Membrii consorțiului eDTLR au avut semnale repetate care arată nerăbdarea cu care este așteptată apariția nerestricționată a eDTLR în internet.

Este de necontestat că eDTLR trebuie să devină un bun public din mai multe motive, cel mai important fiind acela că atât realizarea conținutului cât și realizarea formatului digital au fost plătite din bani publici. Pe de altă parte, finanțarea lui parțială, sub nivelul promis la momentul încheierii contractului, a împiedicat ducerea la îndeplinire a tuturor etapelor de corectare prevăzute, iar ieșirea în spațiul public a unui bun imperfect, sub girul Academiei Române, nu este o opțiune plauzibilă. Drept urmare, intenționăm ca pentru o perioadă, atâta vreme cât se vor mai constata erori de conținut ori structură, eDTLR să fie deschis spre consultare numai specialiștilor, într-o etapă ulterioară *Dicționarul* urmând a fi deschis spre consultare publicului larg.

Mai sunt, de asemenea, de rezolvat câteva probleme de suport tehnologic. Accesul la baza de date în căutare angrenează procese de căutare sofisticate, mari consumatoare de timp. Nu știm cum se va comporta serverul atunci când un număr mare de utilizatori vor face interogări în același timp. Vor fi de considerat soluții de distribuire a bazei de date și de paralelizare a calculelor. Amintim totodată și imensa responsabilitate pe care gestionarea acestei importante baze de date o incumbă, ea trebuid să fie intangibilă la atacuri maligne ori la tentative de scurgeri de informații.

5. Mulțumiri

Proiectul eDTLR a fost finanțat prin contractul CNMP nr. 91_013/18.09.2007. Mulțumim membrilor consorțiului eDTLR, fără a căror dedicație proiectul nu s-ar fi realizat. Adresăm mulțumiri, de asemenea, sutelor de contribuabili (de cele mai multe ori anonimi) care au ajutat la corectarea surselor, utilizând interfața online. Cu prisosință, merită toată recunoștința noastră firma PIM SRL Iași și directorul acesteia d-l Marius Petrariu, care, înțelegând importanța activității noastre și resursele limitate cu care suntem obligați să o realizăm, nu numai că ne-a oferit un preț extrem de convenabil pentru scanări, dar a acceptat să mute, pentru o perioadă de câteva luni, un scanner și să delege o persoană care să se ocupe

exclusiv de scanarea surselor aflate în sediul Bibliotecii Universitare Iaşi. Mulţumim conducerii acestei biblioteci, pentru că a acceptat să ne ofere, pentru a fi scanate la sediul lor, un număr de 308 volume care nu se află sub incidenţa Legii Patrimoniului Naţional şi nici a Legii Drepturilor de autor[9]. Mulţumim doamnei Cătălina Mărănduc pentru cele 144 de volume care au fost scanate de dânsa la sediul Institutului de Lingvistică din Bucureşti şi oferite consorţiului. Mulţumim următorilor cercetători care au acceptat să introducă în baza de date porţiunile care lipseau: Cristina Florescu, Elena Dănilă, Laura Manea, Mioara Dragomir şi Monica Corodeanu, toţi de la Institutul Philippide Iaşi, precum şi Daniela Gîfu şi Iulia Scutariu de la Facultatea de Informatică.

Bibliografie

1. Cristea, Dan, Corina Forăscu, Marius Răschip, Michael Zock, 2008, „How to Evaluate and Raise the Quality in a Collaborative Lexicographic Approach", în *Proceedings of LREC-2008*, Marrakech.
2. Curteanu, Nicolae, Alex Moruz, Diana Trandabăţ, 2008, „Extracting Sense Trees from the Romanian Thesaurus by Sense Segmentation & Dependency Parsing", în Proceedings of CogAlex Cognitive Aspects of the Lexicon: Enhancing the Structure, Indexes and Entry Points of Electronic Dictionaries, COLING 2008, pp. 55–63, ISBN 978-1-905593-56-9
3. Haja, Gabriela, Elena Dănilă, Corina Forăscu, Bogdan-Mihai Aldea, 2005, *Dicţionarul limbii române (DLR) în format electronic. Studii privind achiziţionarea*, Iaşi, Editura Alfa.
4. Iordan, Iorgu, Al. Graur, Ion Coteanu (coord.), 1965, *Dicţionarul limbii române* (DLR). Serie nouă. Tomul VI, *Litera M*, Bucureşti, Editura Academiei R.P.R.
5. Pascu, Ştefan, 1991, *Istoricul Academiei Române. 125 de ani de la înfiinţare*, Editura Academiei Române, Bucureşti.
6. Puşcariu, Sextil, *Dicţionarul limbii române* (DA), 1913, Întocmit şi publicat după îndemnul şi cu cheltuiala Maiestăţii Sale regelui Carol I, Tomul I, Partea 1, *A–B*, Bucureşti, Librăriile Socec & Comp. şi C. Sfetea.
7. Sala, Marius, Gheorghe Mihăilă, 2000, *Cuvânt înainte*, în *Dicţionarul limbii române (DLR)*. Serie nouă. Tomul XIV, *Litera Z*, Bucureşti, Editura Academiei Române.
8. Seche, Mircea, 1969, *Schiţă de istorie a lexicografiei româneşti*, vol. II, Bucureşti, Editura Ştiinţifică.

PARTIAL STATISTICS AT THE END OF THE PROJECT "THESAURUS DICTIONARY OF THE ROMANIAN LANGUAGE IN ELECTRONIC FORM"

Abstract

The paper describes the project aimed at creating the electronic form of the big Dictionary of Romanian Language, an outstanding work elaborated by the Romanian Academy, all along a century. A short comparison with other similar dictionaries is first presented. Then, the two series of the Dictionary are described, the old series, elaborated before the second war and known as the Dictionary of the Academy, and the new series, elaborated between 1965 and 2010, known as the Dictionary of the Romanian Language. The methodology of building the electronic form of the Dictionary is then presented: scanning the volumes of the Dictionary which were found only on paper support, OCR-ising them and correcting the resulting editable form on two passes – one performed by novices, in a collaborating activity, the second operated by experts. The corrected manuscript as well as the newer volumes, originally typed in before printing, were passed to a parsing process which produced the final XML form encoding the structure of the entries. This form is deposited in a large database and can be accessed by the user through an online interface. To show the searching and interpretation possibilities opened by this technology, a number of statistics are presented, as an exercise, on three letters: E, L and S. Finally, come conclusions are drawn.

[9] Conform convenţiei încheiate între Universitatea din Iaşi şi Biblioteca Universitară din Iaşi, aceste surse se găsesc depozitate pe un server aflat la sediul BCU-Iaşi, de unde sunt accesate prin internet.