

Dicționarul limbii române (DLR) în format electronic. Aplicații

Elena DĂNILĂ*, Gabriela HAJA*
Bogdan-Mihai ALDEA**, Corina FORĂSCU**

1.1. Premise

În epoca actuală, a integrării europene a României, pe fondul globalizării tot mai accentuate, tehnologia limbajului uman este esențială în promovarea diversității lingvistice și culturale în Comunitatea Europeană. Păstrarea identității limbilor și a culturilor naționale în cadrul societății informaționale globale a fost încă din anii '90 o realitate, așa cum se arată în avertismentul lui Alain Danzin: „În era electronică, este esențial pentru supraviețuirea unei limbi ca ea să fie folosită în sistemele de informare electronică”¹. Procesarea electronică a unei limbi a devenit de maximă importanță în condițiile evoluției societății moderne, în care comunicațiile prin medii electronice le concurează pe cele tradiționale. Procesarea automată a unei limbi și, în mod special, posibilitățile de traducere automată asigură compatibilizarea unei limbi utilizate de o comunitate națională, medie ca dimensiune, cu limbile de circulație internațională și menținerea identității lingvistice, în condițiile apropierei de o societate fără granițe. Punerea la dispoziție în limba nativă a informației din mediul internațional, precum și transmiterea într-o limbă de circulație mondială a informațiilor naționale au devenit demult o responsabilitate națională². Acest lucru contribuie la înlăturarea barierelor naturale, naționale, culturale și lingvistice, furnizând acces direct la informație în limba maternă a utilizatorilor. Implicit, specificul românesc poate deveni astfel accesibil oricărui dintre locuitorii planetei care are acces la Internet.

În contextul de „unitate prin diversitate” păstrarea limbilor și culturilor actuale se bucură de un interes deosebit, fiind stimulată puternic de organisme internaționale, precum Comisia Europeană³. Mai mult, ținând seama de situația particulară a limbii române, este evident că orice efort în această direcție contribuie, în final, și la îmbunătățirea comunicării între toți vorbitorii de limbă română, inclusiv a celor din afara granițelor (răspândiți în țări precum Republica Moldova, Ucraina, Ungaria, Serbia, Bulgaria, alte țări ale Comunității Europene și ale continentului american), așa cum se dorește și prin actualul program guvernamental: „Guvernul României are datoria de a acționa pentru sprijinirea dezvoltării și exprimarea identității culturale a românilor din afara granițelor României”⁴.

* Institutul de Filologie Română „A. Philippide”, Iași

** Facultatea de Informatică, Iași

¹ A. Danzin, *Towards a European Language Infrastructure*, CEC Doc. 54210/92, 1992, 62 p.

² W. Teubert, *Language Resources for Language Technology*, în Dan Tufiș and Poul Anderson (eds.), *Language and Technology*, Editura Academiei, București, 1997.

³ D. Tufiș, D. Cristea *RO-BALKANET – ontologie lexicalizată, în context multilingv, pentru limba română*, în Dan Tufiș și Fl. Gh. Filip (eds.), *Limba Română în Societatea Informațională - Societatea Cunoașterii*, Editura Expert, București, 2002.

⁴ *Programul de Guvernare*, Capitolul 24 - *Politica în domeniul relațiilor cu românii de pretutindeni*, <http://www.gov.ro/obiective/afis-docdiverse-pg.php?iddoc=268>.

Atingerea acestui scop, în contextul societății informaționale actuale, se realizează numai printr-o strânsă colaborare între lingviști și informaticieni. Versiunea electronică a DLR-ului *integral* – obiectiv de viitor, pentru care proiectul de față înseamnă punctul de plecare și baza de la care se va porni – va oferi lumii, și mai ales comunității românilor de pretutindeni, tezaurul limbii române, precum și mai multe posibilități pentru aplicațiile de procesare a limbajului uman pentru limba română, incluzând traducerea automată și cea asistată de calculator, acces multimedia inteligent la Internet, regăsirea informației (prin căutarea unor cuvinte-cheie sau concepte), extragerea informației (legate de un anumit subiect, deci de un anumit cuvânt-cheie), sumarizare automată, dezambiguizare semantică, achiziția de cunoștințe etc.

Ideea proiectului *Dicționarul limbii române (DLR) în format electronic* se integrează într-un program amplu de informatizare a cercetării lingvistice românești, coordonat în prezent de Comisia de Informatizare pentru Limba Română din cadrul Academiei Române, Secția de Știința și Tehnologia Informației, subsumându-se totodată obiectivelor prioritare privind informatizarea ale Ministerului Educației și Cercetării. Scopul Comisiei de Informatizare pentru Limba Română îl constituie apărarea identității limbii române prin promovarea studiilor dedicate ei dintr-o perspectivă informațională⁵.

Prin sprijinul financiar obținut de la MEC prin intermediul CNCSIS (Consiliul Național al Cercetării Științifice din Învățământul Superior), s-au studiat modalitățile de achiziționare în format electronic a *Dicționarului limbii române*, în cadrul unui proiect desfășurat în perioada 2003 – 2005. Pe lângă atingerea obiectivului central, s-au putut crea și o serie de aplicații ce facilitează utilizarea informatizată a materialului existent în DLR.

În proiect au fost implicați tineri cercetători din domeniul lexicologiei-lexicografiei și din domeniul lingvisticii computaționale, care au format o echipă omogenă, datorită faptului că dificultățile de comunicare au fost depășite prin specializarea interdisciplinară.

Ca formulă de citire și convertire a DLR în format XML s-a optat pentru utilizarea unei euristici specifice DLR, care permite recunoașterea, pe baza caracteristicilor formale ale textului, a unor câmpuri de text, fiecare cu semnificație bine determinată. Realizarea obiectivului inițial, care se referea la definirea unei gramatici lexicografice proprii DLR, a fost amânată; atingerea acestui obiectiv s-a dovedit imposibil de realizat în condițiile date (restricțiile de timp, mijloacele tehnice avute la dispoziție), rămânând ca, după ce vom finaliza achiziționarea DLR cu ajutorul *DLRex* – instrumentul de achiziționare, prelucrare și consultare a DLR, creat în cadrul acestui proiect –, să fie semnalate toate problemele privitoare la normele DLR în vederea rediscutării și adaptării acestora condițiilor de formalizare optimă. Ulterior, va fi posibilă definirea unei gramatici cu un număr rezonabil de reguli, astfel încât funcționalitatea și randamentul să fie îmbunătățite.

1.2. Etapele de lucru

Trecerea DLR din forma tipărită în format electronic a presupus câteva etape:

- scanarea unui eșantion de pagini tipărite din fiecare fasciculă a ediției DLR;
- convertirea imaginii digitale în format .doc; am optat pentru formatul RTF (Word), din motive pe care le vom dezvolta *infra*;
- corectarea manuală a erorilor apărute în urma scanării și a convertirii;
- procesarea textului cu ajutorul *DLRex* și transpunerea lui în format XML.

⁵ D. Cristea, D. Tufiș, *Resurse lingvistice românești și tehnologii informatice aplicate limbii române*, în Ofelia Ichim și Florin-Teodor Olariu (eds.), *Identitatea limbii și literaturii române în perspectiva globalizării*, Academia Română, Institutul de Filologie Română „A. Philippide”, Editura Trinitas, Iași, (2002).

Odată parcurse aceste etape, am obținut fișier XML⁶ care cuprinde peste 400 de pagini tipărite din DLR. Pentru parcurgerea tuturor etapelor necesare procesării unei pagini tipărite au fost necesare, în medie, 45 de minute. Cea mai costisitoare operație, din perspectiva timpului necesar, s-a dovedit a fi aceea de corectare. Numărul mare de erori a fost determinat de calitatea hârtiei și a tiparului din volumele editate înainte de 1990, de performanțele tehnice ale scannerului și ale programului de recunoaștere / convertire a imaginii digitale în text. De aceea, pentru ameliorarea vitezei de lucru, este necesară îmbunătățirea instrumentelor tehnice, utilizarea unor programe de convertire adaptate recunoașterii simultane a mai multor limbi, inclusiv a celor ce nu utilizează litere latine, precum și accesul la textele tehnoredactate computațional după 1990, pentru a se elimina etapele scanării, convertirii și, parțial, a corectării, în cazul acestora.

Chiar în condițiile actuale, este încurajator faptul că s-ar putea realiza achiziționarea integrală a DLR într-un timp real acceptabil, respectiv cca. 5300 de ore, ceea ce reprezintă, raportat la o normă de opt ore, treizeci de luni de lucru pentru o singură persoană.

În condițiile optimizării condițiilor de lucru, am apreciat că timpul s-ar reduce semnificativ: ar fi necesare cca 3000 de ore, adică optsprezece luni.

2.1. Limbajul de programare folosit și formatul fișierelor de intrare

Ideea de bază a aplicației constă în trecerea DLR-ului în format electronic, și anume în format XML. În urma scanării fișierelor de dicționar, a convertirii acestora de către OCR⁷ și a corectării de către lingviști lexicografi (care sunt familiarizați cu formatul standard al DLR-ului) au rezultat fișiere Word cu extensia .doc, deoarece formatarea textului este cea care ne ajută să realizăm trecerea dicționarului în format XML.

Încă de la început s-a încercat eliminarea eventualelor pași intermediari în pregătirea fișierelor de intrare în vederea procesării de către aplicație, motiv pentru care s-a căutat un limbaj de programare care să ne faciliteze lucrul cu astfel de fișiere. În prima fază s-a încercat scrierea aplicației în Delphi 7⁸, pentru că acesta are integrate componente care lucrează (realizează citirea și scrierea) direct din fișierul .doc. S-a constatat că la citirea dintr-un fișier .doc, programul Delphi 7 ține cont și de formatul textului (face diferențele dintre fragmentele scrise în format normal, italic, bold etc.). Fiecare exemplu pentru atestarea unui sens sau subsens al unui cuvânt conține și o siglă, adică o referire către bibliografie. Aceasta este scris cu *smallcaps* (SMALLCAPS). Din păcate Delphi 7 nu poate păstra această formatare la citirea din fișier, interpretând astfel siglele ca și cum ele ar avea stilul fontului normal.

Deoarece sigla este foarte importantă, iar recunoașterea ei în dicționar ar ajuta la parsarea⁹ acestuia, s-a optat pentru JAVA ca limbaj de programare, deoarece acesta

⁶ XML (<http://www.w3.org/XML/>) – *Extensible Markup Language* – este un standard de adnotare a documentelor ce conțin informație structurată.

⁷ Un OCR, *Optical Character Recogniser*, permite transformarea imaginilor (texte scrise pe hârtie, captate de un scanner) în text editabil pe computere.

⁸ <http://www.borland.com/delphi/>.

⁹ Parsarea, din punct de vedere computațional, este procedeul de analizare a unui șir de caractere (uzual text) pentru a determina structura gramaticală a acestuia conform cu un formalism gramatical. În general, parsarea se face în două etape: identificarea *token*-ilor – elementele primitive ale unui text structurat – și construirea arborelui de parsare care captează ierarhia implicită a datelor de intrare.

lucrează mai ușor cu *String*-urile (șiruri de caractere) și prelucrarea acestora este mai facilă în acest mediu de programare. Pentru că JAVA citește foarte greu din binar (modul în care sunt salvate în fișiere documentele Word), s-a optat pentru ca fișierele de intrare să fie fișiere Word, dar de această dată să fie salvate în format RTF (Rich Text Format).

2.2. Modelul

La baza parsării DLR-ului stă atât formatarea textului, cât și prezența unor simboluri speciale, cum ar fi \blacklozenge = romb plin, \lozenge = romb gol (etc.). O intrare din dicționar păstrează, în linii mari, formatul ilustrat în Figura 1.

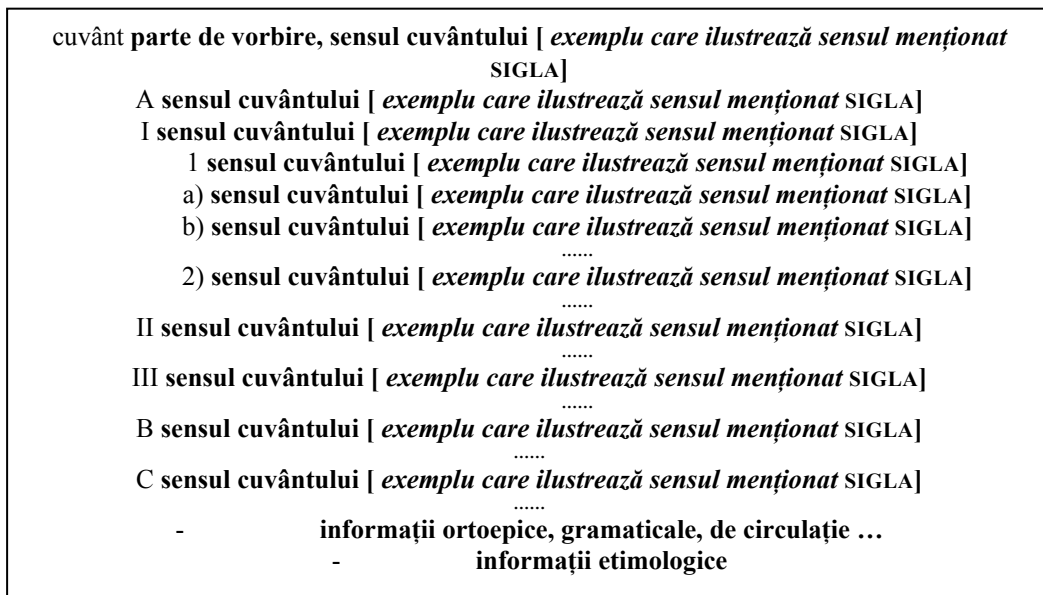


Figura 1. Schema generală a unei intrări în DLR

Apariția unui simbol special reprezintă de fapt introducerea unui nou sens al cuvântului (rombul gol introduce un sens care este apropiat de sensul sub care se află sau un anumit regim gramatical, iar romb plin introduce un sens mai depărtat). Ambele simboluri pot apărea oriunde sub un sens din schema unei intrări prezentate mai sus, nefiind specificată o anumită regulă de apariție a acestora.

După citirea unei intrări din dicționar s-a construit un vector în care este pus (în ordinea citirii din fișier) fiecare fragment ce are o formatare diferită față de fragmentul ce îl precedă și, respectiv, cel de după el, astfel că o parsare a vectorului, ținând seama de modul în care este scrisă o intrare din dicționar, ar putea duce la formarea fișierului XML dorit. După acest pas s-a observat că, deși unele fragmente erau în text unul lângă celălalt și aveau și aceeași formatare, ele erau salvate în interiorul fișierului RTF ca având formatari diferite. Altă anomalie observată o constituie unele caractere care nu își păstrează formatarea. Acestea au apărut în urma scanării și nu au putut fi sesizate cu ochiul liber de către persoanele care au făcut corecția manuală a fișierelor de intrare. Astfel de caractere sunt “ ” (spațiul), “””

(ghilimele), “.” (punctul), “,” (virgula). Aceste caractere nu își păstrează formatarea îndeosebi în interiorul fragmentelor care sunt scrise cu bold sau cu italic, apărând în format normal.

Pe lângă aceste probleme ivite în procesul de construcție a vectorului de formatare, s-a constatat o altă problemă la parsarea vectorului. La sfârșitul fiecărei intrări există o listă în care sunt scrise diferite informații referitoare la cuvântul explicat (plural, pronunție, etimologie etc.). Aceste informații au formatare diferite, care, la parsarea vectorului construit, ar putea duce la apariția unor erori.

2.3. Soluții în vederea evitării erorilor de parsare

Pentru a evita toate aceste probleme ce pot genera erori la parsare, s-a impus o prelucrare a vectorului înainte ca acesta să fie parsat și aducerea acestuia într-o formă mai restrânsă (contopirea în unul singur a elementelor din vector, determinarea formatareii corecte a caracterelor ce nu-și păstrau formatarea corectă și aducerea la aceeași formatare a informațiilor din listele ce încheie o intrare).

Odată adus vectorul în forma dorită, următorul pas este parsarea acestuia și construirea efectivă a fișierului în format XML. Parsarea vectorului are la bază succesiunea stilurilor fragmentelor, prezentată în schema de mai sus, la care se adaugă și tratarea cazurilor particulare ce au fost constatate pe parcursul testării aplicației pe un eșantion cât mai larg de pagini din DLR.

Datorită faptului că volumele DLR-ului au fost scrise de către autori diferiți, precum și a faptului că unele volume sunt mai vechi iar altele au fost scrise mai recent, aplicația ar putea fi mereu îmbunătățită pentru ca parsarea să aibă un procentaj cât mai mare de reușită, fără a fi nevoie de prea multe intervenții din partea factorului uman, care să trateze manual eventualele cazuri particulare.

2.4. Aplicația realizată și funcționalitatea acesteia

Instrumentul creat în cadrul proiectului nostru de cercetare a fost numit **DLR_{ex}** și reprezintă un instrument de achiziționare, prelucrare și consultare a Dicționarului Limbii Române în format electronic. De asemenea, trebuie să precizăm de la început că funcționalitatea acestui instrument nou creat pentru a dezvolta ulterior descrierea caracteristicilor principale:

- permite trecerea textului DLR din format RTF (Word) în format XML;
- permite vizualizarea și corectarea fișierelor XML;
- construiește DLR-ul în format electronic;
- permite actualizarea și unificarea DLR;
- funcționează ca interfață de consultare și realizează interogarea DLR în format electronic.

Funcționalitatea de bază a aplicației este aceea de transpunere a DLR-ului în format electronic (XML). În prima fază, aplicația are un fișier XML gol. Pentru crearea DLR-ului electronic sau pentru adăugarea de noi pagini la cele existente deja, se încarcă în program fișierele RTF, după cum este ilustrat în Figura 2.

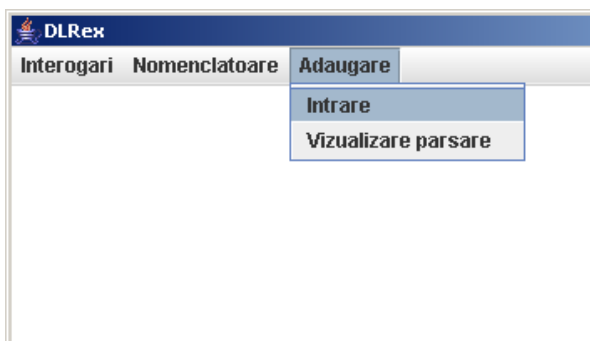


Figura 2: Captură de ecran: încărcarea în DLRex a fișierelor

La intrare se pot selecta mai multe fișiere, în cazul în care scanarea și corectarea paginilor din dicționar s-a făcut pe fiecare pagină în parte și nu s-au pus ulterior în același fișier toate paginile lucrate. Singura condiție, pentru cazul în care se încarcă în aplicație mai multe fișiere, este ca ele să fie selectate în ordinea numărului paginilor din dicționar, ca în Figura 3, pentru ca, în cazul în care o intrare în dicționar începe la sfârșitul unei pagini și continuă pe următoarea, aplicația să identifice pagina următoare și să facă saltul necesar pentru realizarea corectă a parsării. În urma parsării fișierelor, aplicația construiește un fișier XML temporar, al cărui nume este de forma "DLR_" + data și ora la care a fost scris + ".xml". Acest lucru este conceput pentru a evita introducerea datelor nedorite în fișierul XML în care este păstrat DLR-ul creat până în momentul respectiv. Fișierul XML rezultat este salvat în subdirectorul "data" al aplicației.

Pentru adăugarea la fișierul XML ce conține DLR, trebuie să se verifice mai întâi dacă parsarea s-a realizat cu succes. Pentru aceasta se deschide fișierul XML rezultat în urma parsării.

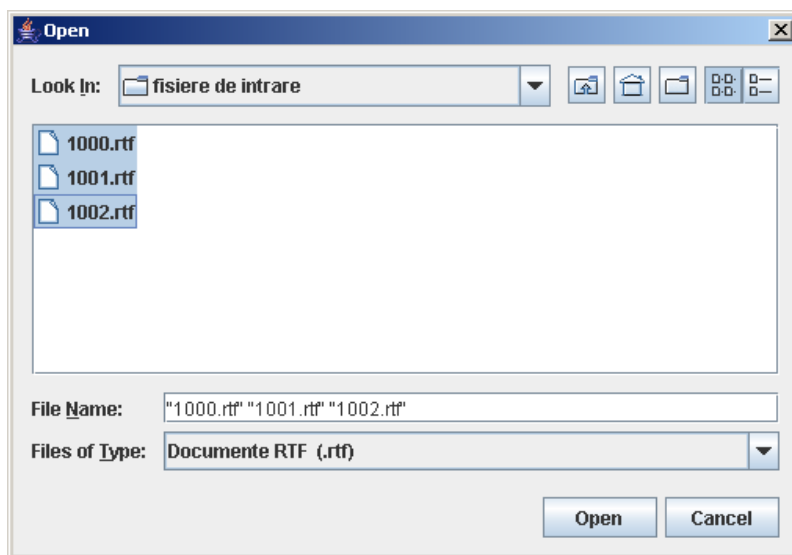


Figura 3: Captură de ecran: ordinea de încărcare a fișierelor în DLRex

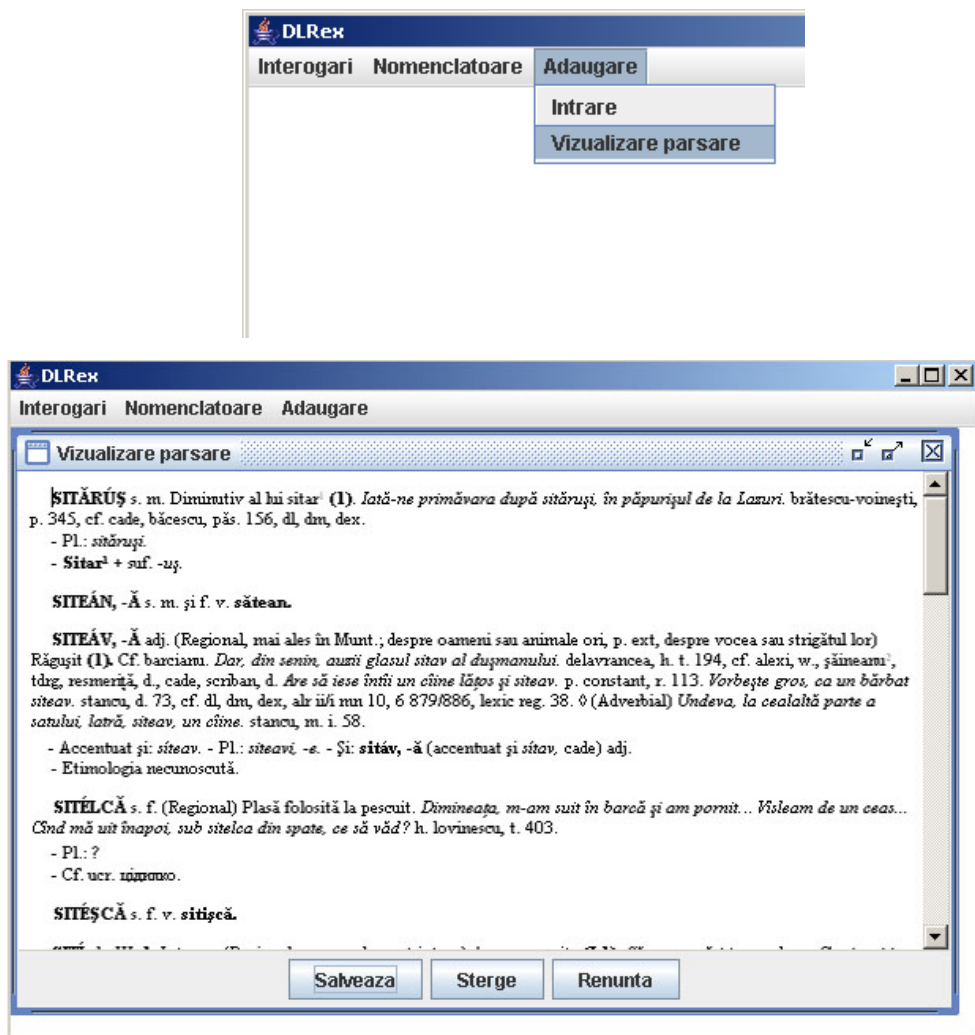


Figura 4: Capturi de ecran: deschiderea și vizualizarea unei parsări

În cazul în care rezultatul în urma parsării este corect, trebuie apăsat butonul “Salvează” pentru a adăuga noile pagini parsate la DLR construit până în acel moment. Dacă parsarea nu este corectă, trebuie apăsat butonul “Șterge”, pentru a muta fișierul temporar construit în urma parsării în directorul de backup, urmând ca ulterior să corecteze în fișierul RTF greșelile de formatare, apoi să se încerce o nouă parsare. În situația în care utilizatorul nu este sigur de acțiunea pe care dorește să o realizeze, sau vrea să amâne decizia, poate să apese butonul “Renunță”, ceea ce permite ca operațiunea să se poată relua în orice moment.

Pentru adăugarea de noi părți de vorbire existente în DLR sau pentru editarea cotelor ori a cronologiilor, se folosește submeniul din mijloc al aplicației.

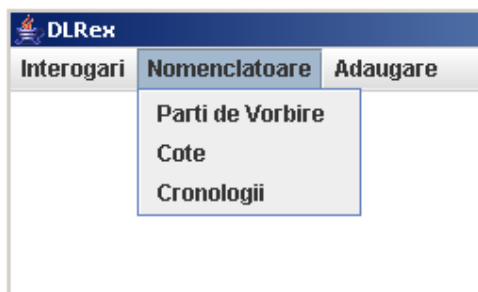


Figura 5: Captură de ecran: Meniul *Nomenclatoare* din DLReX

Aplicația are de asemenea și un modul de interogări (căutare a cuvintelor în DLR), ce poate fi accesat din primul submeniu al aplicației.

Căutarea unui cuvânt în DLR are și un modul mai avansat în care putem filtra modul după care să se realizeze căutarea (Figura 7).

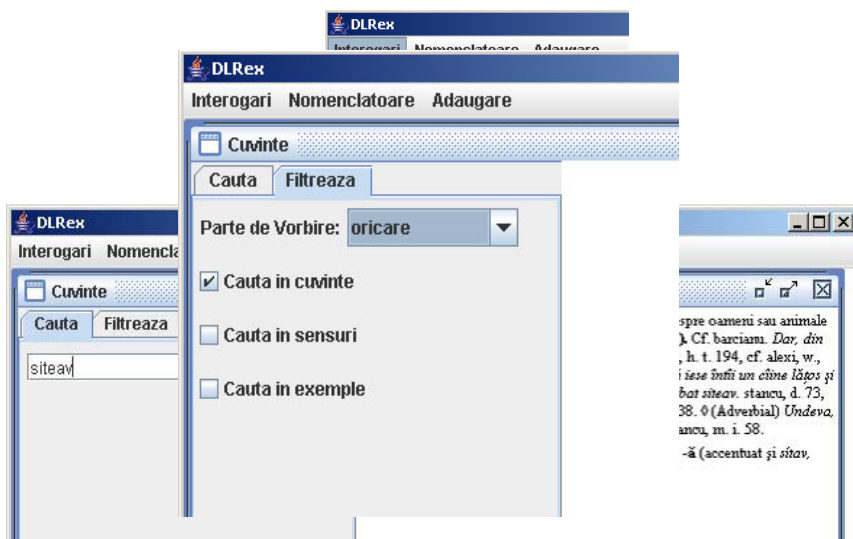


Figura 7: Captură de ecran: Căutarea cuvintelor din DLR folosind filtrări

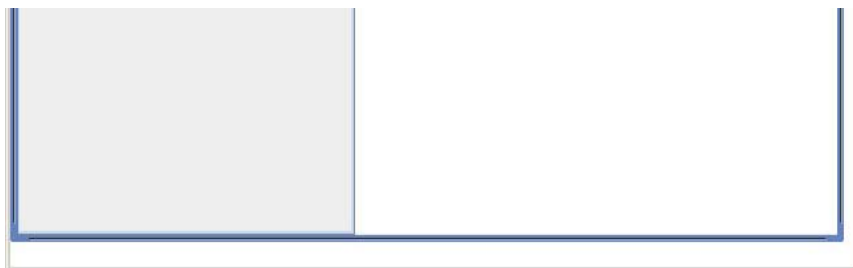


Figura 6: Capturi de ecran: Căutarea cuvintelor din DLR folosind DLReX

Timpul mediu de adăugare (parsarea împreună cu verificarea corectitudinii) a unei pagini din DLR în format RTF variază între 30 și 60 de secunde. Acest timp poate crește în situația apariției unor cazuri particulare de formatare, care nu au fost întâlnite în paginile de test ale aplicației.

3. Concluzii

Prin realizarea acestui proiect s-a făcut un pas important în favoarea eficientizării procesului de realizare a *Dicționarului Limbii Române (DLR)* în format electronic, instrument și bază de date fundamentale pentru informatizarea cercetării lingvistice românești, a celei lexicografice în special, dar și pentru crearea resurselor și instrumentelor necesare integrării limbii române în rândul limbilor de circulație internațională prin intermediul mijloacelor electronice actuale.

Poate părea ambițioasă această perspectivă, dar eforturile realizate până în prezent de membrii Consorțiului de informatizare pentru limba română (<http://consilr.info.uaic.ro/>), precum și cadrul științific construit de acesta ne îndreptătesc să sperăm că, prin cooperare interinstituțională și transdisciplinarite, prin conjugarea rezultatelor cercetărilor desfășurate în mai multe centre și instituții, perspectiva în care ne situăm va deveni realitate.

DLR^{ex} deschide posibilități încurajatoare. Cu ajutorul acestui instrument este realizabilă achiziționarea integrală a *Dicționarului Limbii Române (DLR)* și, ulterior, achiziționarea *Dicționarului Academiei (DA)*, editat sub coordonarea lui Sextil Pușcariu în prima jumătate a secolului al XX-lea, potrivit unei euristici analoage celei care a permis transpunerea DLR în format XML; în fine, va fi posibilă actualizarea DLR integral, de la A la Z, și realizarea unor ediții viitoare, cu ritmicitatea și eficiența pe care mijloacele actuale le permit, în concordanță cu nivelul cercetării avansate în domeniu. Acest deziderat nu poate fi atins fără a se crea, paralel, un corpus de texte românești suficient de mare și de performant pentru respectarea standardelor științifice ale acestei lucrări fundamentale a culturii române.

Proiectul nostru propune căi de realizare a unui obiectiv ce nu poate fi atins fără suport instituțional – ne referim în primul rând la Academia Română și la Ministerul Educației și Cercetării –, care să promoveze și să susțină un atare demers, ce presupune implicarea cercetării interdisciplinare, formarea de cercetători cu specializare multiplă, crearea unui cadru juridic care să permită accesul liber al cercetătorului la resursele lingvistice (publicații, instrumente etc.) realizate sub egida Academiei, precum și a celor realizate în cadrul cercetării din sistemul de învățământ universitar. Fără acest sprijin, entuziasmul care a dus la realizările de până în prezent își pierde forța sau rămâne, cel mult, o manifestare epigonică a mitului sisific.

The Dictionary of the Romanian Language (DLR) in Electronic Format. Studies regarding its acquisition

The paper presents the results of our research pursued during a project financed by the Romanian Ministry of Education and Research, through CNCSIS - the National University Research Council. The main objective of the project was to study and optimise the possibilities to obtain through computer-aided acquisition the electronic format of the thesaurus dictionary of Romanian (DLR). Due to the financial and technical conditions offered by the project, our studies show that an acceptable solution is to define a parsing strategy, specific to DLR. An

output of the project is a Java-implemented tool, DLReX, an instrument for acquisition, processing and browsing the DLR. The computer-aided acquisition of DLR implies scanning, converting through an OCR, and manually correcting each printed file of DLR; the files obtained this way are further processed by DLReX in order to finally have them in an XML format. The results of the project show that the integral acquisition of DLR in an electronic format is feasible.