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 şi culturale în Comunitatea Europeană. Păstrarea identității limbilor şi a culturilor naționale în cadrul societății informaționale globale a fost încă din anii '90 o realitate, aşa 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 şi, în mod special, posibilitățile de traducere automată asigură compatibilizarea unei limbi utilizate de o comunitate națională, medie ca dimensiune, cu limbile de circulație internațională şi menținerea identității lingvistice, în condițiile apropierii de o societate fără granițe. Punerea la dispoziție în limba nativă a informației din mediul internațional, precum şi transmiterea într-o limbă de circulație mondială a informațiilor naționale au devenit demult o responsabilitate națională[2]. Acest lucru contribuie la înlăturarea barierelor naturale, naționale, culturale şi lingvistice, furnizând acces direct la informație în limba maternă a utilizatorilor. Implicit, specificul românesc poate deveni astfel accesibil oricăruia dintre locuitorii planetei care are acces la Internet.

În contextul de „unitate prin diversitate" păstrarea limbilor şi culturilor actuale se bucură de un interes deosebit, fiind stimulată puternic de organisme internaționale, precum Comisia Europeană[3]. Mai mult, ținând seama de situația particulară a limbii române, este evident că orice efort în această direcție contribuie, în final, şi 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 şi ale continentului american), aşa cum se doreşte şi prin actualul program guvernamental: „Guvernul României are datoria de a acționa pentru sprijinirea dezvoltării şi exprimarea identității culturale a românilor din afara granițelor României"[4].

---

* Institutul de Filologie Română „A. Philippide", Iaşi
** Facultatea de Informatică, Iaşi

[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 Tufiş and Poul Anderson (eds.), *Language and Technology*, Editura Academiei, Bucureşti, 1997.

[3] D. Tufiş, D. Cristea *RO-BALKANET – ontologie lexicalizată, în context multilingv, pentru limba română*, în Dan Tufiş şi Fl. Gh. Filip (eds.), *Limba Română în Societatea Informațională - Societatea Cunoaşterii*, Editura Expert, Bucureşti, 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 societăţii informaţionale actuale, se realizează numai printr-o strânsă colaborare între lingvişti şi informaticieni. Versiunea electronică a DLR-ului *integral* – obiectiv de viitor, pentru care proiectul de faţă înseamnă punctul de plecare şi baza de la care se va porni – va oferi lumii, şi mai ales comunităţii românilor de pretutindeni, tezaurul limbii române, precum şi mai multe posibilităţi pentru aplicaţiile de procesare a limbajului uman pentru limba română, incluzând traducerea automată şi cea asistată de calculator, acces multimedia inteligent la Internet, regăsirea informaţiei (prin căutarea unor cuvinte-cheie sau concepte), extragerea informaţiei (legate de un anumit subiect, deci de un anumit cuvânt-cheie), sumarizare automată, dezambiguizare semantică, achiziţia de cunoştinţe etc.

Ideea proiectului *Dicţionarul limbii române (DLR) în format electronic* se integrează într-un program amplu de informatizare a cercetării lingvistice româneşti, coordonat în prezent de Comisia de Informatizare pentru Limba Română din cadrul Academiei Române, Secţia de Ştiinţa şi Tehnologia Informaţiei, subsumându-se totodată obiectivelor prioritare privind informatizarea ale Ministerului Educaţiei şi Cercetării. Scopul Comisiei de Informatizare pentru Limba Română îl constituie apărarea identităţii limbii române prin promovarea studiilor dedicate ei dintr-o perspectivă informaţională[5].

Prin sprijinul financiar obţinut de la MEC prin intermediul CNCSIS (Consiliul Naţional al Cercetării Ştiinţifice din Învăţământul Superior), s-au studiat modalităţile de achiziţionare în format electronic a *Dicţionarului limbii române*, în cadrul unui proiect desfăşurat în perioada 2003 – 2005. Pe lângă atingerea obiectivului central, s-au putut crea şi o serie de aplicaţii ce facilitează utilizarea informatizată a materialului existent în DLR.

În proiect au fost implicaţi tineri cercetători din domeniul lexicologiei-lexicografiei şi din domeniul lingvisticii computaţionale, care au format o echipă omogenă, datorită faptului că dificultăţile de comunicare au fost depăşite prin specializarea interdisciplinară.

Ca formulă de citire şi convertire a DLR în format XML s-a optat pentru utilizarea unei euristici specifice DLR, care permite recunoaşterea, pe baza caracteristicilor formale ale textului, a unor câmpuri de text, fiecare cu semnificaţie bine determinată. Realizarea obiectivului iniţial, care se referea la definirea unei gramatici lexicografice proprii DLR, a fost amânată; atingerea acestui obiectiv s-a dovedit imposibil de realizat în condiţiile date (restricţiile de timp, mijloacele tehnice avute la dispoziţie), rămânând ca, după ce vom finaliza achiziţionarea DLR cu ajutorul DLRex – instrumentul de achiziţionare, prelucrare şi consultare a DLR, creat în cadrul acestui proiect –, să fie semnalate toate problemele privitoare la normele DLR în vederea rediscutării şi adaptării acestora condiţiilor de formalizare optimă. Ulterior, va fi posibilă definirea unei gramatici cu un număr rezonabil de reguli, astfel încât funcţionalitatea şi randamentul să fie îmbunătăţite.

1.2. Etapele de lucru

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

- scanarea unui eşantion de pagini tipărite din fiecare fasciculă a ediţiei 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 şi a convertirii;
- procesarea textului cu ajutorul DLRex şi transpunerea lui în format XML.

[5] D. Cristea, D. Tufiş, *Resurse lingvistice româneşti şi tehnologii informatice aplicate limbii române*, în Ofelia Ichim şi Florin-Teodor Olariu (eds.), *Identitatea limbii şi literaturii române în perspectiva globalizării*, Academia Română, Institutul de Filologie Română „A. Philippide", Editura Trinitas, Iaşi, (2002).

Odată parcurse aceste etape, am obținut fișier XML[6] 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 şi a tiparului din volumele editate înainte de 1990, de performanțele tehnice ale scannerului şi ale programului de recunoaştere / convertire a imaginii digitale în text. De aceea, pentru ameliorarea vitezei de lucru, este necesară îmbunătățirea instrumentelor tehnice, utilizarea unor programe de convertire adaptate recunoaşterii simultane a mai multor limbi, inclusiv a celor ce nu utilizează litere latine, precum şi accesul la textele tehnoredactate computațional după 1990, pentru a se elimina etapele scanării, convertirii şi, 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 şi formatul fişierelor de intrare

Ideea de bază a aplicației constă în trecerea DLR-ului în format electronic, şi anume în format XML. În urma scanării filelor de dicționar, a convertirii acestora de către OCR[7] şi a corectării de către lingvişti lexicografi (care sunt familiarizați cu formatul standard al DLR-ului) au rezultat fişiere 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 eventualilor paşi intermediari în pregătirea fişierelor 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 fişiere. În prima fază s-a încercat scrierea aplicației în Delphi 7[8], pentru că acesta are integrate componente care lucrează (realizează citirea şi scrierea) direct din fişierul .doc. S-a constatat că la citirea dintr-un fişier .doc, programul Delphi 7 ține cont şi 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 şi 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 fişier, interpretând astfel siglele ca şi cum ele ar avea stilul fontului normal.

Deoarece sigla este foarte importantă, iar recunoaşterea ei în dicționar ar ajuta la parsarea[9] 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 şir 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 – şi construirea arborelui de parsare care captează ierarhia implicită a datelor de intrare.

lucrează mai uşor cu *String*-urile (şiruri de caractere) şi prelucrarea acestora este mai facilă în acest mediu de programare. Pentru că JAVA citeşte foarte greu din binar (modul în care sunt salvate în fişiere documentele Word), s-a optat pentru ca fişierele de intrare să fie fişiere Word, dar de acestă dată să fie salvate în format RTF (Rich Text Format).

2.2. Modelul

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

cuvânt **parte de vorbire, sensul cuvântului [** ***exemplu care ilustrează sensul menționat*** **SIGLA]**

A **sensul cuvântului [** ***exemplu care ilustrează sensul menționat*** **SIGLA]**

I **sensul cuvântului [** ***exemplu care ilustrează sensul menționat*** **SIGLA]**

1 **sensul cuvântului [** ***exemplu care ilustrează sensul menționat*** **SIGLA]**

a) **sensul cuvântului [** ***exemplu care ilustrează sensul menționat*** **SIGLA]**

b) **sensul cuvântului [** ***exemplu care ilustrează sensul menționat*** **SIGLA]**

.....

2) **sensul cuvântului [** ***exemplu care ilustrează sensul menționat*** **SIGLA]**

.....

II **sensul cuvântului [** ***exemplu care ilustrează sensul menționat*** **SIGLA]**

.....

III **sensul cuvântului [** ***exemplu care ilustrează sensul menționat*** **SIGLA]**

.....

B **sensul cuvântului [** ***exemplu care ilustrează sensul menționat*** **SIGLA]**

.....

C **sensul cuvântului [** ***exemplu care ilustrează sensul menționat*** **SIGLA]**

.....

- **informații ortoepice, gramaticale, de circulație …**
- **informații etimologice**

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 rombul 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 fişier) fiecare fragment ce are o formatare diferită față de fragmentul ce îl precedă şi, 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 fişierului XML dorit. După acest pas s-a observat că, deşi unele fragmente erau în text unul lângă celălalt şi aveau şi aceeaşi formatare, ele erau salvate în interiorul fişierului RTF ca având formatări diferite. Altă anomalie observată o constituie unele caractere care nu îşi păstrează formatarea. Acestea au apărut în urma scanării şi nu au putut fi sesizate cu ochiul liber de către persoanele care au făcut corecția manuală a fişierelor de intrare. Astfel de caractere sunt “ ” (spațiul), “”””

(ghilimele), "." (punctul), "," (virgula). Aceste caractere nu îşi 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 formatări, s-a constatat o altă problemă la parsarea vectorului. La sfârşitul fiecărei intrări există o listă în care sunt scrise diferite informații referitoare la cuvântul explicat (plural, pronunție, etimologie etc.). Aceste informații au formatări 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 şi aducerea acestuia într-o formă mai restrânsă (contopirea în unul singur a elementelor din vector, determinarea formatării corecte a caracterelor ce nu-şi păstrau formatarea corectă şi aducerea la aceeaşi fomatare a informațiilor din listele ce încheie o intrare).

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

Datorită faptului că volumele DLR-ului au fost scrise de către autori diferiți, precum şi 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 reuşită, fără a fi nevoie de prea multe intervenții din partea factorului uman, care să trateze manual eventualele cazuri particulare.

2.4. Aplicația realizată şi funcționalitatea acesteia

Instrumentul creat în cadrul proiectului nostru de cercetare a fost numit DLRex şi reprezintă un instrument de achiziționare, prelucrare şi 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 şi corectarea fişierelor XML;
- construieşte DLR-ul în format electronic;
- permite actualizarea şi unificarea DLR;
- funcționează ca interfață de consultare şi 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 fişier XML gol. Pentru crearea DLR-ului electronic sau pentru adăugarea de noi pagini la cele existente deja, se încarcă în program fişierele RTF, după cum este ilustrat în Figura 2.

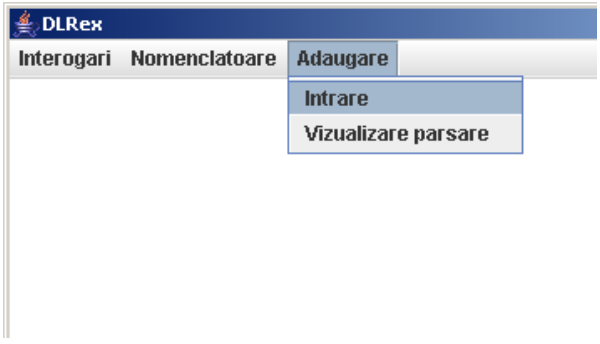

Figura 2: Captură de ecran: încărcarea în DLRex a fişierelor

La intrare se pot selecta mai multe fişiere, în cazul în care scanarea şi corectarea paginilor din dicţionar s-a făcut pe fiecare pagină în parte şi nu s-au pus ulterior în acelaşi fişier toate paginile lucrate. Singura condiţie, pentru cazul în care se încarcă în aplicaţie mai multe fişiere, este ca ele să fie selectate în ordinea numărului paginilor din dicţionar, ca în Figura 3, pentru ca, în cazul în care o intrare în dicţionar începe la sfârşitul unei pagini şi continuă pe următoarea, aplicaţia să identifice pagina următoare şi să facă saltul necesar pentru realizarea corectă a parsării. În urma parsării fişierelor, aplicaţia construieşte un fişier XML temporar, al cărui nume este de forma "DLR_" + data şi ora la care a fost scris + ".xml". Acest lucru este conceput pentru a evita introducerea datelor nedorite în fişierul XML în care este păstrat DLR-ul creat până în momentul respectiv. Fişierul XML rezultat este salvat în subdirectorul "data" al aplicaţiei.

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

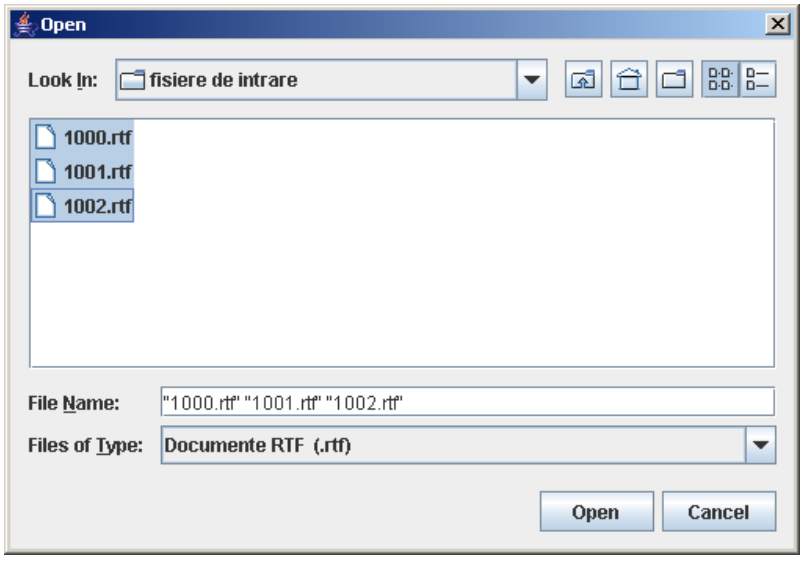

Figura 3: Captură de ecran: ordinea de încărcare a fişierelor în DLRex

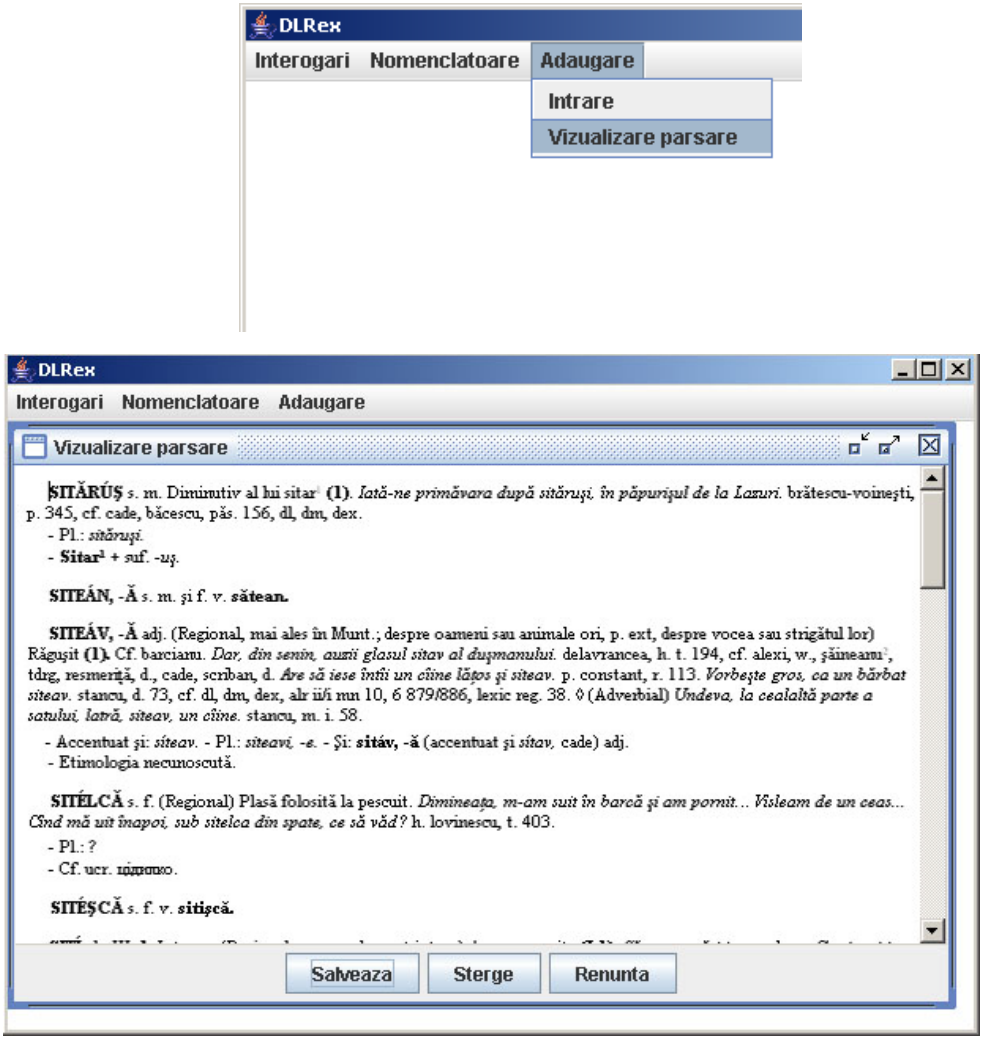

Figura 4: Capturi de ecran: deschiderea şi vizualizarea unei parsări

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 bakup, 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 sa apese butonul "Renunță", ceea ce permite ca operațiunea să se poată relua în orice moment.

Pentru adăugarea de noi părţi de vorbire existente în DLR sau pentru editarea cotelor ori a cronologiilor, se foloseşte submeniul din mijloc al aplicaţiei.

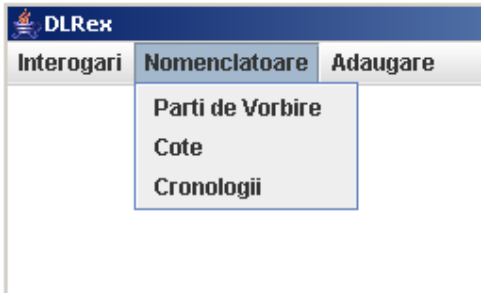

Figura 5: Captură de ecran: Meniul ***Nomenclatoare*** din 𝔻𝕃ℝ𝕖𝕩

Aplicaţia are de asemenea şi un modul de interogări (căutare a cuvintelor în DLR), ce poate fi accesat din primul submeniu al aplicaţiei.

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

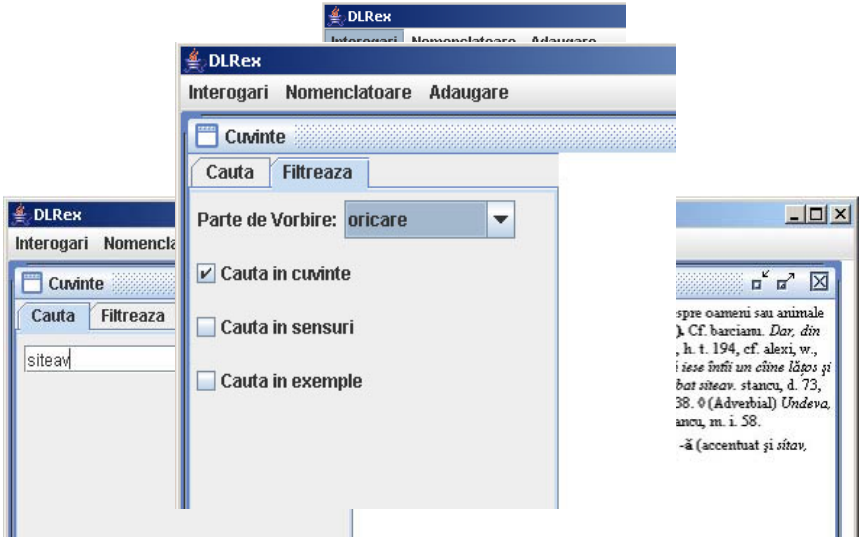

Figura 7: Captură de ecran: Căutarea cuvintelor din DLR folosind filtrări

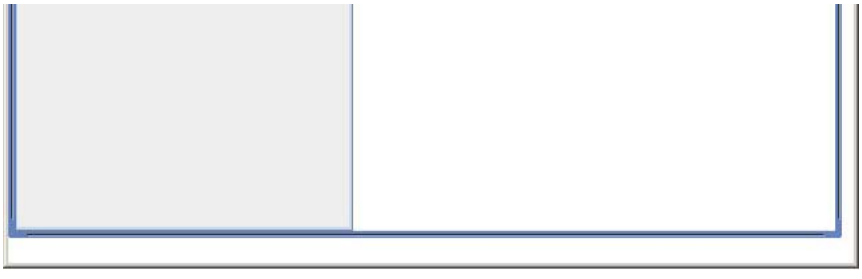

Figura 6: Capturi de ecran: Căutarea cuvintelor din DLR folosind 𝔻𝕃ℝ𝕖𝕩

Timpul mediu de adăugare (parsarea împreună cu verificarea corectitudinii) a unei pagini din DLR în format RTF variază între 30 si 60 de secunde. Acest timp poate creşte 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 şi bază de date fundamentale pentru informatizarea cercetării lingvistice româneşti, a celei lexicografice în special, dar şi pentru crearea resurselor şi 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 şi cadrul ştiințific construit de acesta ne îndreptățesc să sperăm că, prin cooperare interinstituțională şi transdisciplinarite, prin conjugarea rezultatelor cercetărilor desfăşurate în mai multe centre şi instituții, perspectiva în care ne situăm va deveni realitate.

DLRex deschide posibilități încurajatoare. Cu ajutorul acestui instrument este realizabilă achiziționarea integrală a *Dicționarului Limbii Române (DLR)* şi, ulterior, achiziționarea *Dicționarului Academiei (DA)*, editat sub coordonarea lui Sextil Puşcariu 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, şi realizarea unor ediții viitoare, cu ritmicitatea şi 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 româneşti suficient de mare şi de performant pentru respectarea standardelor ştiințifice 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 şi la Ministerul Educației şi Cercetării –, care să promoveze şi 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 şi a celor realizate în cadrul cercetării din sistemul de învățământ universitar. Fără acest sprijin, entuziasmul care a dus la realizările de până în prezent îşi 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.