Despre necesitatea realizării unui corpus lexicografic românesc esențial[1]

Elena DĂNILĂ

Key-words: *lexicography, computerized lexicography, linguistic resources, computerized lexicographic instruments*

Trăim într-o epocă în care comunicarea se face cu o viteză fantastică. Schimbul de informații, de orice fel, este aproape instantaneu. Apariția internetului în anii '60 a schimbat radical şi perspectiva din care este privit atât accesul la informații, cât şi schimbul de informații. Şi acest lucru este cu atât mai evident dacă ne gândim la cum circula informația în Antichitate sau în Evul Mediu – de la tradiția orală şi papirusuri, de la bibliotecile care erau accesibile doar călugărilor învățați ai momentului (vezi Umberto Eco, *Numele trandafirului*) la bibliotecile publice, deschise aproape oricărei categorii de cititor etc.; de la mesajele transmise cu ajutorul fumului sau al porumbeilor la mesajele instantanee trimise prin fibră optică sau wireless, prin email sau mass/ mess etc.

Astfel, astăzi oricine poate afla foarte multe lucruri la o simplă căutare pe internet! Deci informația este cu atât mai mult un bun public. Bibliotecile virtuale dau posibilitatea mult mai multor cititori de a ajunge la cărți pe care, înainte vreme, le găseau foarte greu sau, poate, niciodată! Cu atât mai important este accesul la informația din domeniu din toată lumea şi din toate timpurile, aproape pentru cei care studiază/ cercetează!

Viteza fantastică de care vorbeam mai sus a început să se manifeste şi în domeniul al studiului, din diferite perspective, al limbii române, cu toate că începutul a fost mai „timid"! Deşi conectarea României la internet s-a făcut în 1993, primele proiecte de informatizare a cercetării filologice româneşti au început mai târziu. Astfel, pentru limba română, nu cu mulți ani în urmă, nu exista aproape nici un corpus de limbă română (cu acces liber) pe internet!

Între timp s-au produs diferite „mişcări", inițiativa datorându-se, în primul rând, informaticienilor cu preocupări în domeniul prelucrării limbajului natural. Însă colaborarea între informaticieni şi filologi a dus, şi în România, la rezultate spectaculoase.

Pentru realizarea unor corpusuri de limba română, în ultimii anii s-au înregistrat:

[1] Acest articol a fost redactat în cadrul proiectului cu numărul PN II-RU 246/2010, finanțat de CNCSIS–UEFISCSU.

1) inițiative instituționale – cu *acces restricționat/ limitat*, doar în scop de cercetare (în funcție de legislația în vigoare privitoare la drepturile de autor), ceea ce demonstrează necesitatea unor *politici lingvistice* coerente pentru reglementarea acestei situații. Astfel, inițiativele instituționale s-au concretizat în:

a) granturi/ proiecte cu finanțare din diverse surse (de exemplu, după desfăşurarea a 4 granturi[2], cu finanțare CNCSIS, a fost inițiat şi definitivat proiectul complex eDTLR[3], cu finanțare CNMP, cu participare națională, de la care s-a plecat pentru realizarea proiectului CLRE – cu finanțare CNCSIS şi a unor eventuale alte proiecte);

b) inițiative „locale", ale unor anumite instituții (Institute, Centre de cercetare, Universități), realizate în planul de lucru al acestora, s-au transformat ulterior în granturi cu finanțare națională (de exemplu, proiectul CNR – Bucureşti)

2) inițiative particulare – ale căror rezultate sunt oferite, de obicei, cu acces liber la informație (v. http://dexonline.ro/ sau activitatea unor edituri care pun la dispoziția publicului, pe site-urile proprii, accesul la dicționarele sau la formatul electronic al cărților pe care le-au editat – de exemplu, Editura Litera[4]).

Corpusurile de limbă pot fi clasificate, în funcție de diferite criterii, în:

– corpusuri generale de limbă (scrisă, vorbită etc.);
– corpusuri pe anumite teme;
– corpusuri lexicografice;
– corpusuri de limbă scrisă;
– corpusuri video/ audio/ multimedia.

Pe Internet sunt accesibile, cu acces liber sau parțial liber, variate corpusuri de texte, pentru diferite limbi.

Pentru limba *franceză*, cele mai importantă resurse de acest tip sunt cele oferite de laboratorul ATILF – http://www.atilf.fr/, care prezintă *Les ressources linguistiques informatisés pour l'étude du français* – corpus ce include diverse resurse lingvistice pentru studiul francezei:

– *Trésor de la Langue Française informatisé (TLFi)* care reprezintă varianta digitizată a TLF (*Trésor de la Langue Française*), sec. XIX-XX, în 16 volume, 100.000 intrări, 270.000 definiții, 430.000 exemple, şi la care accesul este liber, permițându-se trei niveluri de consultare (vizualizare simplă a unui articol, consultare transversală, interogări complexe);

– *Les Dictionnaires de l'Académie française informatisés* – colecție de dicționare cu acces liber la TLFi, incluzând patru dicționare vechi şi şase dicționare ale Academiei Franceze;

[2] *Dicționarul limbii române (DLR) în format electronic. Studii privind achiziționarea*, proiect finanțat de CNCSIS, 2003–2005, Iaşi; *Resurse lingvistice în format electronic: Monumenta linguae Dacoromanorum. Biblia 1688. Regum I, Regum II – Ediție critică şi corpus adnotat*, proiect finanțat de CNCSIS, 2006–2007, Iaşi; *DLRI. Bază lexicală informatizată. Derivate*, proiect finanțat de CNCSIS, 2007–2008, Iaşi; *CNR – Corpus de referință al limbii române pentru constituirea de dicționare academice*, proiect finanțat de CNCSIS, 2007–2008, Bucureşti.

[3] *eDTLR – Dicționarul tezaur al limbii române în format electronic*, proiect finanțat de Consiliul Național al Managementului Proiectelor (CNMP), 2007–2010, cu participarea a şapte instituții de cercetare din țară.

[4] http://nodex.litera.ro/cautare_dex/dictionar.

– *Les bases de données textuelles Frantext* – bază de date textuale, care cuprinde peste 3650 de texte din sec. al XVI-lea – al XX-lea, 218 texte din secolele al al XIV-lea şi al XV-lea, 22 texte medievale; acces se face pe bază de abonament;

– *Encyclopédie Diderot et d'Alembert* – 72.000 articole;

– *Matériaux pour le dictionnaire du Moyen Français* – 69.089 articole corespunzătoare unui nomenclator de 26 354 unităţi, completate de un lematizator on-line;

Condiţiile de acces pe care Laboratorul ATILF le propune celor interesaţi sunt:

– cu acces liber (pentru *TLFi, rayon des dictionnaires* şi pentru *base des lexiques du Moyen Francais*);

– cu abonament (pentru *FRANTEXT* şi pentru *Encyclopédie Diderot et d'Alembert*);

– prin convenţii specifice (pentru accesul la textele integrale sau la alte resurse).

Pentru limba *engleză* sunt disponibile pe internet mai multe resurse on-line:

– *Oxford English Corpus* – corpus pentru engleza britanică, care include două bilioane cuvinte; accesul este liber

(http://www.oxforddictionaries.com/page/oec?view=uk);

– *British National Corpus* (BNC) – corpus pentru engleza britanică, care include 22.000.000 cuvinte, în varianta scrisă sau orală, din secolul al XX-lea, cea mai recentă ediţie fiind din martie 2007; acces se face prin licenţă, în anumite condiţii (http://www.natcorp.ox.ac.uk/);

– *American National Corpus* (ANC) – corpus pentru engleza americană, care include 22.000.000 cuvinte, în varianta scrisă sau orală, începând cu anul 1990; acces la cca 10.000.000 cuvinte este permis prin licenţă nominală, în scop de cercetare sau în scop educaţional şi contra cost la toată informaţia din corpus (http://www.anc.org/).

Şi pentru unele limbi mai apropiate de română, geografic vorbind, s-au realizat corpusuri de limbă. Astfel, de exemplu, pentru limba *croată* există *Croatian National Corpus* (http://hnk.ffzg.hr/default_en.htm), pentru limba *rusă* – *Russian National Corpus* – corpus care include peste 150.000.000 de forme (http://ruscorpora.ru/en/index.html) ş.a.

Pentru limba română există câteva corpusuri care pot fi subcategorizate în două tipuri.

1) *Corpusuri de texte* în limba română, unele datorate iniţiativei unor iubitori de limba română „individuali" (vezi exemplul Radei Mihalcea), altele – proiectelor unor instituţii:

– *DACOROMANICA* (http://www.dacoromanica.ro/) – proiect realizat de *Biblioteca Metropolitană Bucureşti* şi *Biblioteca Academiei Române* – cea mai importantă bibliotecă digitală românească[5], la ora actuală, accesibilă gratuit în internet; care oferă un fond de 2000 de volume ce însumează 1.000.000 pagini, 400 de imagini şi câteva resurse sonore digitizate[6] şi care are ca obiectiv pentru anul

[5] *Dacoromanica* şi-a propus pentru anul 2010 adăugarea a cca 1.500.000 de pagini, 1.500 imagini şi resurse sonore.

[6] *Dacoromanica* este partener oficial al celei mai importante biblioteci digitale la nivel european – Biblioteca Virtuală Europeană EUROPEANA http://www.europeana.eu/portal/.

2010 Documentele aparţin perioadei cuprinse între Evul Mediu şi începutul sec. XX. Obiectivul este de a constitui o bibliotecă patrimonială şi enciclopedică. Politica documentară se bazează pe un corpus de autori, pe colecţii de periodice şi de serii editoriale, pe instrumente bibliografice, enciclopedice şi lingvistice, pe corpus-uri realizate în baza dosarelor tematice pe cât posibil în format multimedia. Accesul la informaţiile de pe platforma DACOROMANICA este liber la cea mai mare parte a resurselor;

– *CONSILR* – Consorţiul pentru Informatizarea Limbii Române (http://consilr.info.uaic.ro/ro/) – portal care permite şi accesul la unele resurse în limba română, accesul făcându-se în condiţii care sunt specificate de la caz la caz;

– *eDTLR*[7] – corpus care include, pe de o parte, formatul electronic al *Dicţionarului tezaur al limbii române*, şi, pe de altă parte, resursele din Bibliografia *Dicţionarului*, precum şi programele de interogare/ consultare;

– proiectul *CNR – Corpus de referinţă al limbii române pentru constituirea de dicţionare academice*, realizat în perioada 2007–2008 – Institutul de Lingvistică „Iorgu Iordan – Al. Rosetti", Bucureşti; accesul la resursele create prin acest proiect este strict, acestea fiind destinate deocamdată exclusiv cercetătorilor de la institutul menţionat;

2) *Corpusuri lexicografice* pentru limba română, care pot fi subcategorizate, la rândul lor, în:

– corpusuri monoligve: *DEXONLINE* – 32 de dicţionare transpuse textual pe internet; proiect datorat unei iniţiative personale (Cătălin Frâncu, Bucureşti – 2001);

– bi- sau plurilingve – *RoWordNet* (reţea semantică lexicală a limbii române; aliniată la nivel de concept cu cea a limbii engleze); *EUROVOC* (tezaur multilingv – numeroase dicţionare bilingve extrase automat din corpusuri paralele).

În privinţa corpusurilor de texte pentru limba română (şi nu numai), principalele probleme care apar sunt legate de:

– finanţare şi iniţiativă, ceea ce presupune, din nou, necesitatea rafinării politicii lingvistice a României;

– respectarea legislaţiei în vigoare în ceea ce priveşte proprietatea intelectuală;

– securizarea informaţiilor din corpusuri;

– accesul publicului la informaţiile conţinute. Soluţia DACOROMANICA ni se pare una dintre cele inspirate. Astfel, pentru unele documente, care nu se află sub incidenţa dreptului de autor sau pentru care drepturile au fost negociate cu moştenitorii de drept, acces este liber, iar documentele aflate sub incidenţa dreptului de autor sunt disponibile exclusiv *in situ* şi fără drepturi de copiere, pe site fiind afişate doar metadatele, pictograma şi, selectiv, primele 5 pagini din document, pornind de la pagina de titlu.

Publicul-ţintă căruia îi este destinată informaţia cuprinsă în corpusurile de limbă română se împarte în două categorii:

– obişnuit (neavizat, în sens nepeiorativ) – orice cunoscător sau persoană interesată de limba română, din ţară sau din străinătate;

[7] Proiect in curs de definitivare. Informaţii despre proiect se găsesc la următoarea adresă: https://consilr.info.uaic.ro/ edtlr/wiki/index.php?title=Despre_proiect.

– specializat (specialişti din diferite domenii, mai ales din domeniul filologiei şi al informaticii (avem în vedere mai ales specialiştii în prelucrarea limbajului natural).

O situație specială este reprezentată de situația informatizării *Dicționarului Tezaur al limbii române*. Aşa cum am arătat deja într-un articol anterior (Dănilă 2010), acest proiect, care este în curs de finalizare, ar trebui, pe de o parte, să pună la dispoziția tuturor cunoscătorilor sau celor interesați de limba română formatul electronic al *Dicționarului Academiei*, pe suport electronic – şi, poate ulterior, în funcție de schimbarea politicii lingvistice din România, şi on-line, în acces liber sau condiționat – şi, pe de altă parte, să pună la dispoziția deocamdată doar a cercetătorilor o arhivă electronică care să cuprindă toate textele din Bibliografia DLR.

Acest corpus uriaş pentru limba română pune şi limba noastră într-o situație de egalitate cu celelalte limbi care au făcut deja acest lucru (*Le Trésor de la Langue Française Informatisé* (TLFi – http://atilf.atilf.fr/); *Diccionario de la lengua espanola de la Real Academia Espagnola* (DRAE – http://buscon.rae.es/draeI/); *Tesoro della lingua italiana delle origini* (TLIO – http://tlio.ovi.cnr.it/TLIO/index2.html); *Oxford English Dictionary* (OED – http://www.oed.com/).

Rezultatele obținute prin eDTLR pot şi trebuie să fie continuate. Astfel, a apărut ideea necesității unui corpus lexicografic esențial. A fost propus un proiect cu titlul *CLRE. Corpus lexicografic românesc esențial. 100 de dicționare din bibliografia DLR aliniate la nivel de intrare şi la nivel de sens*, care este finanțat de CNCSIS (pentru perioada 2010 – 2013).

Colectivul de cercetare este format din:

– Elena Dănilă – lexicograf la Institutul de Filologie Română „A. Philippide", Iaşi;

– Marius-Radu Clim – lexicograf la Institutul de Filologie Română „A. Philippide", Iaşi;

– Ana-Veronica Catană-Spenchiu – doctorand la Facultatea de Litere, de la Universitatea „Alexandru Ioan Cuza", Iaşi;

– Marius Iulian Răschip – doctorand informatician la Facultatea de Informatică, de la Universitatea „Alexandru Ioan Cuza", Iaşi.

Obiectivele proiectului *CLRE* sunt:

1) realizarea unei baze de date care să cuprindă dicționarele esențiale din Bibliografia DLR, aliniate la nivel de intrare şi parțial la nivel de sens;

2) construirea unui mediu de programe care să permită consultarea interactivă a acestui corpus, care să se constituie într-un cadru modern de lucru şi cercetare lexicografică, uşor adaptabil la o diversitate de obiective (de exemplu, corectarea şi actualizarea prezentei ediții a DLR, precum şi redactarea viitoarelor ediții ale Dicționarului Academiei);

3) realizarea unei liste de cuvinte cvasi-exhaustive, pentru limba română, pornind de la corpusul aliniat;

4) creşterea vizibilității activității de cercetare a specialiştilor lingvişti şi informaticieni, în domeniul limbii române, prin promovarea mijloacelor informatice de prelucrare lingvistică create în proiect, prin studii, articole publicate în reviste de specialitate sau prin participarea la manifestări ştiințifice.

Astfel, prin acest proiect se vizează: realizarea unui corpus scanat, format din dicționarele de referință ale DLR (cu respectarea legislației în vigoare în ceea ce privește drepturile de proprietate intelectuala); scanarea si prelucrarea (OCR-izarea[8]; parsarea[9] textului la nivel de intrare şi, parțial, la nivel de sens) a acestor dicționare; realizarea unei interfețe on-line pentru validarea/corectarea parsării, precum şi validarea alinierii între textul *Dictionarului Tezaur al limbii romane* (in format electronic, rezultat al proiectului eDTLR) şi dicționarele de referință din Bibliografia DLR.

În lume există deja astfel de corpusuri:

– *Le rayon des dictionnaires* (http://www.atilf.fr/) – colecție de dicționare informatizate franceze, din secolul al XVI-lea până în secolul al XX-lea;

– *Nuevo tesoro lexicográfico de la lengua española* (http://buscon.rae.es/ntlle/SrvltGUILoginNtlle) – bază de date cuprinzând versiunile facsimilate ale tuturor dicționarelor editate şi publicate de Real Academia Española;

– *Das Wörterbuchnetz* (http://germazope.uni-trier.de/Projects/WBB/) – rețea de dicționare de limba germană, creată la universitatea Trier din Germania.

Pe lângă corpusul complex eDTLR şi pe lângă corpusul lexicografic creat pe bază de voluntariat, de nespecialişti (http://dexonline.ro/), limba română va dispune şi de un alt corpus lexicografic, profesionist, în sensul că va fi creat şi supravegheat de specialişti lexicografi şi informaticieni.

Diferența specifică pe care o aduce proiectul pe care îl propunem față de varianta http://dexonline.ro/ este reprezentată de:

– numărul mare de dicționare avute în vedere, însumând cca 150.000 pagini de dicționare, care trebuie incluse în baza de date şi aliniate la nivel de intrare şi parțial al nivel de sens;

– perspectiva istorică asupra lexicografiei româneşti şi, implicit, asupra limbii române pe care selecția de dicționare propusă de noi o presupune;

– faptul că ceea ce propunem noi este un corpus realizat după toate normele ştiințifice, realizarea lui de către o echipă mixtă, formată din lingvişti şi informaticieni, şi evaluarea rezultatelor de către specialişti români, consacrați în domeniu, asigurând calitatea cerută de exigențele unei cercetări competente şi adecvate.

Aşadar, ceea ce ne propunem să obținem prin intermediul CLRE este un corpus bine individualizat, superior atât cantitativ (32 dicționare în http://dexonline.ro/ şi 100 dicționare în CLRE), cât şi calitativ tuturor încercărilor de până acum.

În acest corpus vom include mai multe tipuri de dicționare:

1) *Dicționare generale*:

– DA = *Dicționarul limbii române*, tom I–II, Tipografia ziarului „Universul", Imprimeria Națională, Bucureşti, 1907-1944;

– DLR = *Dicționarul limbii române*, Serie nouă, tom VI–XIV, Editura Academiei, Bucureşti, 1965–2010;

– DEX = *Dictionarul explicativ al limbii romane*. Bucureşti, Editura Academiei, 1975;

[8] Convertirea din format imagine in format text.

[9] Identificarea automată a intrărilor din dicționarele scanate şi ocerizate anterior.

– DEXI = *Dicționarul explicativ ilustrat al limbii române*, Autori: Eugenia Dima, Doina Cobeț, Laura Manea, Elena Dănilă, Gabriela E. Dima, Andrei Dănilă, Luminița Botoşineanu, Chişinău, Editurile Arc şi Gunivas, 2007;

– MDA = *Micul dicționar academic*. Vol. I–IV. Bucureşti, Editura Univers Enciclopedic. Volumul I: A–C (2001); volumul al II-lea: D–H (2002); volumul al III-lea: I–Pr (2003); volumul al IV-lea: Pr–Z (2003);

– NDU = Ioan Oprea, Carmen-Gabriela Pamfil, Rodica Radu, Victoria Zăstroiu, *Noul dicționar universal al limbii române*. Bucureşti – Chişinău, Litera Internațional, 2006.

2) *Dicționare auxiliare* (care sunt strâns legate de redactarea Dicționarului Tezaur):

– A. de Cihac, *Dictionnaire d'etymologie daco-romane*. Vol. I. *Elements latins, comparés avec les autres langues romanes*, Francfort A.-M., Ludolphe St. Goar; Berlin, A. Asher; Bucarest, Socec, 1870. Vol. II. *Elements slaves, magyars, turcs, grecs-moderne et albanais*, Francfort, Ludolphe St. Goar; Berlin, S. Calvary; Bucureşti, Sotschek, 1879

– Alexandru Ciorănescu, *Dicționarul etimologic al limbii române*. Ediție îngrijită şi traducere din limba spaniolă de Tudora Sandru-Mehedinți şi Magdalena Popescu Marin. Bucureşti, Editura Saeculum I. O., 2002.

– *** *Dicționarul ortografic, ortoepic şi morfologic al limbii române*. Ediția a II-a revăzută si adăugită, Bucureşti, Univers Enciclopedic, 2005.

– Florin Marcu, *Noul dicționar de neologisme*. Bucureşti, Editura Academiei Române, 1997.

3) *Dicționare speciale* (enciclopedice ori dicționare speciale, alese după criteriul cantitativ ori după criteriul importanței lor pentru perspectiva diacronică asupra limbii):

– *Dicționar enciclopedic*. [Vol.] I: A–C (1993), [vol.] II: D–G (1996), [vol.] III: H–K (2000), [vol.] IV: L–N (2001), [vol. V]: O–Q (2004). [vol.] VI: R–S (2006). [vol.] VII: T–Z (2009). Bucureşti, Editura Enciclopedica;

– I.–Aurel Candrea – Gh. Adamescu, *Dicționarul enciclopedic ilustrat. Partea I: Dicționarul limbii române din trecut şi de astăzi* de I.-Aurel Candrea. *Partea II: Dicționarul istoric şi geografic universal* de Gh. Adamescu. Bucureşti, Editura Cartea Românească, [1926–1931];

– *Lexiconul tehnic român*. I ş. u. Elaborare nouă. Bucureşti, Editura Tehnică, 1957 ş.u.

În proiect se vor utiliza, astfel, atât metode lingvistice clasice / tradiționale (de exemplu, transliterarea intrărilor în alfabet chirilic sau de tranziție ori studiul comparativ, la nivel semantic, al dicționarelor), cât şi metode noi, de lexicografie computațională.

Dificultățile unui astfel de demers sunt legate de cantitatea mare de material care trebuie să fie tratat din punct de vedere informatic – parsat şi aliniat (cca 100 de dicționare – aproximativ 150.000 de pagini de dicționar), de situația dicționarelor scrise în alfabet chirilic sau în alfabet de tranziție, de problemele de copyright în cazul dicționarelor mai recente etc.

Perspective după încheierea proiectului CLRE:

– dezvoltarea de aplicații de anvergură privind dezambiguizarea semantică a cuvintelor;

– selecții de tipuri de intrări în vederea elaborării de noi dicționare specializate (tematice, etimologice etc.);

– corelarea cu alte resurse lingvistice ori multimedia, ceea ce ar aduce lexicografia românească la un nivel comparabil cu lexicografia europeană.

CLRE reprezintă, astfel, şi un punct de plecare pentru cercetări viitoare.

Concluzii

Importanța acestui corpus care ar trebui să fie accesibil tuturor cercetătorilor şi nu doar celor care lucrează la *Dicționar* – este indiscutabilă.

Rezultatul final al acestui proiect îl constituie un *Corpus lexicografic românesc esențial*, care va include un număr semnificativ de dicționare de bază ale limbii române, aliniate formal şi, parțial, semantic, ceea ce va pune la dispoziția specialiştilor români şi străini un excelent instrument de lucru indispensabil, mult timp aşteptat, foarte util pentru promovarea şi menținerea limbii române, ceea ce permite alinierea cercetării româneşti la standardele internaționale din domeniu.

Bibliografie

DA = *Dicționarul limbii române*, tom I–II, Tipografia ziarului „Universul”, Imprimeria Națională, Bucureşti, 1907–1944.

DLR = *Dicționarul limbii române*, Serie nouă, tom VI-XIV, Editura Academiei, Bucureşti, 1965–2010.

DRAE = *Diccionario de la lengua española de la Real Academia Española* – http://buscon.rae.es/draeI/.

TLFi = *Le Trésor de la Langue Française Informatisé* – http://atilf.atilf.fr/.

TLIO = *Tesoro della lingua italiana delle origini* – http://tlio.ovi.cnr.it/TLIO/index2.html.

OED = *Oxford English Dictionary* – http://www.oed.com/.

DWB = *Deutsches Wörterbuch "der Grimm"* – http://germazope.uni-trier.de/Projects/DWB.

Aldea, Dănilă *et alii* 2006: Bogdan-Mihai Aldea, Elena Dănilă, Corina Forăscu, Gabriela Haja, *Dicționarul limbii române (DLR) în format electronic. Aplicații*, în volumul *Comunicare interculturală şi integrare europeană*, volum îngrijit de Elena Dănilă, Ofelia Ichim, Florin-Teodor Olariu, Iaşi, Editura Alfa, p. 7–17.

Clim, Dănilă *et alii* 2008: Marius Clim, Elena Dănilă, Gabriela Haja, *Premise ale informatizării cercetării lexicografice academice româneşti* în volumul *Limba română. Dinamica limbii, dinamica interpretării*, Editura Universității din Bucureşti, p. 585–591.

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 vol. *Advances in Spoken Language Technology* (editors Corneliu Burileanu, Horia-Nicolai Teodorescu), Bucureşti, Editura Academiei Române, p. 195–206.

Curteanu, Moruz *et alii* 2008: Neculai Curteanu, Alexandru-Mihai Moruz, Diana Trandabăț, *Extracting Sense Trees from the Romanian Thesaurus by Sense Segmentation & Dependency Parsing*. In *Proceedings of the COLING 2008 Workshop on Cognitive Aspects of the Lexicon (COGALEX 2008)*, p. 55–63, Manchester, UK, 24 August, 2008.

Dănilă 2007: Elena Dănilă, *Tradiție și inovație în cercetarea lexicografică românească*, în volumul de la Colocviul Internațional de Științele Limbajului „Eugen Coșeriu", ediția a IX-a, Suceava (partea I), *Evoluția și funcționarea limbii – perspective normative în noul context european*, Editura Universității Suceava, p. 213–215.

Dănilă 2010: Elena Dănilă*, eDTLR – base de données et instrument pour la recherche lexicographique roumaine*, in „Philologica Jassyensia", An VI, Nr. 1 (11), 2010, p. 37–46.

Dănilă, Haja 2009: Elena Dănilă, Gabriela Haja, *Dicționarul limbii române în format electronic (eDTLR) în perspectiva globalizării*, în Actele Conferinței Internaționale *Paradigma discursului ideologic* (Galați, 29–30 aprilie 2009), publicat în «Communication interculturelle et littérature», nr. 1 (6), aprilie–mai–iunie, Galați, Editura Europlus, p. 269–273.

Haja 2007: Gabriela Haja, *Resurse electronice pentru cercetarea lexicografică românească*, în vol. *Limba română azi*, Actele celei de-a X-a Conferințe Naționale *Limba română – azi* (Iași – Chișinău, 3–7 noiembrie 2006), Iași, Editura Universității „Alexandru Ioan Cuza", p. 129–134.

Haja, Dănilă *et alii* 2005: Gabriela Haja, Elena Dănilă, Corina Forăscu, Bogdan-Mihai Aldea, *Dicționarul limbii române (DLR) în format electronic. Studii privind achiziționarea*, Iași, Editura Alfa, publicat și electronic pe www.consilr.info.uaic.ro.

Haja, Forăscu *et alii* 2006: Gabriela Haja, Corina Forăscu, Bogdan-Mihai Aldea, Elena Dănilă, *The dictionary of Romanian Language: steps toward the electronic version*. In *Proceedings of EURALEX 2006*, Torino, Italy, september 2006.

On the Importance of Creating a Romanian Essential Lexicography Corpus

This paper aims at highlighting the importance of creating a romanian essential lexicography corpus. The project presented in this paper has as purpose the valorisation of certain results from the complex project eDTLR, by using, as reference text for the alignment, the Thesaurus Dictionary in electronic format and especially creating a Romanian lexicographic corpus, which will contain 100 dictionaries (from the XVIth century to present day) aligned at entry and, partially, at meaning level.

At the same time, the lexicographic corpus realized by us will constitute an useful database for linguistic research, in the country or abroad, having the potential of being useful, at its turn, in future project for implementing and developing the Romanian fundamental humanist research according to international standards.

Institutul de Filologie Română „A. Philippide", Iaşi
România