

REDACTAREA ÎN FORMAT ELECTRONIC A DLR. CÂMPUL „DEFINIȚIE” – POSIBILITĂȚI DE ORGANIZARE

DIGITALISING THE ROMANIAN DICTIONARY: ALTERNATIVES FOR
ORGANISING THE DEFINITION FIELD.

(Abstract)

The present article discusses how different methods of electronic lexicography can be applied to the projects run by the Lexicology and Lexicography Department of the ‘Iorgu Iordan – Al. Rosetti’ Institute of Linguistics. Firstly, we briefly present the current state of the digitalization process that the *Dictionary of Romanian* (DLR) is undergoing. Then, we take a look at the ways in which other e-dictionaries and international projects organise the definition field. Finally, we analyse several examples in our attempt to see whether such means can be used for the Dictionary of Romanian (DLR) and the Explanatory Dictionary of Romanian (DEX).

Cuvinte-cheie: lexicografie electronică, definiție, informatizarea, dicționar.

Key-words: e-lexicography, definition, digitalisation, dictionary.

Dezvoltarea fără precedent a științei și a tehnologiei a dus la apariția unor noi subdomenii ale lingvisticii, cum ar fi neurolingvistica, dar și la noi direcții de cercetare în cadrul disciplinelor lingvistice deja consacrate. Nici lexicografia nu a scăpat de „virusul informatizării”: de la dicționare explicative unilingve și/sau bilingve la dicționare tezaur, toate tind să aibă cel puțin o variantă electronică, astfel încât acest tip de redactare a dicționarelor este o practică deja verificată de tradiția lexicografică occidentală.

În 2013, departamentul de Lexicologie și lexicografie al Institutului de Lingvistică „Iorgu Iordan – Al. Rosetti” a început să folosească aceste noi tehnologii pentru redactarea *Dicționarul limbii române* (DLR). Finalitatea

acestei întreprinderi este o formă electronică a DLR, care va fi accesibilă și online.

Redactarea în format XML presupune un efort suplimentar de organizare a informației, pentru ca varianta electronică să permită utilizatorului mai multe tipuri de căutare și, deci, mai multe modalități de cercetare. Acest efort de organizare a informației îi este util și redactorului din motivele arătate mai jos, sintetizate într-o prezentare succintă a procesului de redactare în această nouă variantă.

Dacă în prima parte am arătat „ce se poate face deja”, în a doua parte am îndrăznit să ne gândim „ce s-ar mai putea face” în plus. Cum această întreprindere este abia la început de drum, am încercat să propunem câteva idei în ceea ce privește modul de organizare a câmpului „definiție” în acest nou format.

Plecând de la tipurile de definiții existente în forma tradițională a DLR, am sugerat moduri în care utilizarea programelor lexicografice ar putea ușura munca de redactare, eliminând, pe cât posibil, eroarea umană. Am arătat cum restricțiile pe care aceste programe le impun pot fi folosite drept avantaj și nu ca elemente care să îngreuneze redactarea. Am avut ca exemple modul în care alte dicționare în format electronic au ales să organizeze câmpul „definiție” și, păstrând specificul DLR, am testat dacă este posibilă și o sistematizare (și rigidizare) a tipului de definiții întâlnite în acest dicționar și, apoi, dacă vreuna dintre soluțiile oferite de lexicografii englezi sau francezi poate fi profitabilă pentru a doua lucrare importantă a departamentului de Lexicografie, anume *Dicționarul explicativ al limbii române* (DEX).

1. Forma electronică a DLR: câteva concepte și stadiul actual

Informatica s-a insinuat în redactarea dicționarelor încă din anii 1960, prin informatizarea dicționarelor Webster (site-ul Webster); în același timp a început și folosirea băncilor de texte electronice, care au permis extragerea automată a citatelor din sute, chiar mii de opere literare – proiectul GAMMA 60, utilizat din 1964, strămoșul actualului FRANTEXT, a permis extragerea automată a 430 000 de citate (TLFi 2004: 2). Dimensiunea bazei de date obținute prin despuieră automată a textelor nu a constituit un obstacol pentru redactori, întrucât datele puteau fi ordonate după mai multe criterii: ordinea cronologică a extraselor, ordinea alfabetică a contextelor de la stânga sau de la dreapta cuvântului, tipuri de construcții sintactice.

Practica lexicografică este cu mult mai veche decât era informatică, astfel încât dicționarele de referință din lexicografia europeană au apărut mai întâi în format tipărit. Informatizarea unui dicționar nu înseamnă culegerea lui ca simplu document electronic, în care stilurile (ex.: caracterele aldine, cursive,

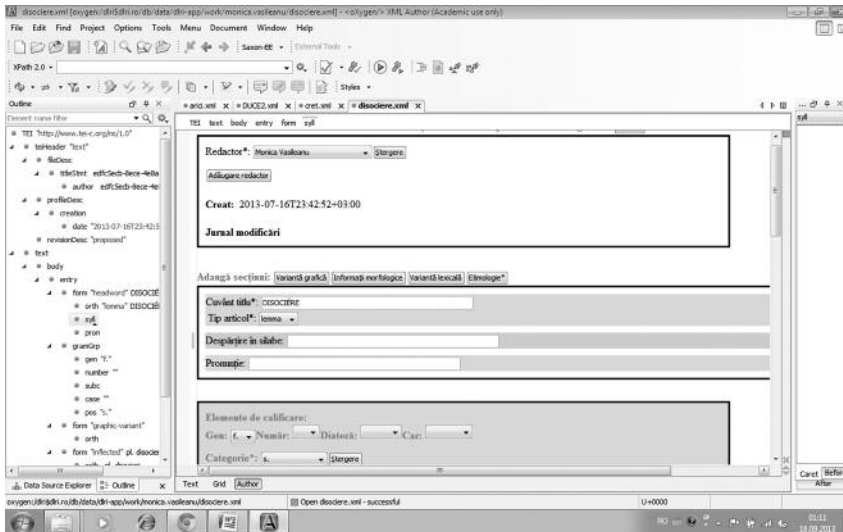
majuscule etc.) fiecărui tip de informație ar trebui aplicate manual, ci transformarea lui într-un document structurat (TLFi 2004: 10; Queens, Reker-Hamm 2003: 1). Pentru a transforma un document electronic simplu într-un document structurat este nevoie: a) de o segmentare a documentului, b) de adnotarea segmentelor, c) de un program care să transforme adnotările în stiluri aplicate segmentului adnotat. Odată adnotate segmentele, aceste adnotări pot fi folosite de program în mult mai multe scopuri decât simpla aplicare a stilurilor – stiluri care într-un dicționar sunt purtătoare de semnificații (ex. cuvântul-titlu se marchează în DLR prin majuscule aldine, prin urmare cititorul știe că acele cuvinte scrise cu majuscule aldine sunt cuvinte-titlu, așadar în funcție de aceste stiluri se cataloghează informația). Pentru ca aceste adnotări să poată fi procesate de mai multe programe, a fost necesară standardizarea lor și deci crearea unei liste de asemenea adnotări. Întrucât aceste adnotări transformă documentul într-un fel de hartă, marcând și delimitând tipurile de informație, au mai fost numite și balize între lexicografi.

Standardizarea lor s-a petrecut în anii 1980, prin recunoașterea de către ISO (International Organization for Standardization) a SGML (=Standard General Markup Language) drept un standard în tehnologia informatică (Renear 2004: 226). Din această tehnologie standard s-au dezvoltat limbajele HTML și XML, utilizate pe scară largă astăzi mai ales în crearea paginilor de internet. Limbajul XML și dezvoltat de TEI (Text Encoding Initiative), un consorțiu care menține și dezvoltă standardele pentru reprezentarea textelor în formă digitală, fiind axat pe disciplinele umaniste; limbajul XML este utilizat în cataloagele bibliotecilor, ale muzeelor, în băncile de texte și în lexicografie. Comunitatea TEI organizează conferințe și ateliere, punând la dispoziția utilizatorilor și programele care pot procesa balizele-standard acceptate de TEI (www.tei-c.org). O prezentare a acestor balize și a ierarhiei lor există pe site-ul TEI, sub forma unui ghid foarte cuprinzător și... stufos.

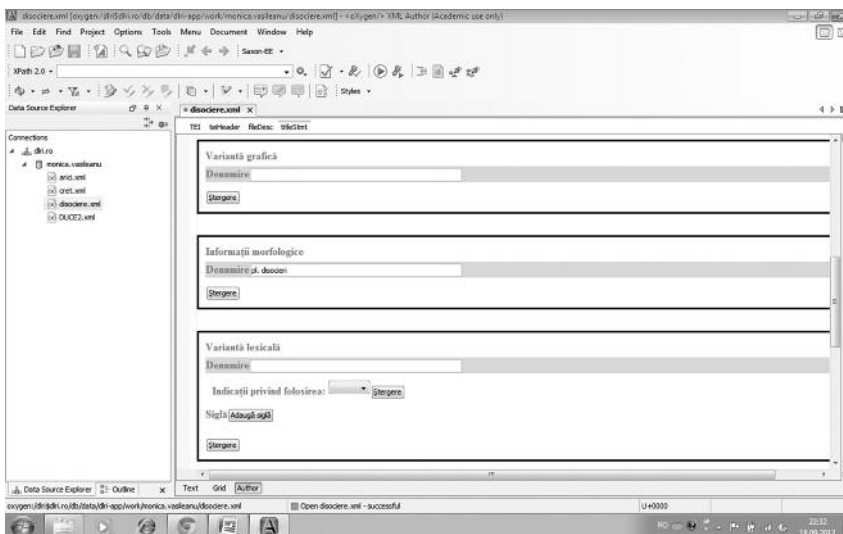
În informatizarea DLR se folosește limbajul XML, procesat prin programul Oxygen. Astfel, balizele folosite pentru transformarea articolelor de dicționar în documente structurate sunt cele din ghidul TEI. Întrucât DLR este un dicționar complex și conține multe tipuri de informație, informaticianul Claudiu Teodorescu a pregătit o interfață care să ascundă etichetele și să faciliteze astfel munca redactorilor. Prezentăm mai jos câteva capturi de ecran, așa cum arătau ele în luna noiembrie 2013. De atunci, interfața a suferit și va mai suferi modificări datorate specificului *DLR-ului*.

(1) Există posibilitatea de a vedea și partea tehnică, în format pur XML, dar lexicografului, lingvist de profesie și nu informatician, îi este mai util să lucreze în interfața „Author”. Pentru a avea o evidență riguroasă a muncii fiecărui redactor, primele câmpuri conțin metadatele: Se înregistrează

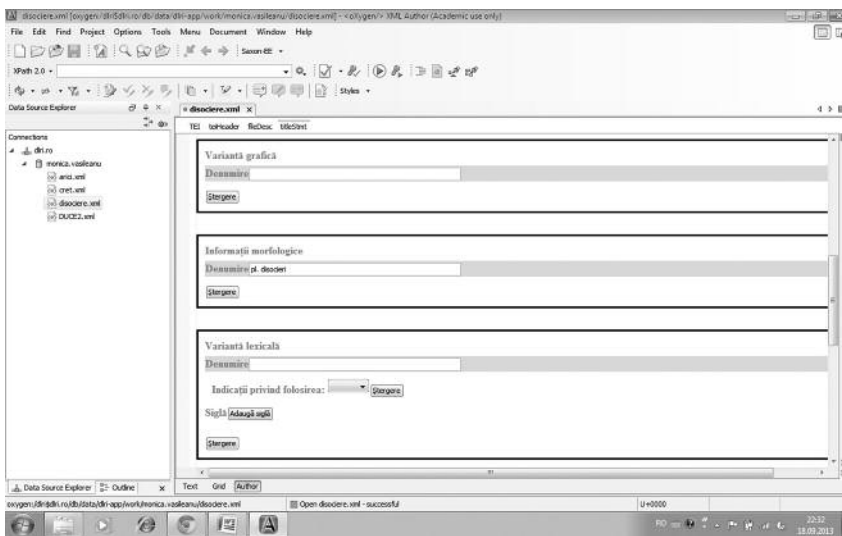
numele redactorului (într-o fază ulterioară al revizorului), data creării, data modificării etc.



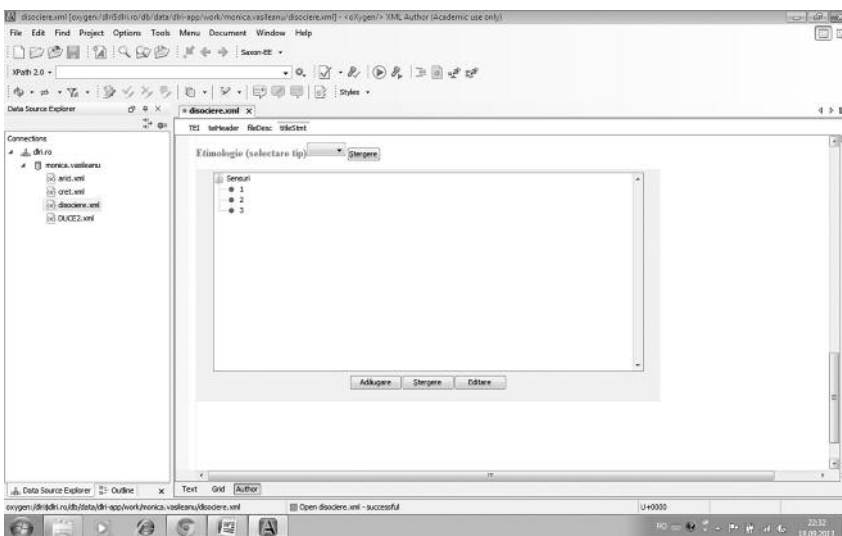
(2) Urmează câmpurile specifice dicționarului propriu-zis: cuvântul-titlu, tipul articolului (lemă sau variantă); astfel, variantele ar putea fi inserate automat la ordinea alfabetică, iar corelarea cu cuvântul-titlu s-ar face și ea automat. Sunt prezente și câteva elemente de ortoepie, cum ar fi despărțirea în silabe, de obicei menționată în cazul succesiunilor vocalice pentru a distinge diftongii de hiat.



(3) Detaliile din paragraful penultim: variantă grafică („Scris și:”), informația morfologică (forme de plural pentru substantive, de prezent, conjunctiv, perfect simplu, participiu pentru verbe etc.), variantele lexicale („Și:”) sunt urmate de etimologie.



(4) Informația semantică are aspectul unui arbore de sensuri. Butoanele permit adăugarea de subsensuri și subsensuri – numărul de trepte este probabil infinit, însă în DLR există maximum 5 trepte.



(5) Definiția propriu-zisă este introdusă într-un câmp unic, nesegmentat. Fiecărei definiții îi urmează o trimitere (opțională) către alt cuvânt-întrare înrudit, indicații privind folosirea (mărci diastratice, indicații de domeniu, restricții combinatorice etc.) – în forma tipărită, acestea se menționează înaintea definiției) și citatele ilustrative. Siglele citatelor vor fi introduse automat, prin selectare dintr-un meniu derulant.

Cele mai multe dintre aceste câmpuri pot fi multiplicare, pentru a putea astfel cuprinde informația complexă din DLR.



2. Câmpul definiție: posibilități de structurare

În secțiunea precedentă am arătat ce s-a făcut deja pentru informatizarea DLR. Adaptorul mai are nevoie de câteva adaosuri, majoritatea fiind conținuturi pe care programul să le prezinte în meniuri derulante din care redactorul trebuie să aleagă (ex. în câmpul etimologie, în câmpul domeniu, și, mai ales, în lista de sigle), însă „gramatica” programului a fost definită.

Am văzut, în capturile de ecran de mai sus, că definiția este redactată într-un câmp unic, nesegmentat, nerestricționat. Ne-am gândit la două posibilități de structurare a acestui câmp, posibilități exploatare de alte dicționare europene sau aflate încă în stadiul de cercetare, cu aplicații multiple:

a. Segmentarea câmpului „definiție” în mai multe subcâmpuri, precum „gen proxim” + „diferențe specifice”. Această idee presupune segmentarea

și balizarea informației semantice, operațiune dificilă pentru redactor, dar care deschide noi perspective de cercetare;

b. Restricționarea câmpului „definiție” prin stabilirea unui număr finit de cuvinte care să poată fi folosite în definiții. Lexicograful presupune că aceste cuvinte folosite în „definiția controlată” sunt cunoscute deja de utilizator. Această practică este specifică dicționarelor cu scop didactic – *learners' dictionaries* (Geeraerts 2003: 91).

2.1 Segmentarea câmpului definiție

Întrucât dicționarele de referință din lexicografia europeană au apărut mai întâi în format tipărit și abia apoi în format electronic, definițiile au fost redactate fără ajutorul uneltelor electronice. Redactorii au aplicat „șabloane de definire”, redactând definiții-tip de fiecare dată când a fost posibil, însă acesta a fost un proces efectuat de oameni în mod tradițional, nu automatizat. Asemenea definiții-tip au fost aplicate numai unui număr finit de cuvinte. În ce măsură se poate extinde această practică și în ce măsură se poate automatiza vom vedea pornind de la două proiecte, unul al lexicografiei anglo-saxone, DANTE, iar celălalt al lexicografiei franceze, Definiens.

2.1.1 DANTE și Definiens – două modele

DANTE (www.webdante.com), probabil baza de date care conține cea mai rafinată analiză lexicală a limbii engleze (Rundell 2012: 20), este folosit actualmente de lexicografi și cu acest scop, de a crea definiții-tip („proforma definitions”) pentru un număr cât mai mare de cuvinte. Proiect complex, care combină segmentarea definițiilor deja existente cu analiza pe un corpus imens de limbă engleză, DANTE a fost creat și dezvoltat ca o unealtă lexicografică prin care eficiența lexicografului poate fi sporită, cu costuri minime. Două dicționare relativ recente (din 2007) au folosit resursele DANTE pentru a crea definiții-șablon: *Oxford-Hachette English-French Dictionary* și *Macmillan English Dictionary for Advanced Learners*. Nu numai zilele săptămânii, elementele chimice sau literele alfabetului au primit definiții care urmează același tipar pentru toți membrii câmpului lexico-semantic. DANTE a creat 68 de definiții-șablon, aplicabile nu doar cuvintelor monosemantice, ci și celor polisemantice. Spre exemplu, în cazul băuturilor, șablonul includea un sens al substantivului masiv, defectiv de plural (băutura propriu-zisă) și un alt sens, obținut prin metonimie, „un pahar/o sticlă de” + băutura, situație în care substantivul are și forme de plural (Rundell 2012: 24). Mai mult decât atât, programele de redactare a dicționarelor sunt capabile să identifice automat membrii câmpului și încarcă șablonul automat, astfel încât redactorul care deschide cuvântul-intrare găsește deja șablonul de definiție cu câteva date gata

introduse (Rundell 2012: 24). Munca lexicografului devine cu totul diferită cu asemenea programe: el nu trebuie să creeze definiții, ci să le corecteze – uneori destul de sever – și să le completeze pe cele generate automat. Performanțele resurselor grupate în proiectul DANTE au schimbat modul de a concepe lexicografia, însă instrumentele folosite sunt aplicate doar limbii engleze.

Lexicografia franceză (și nu numai) parcurge un drum invers nu de redactare a definiției după un șablon, ci de extragere a șablonului din definiții deja redactate. Un grup de cercetători ai celebrului Institut de linguistique française se ocupă, în cadrul programului *Definiens*, cu segmentarea și adnotarea definițiilor din *Trésor de la langue française informatisé*, fără a rescrie sau a revizui aceste definiții (cel puțin pentru moment), cu scopul de a forma o bază de date cu definiții structurate, de mare importanță pentru toate cercetările lexicale și privitoare la tratamentul automat al francezei (site-ul *Definiens*). Spre deosebire de programele de parsare folosite asupra limbii engleze, proiectul *Definiens* nu a putut folosi cu succes un *parser*: segmentarea și adnotarea se fac în mare parte manual, întrucât programele de adnotare automată au avut rezultate în mare parte eronate, iar corectarea acestora durează la fel de mult ca adnotarea manuală (Barque, Nasr, Pologuère 2010: 5).

Ghidul de adnotare (Barque, Pologuère 2012) cuprinde instrucțiuni pentru lingviștii implicați în proiect și o listă a balizelor folosite. Segmentarea definițiilor se petrece doar în cazul definițiilor analitice, definițiile prin sinonime sau prin serie sinonimică fiind balizate ca atare:

FLAGADA =

<DEFI><PARAPH_SYNOS>Fatigué, avachi, sans force</PARAPH_SYNOS>.</DEFI>

(Barque, Pologuère 2012: 5)

Segmentarea definițiilor analitice are mai multe etape. Într-o primă etapă, lingvistul trebuie să detașeze „componenta centrală” de „componentele periferice”, adică „genul proxim” de „diferențele specifice” (Barque 2010: 3):

BROUETTE (sense B.1): <PARAPH><CC>Véhicule</CC> <CP>à une roue et à deux
brancards</CP> <CP>servant au transport des matériaux</CP></PARAPH>

(Barque, Nasr, Pologuère 2010: 3)

Alegerea componentei centrale se dovedește o acțiune dificilă întrucât presupune o bună cunoaștere a dicționarului. Componenta centrală nu se stabilește doar pe criteriile sintactice (să fie centrul de grup al parafrizei-definiției), ci prin comparație cu alte cuvinte din același câmp lexico-semantic. Spre exemplu, în cazul cuvântului *brouette*, lingvistul oscilează între „véhicule” și „véhicule à une roue”. Dacă „véhicule à une roue” ar mai fi găsită ca centru al altei definiții, atunci întreaga sintagmă ar fi acceptată ca gen proxim. Verificarea se poate face prin căutarea sintagmei în câmpul definițiilor din TLF:

Recherche d'un mot

Recherche assistée

Cherchez tous les passages de TLF ayant une ou plusieurs des propriétés ci-dessous.
(Voir des exemples d'allocation de formulaire)

1) Le passage est consacré à une vedette égale ou contenant un mot donné

2) Le passage est consacré à une vedette ayant un code grammatical donné
Choisissez le code : INDIFFERENT

3) Le passage est consacré à une discipline donnée
Choisissez pour choisir la discipline
Nombre de disciplines actuellement sélectionnées : 0

4) Le passage est consacré à un indicateur d'emploi
Choisissez le type d'emploi : INDIFFERENT

5) Le passage doit contenir au moins un objet textuel de type et de contenu donnés

5.a) Indiquez le type de l'objet recherché : Définition (Voir la signification des types d'objets)

5.b) Indiquez le ou les contenus que l'on doit trouver dans l'objet (ligne "Oui") ou que l'on ne doit pas trouver (ligne "Non").

	Contenu 1 <input type="text"/>	Contenu 2 <input type="text"/>	Contenu 3 <input type="text"/>
Oui	lettre	à une roue	
Non			

Valider

Întrucât singurul articol găsit în urma acestei căutări a fost *brouette*, genul proxim indicat rămâne „véhicule”. Însă și această căutare poate fi înșelătoare, întrucât adeseori același conținut semantic poate fi redat prin cuvinte diferite:

LIEU(2) =

<DEFI><PARAPH><CC>Poisson marin</CC>...</PARAPH></DEFI>

BAR(1) =

<DEFI><PARAPH><CC>Poisson de mer</CC>...</PARAPH></DEFI>

(Barque, Pologuère 2012: 10)

Sau aceeași componentă centrală, exprimată prin aceleași cuvinte, poate fi discontinuă într-un caz și continuă în altul:

RÉGLLETTE =

<DEFI><PARAPH><CP>Petite</CP> <CC>règle graduée</CC>...</PARAPH></DEFI>

ÉCHELLE DE MARÉE =

<DEFI><PARAPH><CC>Règle <CP>verticale</CP>graduée</CC>...</PARAPH></DEFI>

(Barque, Pologuère 2012: 10)

Urmează apoi specificarea fiecărei componente, centrale sau periferice printr-o etichetă:

BROUETTE (sense B.1): <PARAPH><CC=véhicule>Véhicule</CC> <CP=parties caractéristiques>à une roue et à deux brancards</CP> <CP=fonction>servant au transport des matériaux</CP></PARAPH>

(Barque, Nasr, Pologuère 2010: 6)

ACCÉLÉRATEUR (acception II)

```
<DEFI><PARAPH étiqu=dispositif><CC>Dispositif </CC>
<CP rôle=fonction>contrôlant la puissance du moteur
<CP rôle=fonctionnement>en réglant son alimentation en gaz carburés</CP>
</CP></PARAPH></DEFI>
```

(Barque, Pologuère 2012: 5)

În 2010, existau deja 790 de etichete formalizate (Barque, Nasr, Pologuère 2010: 6), obținute în urma segmentării unui eşantion mare de definiții; cu siguranță, lista nu este exhaustivă și până la balizarea tuturor definițiilor analitice din TLFi vor mai fi adăugate și altele.

Următoarea etapă constă în asocierea fiecărei etichete cu un rol (Barque, Nasr, Pologuère 2010: 6), etapă încă neaplicată.

Prin această balizare se obține o bază de date cu definiții formalizate, care pot fi folosite de alte programe de procesare a limbajului natural, programe care sunt mult mai avansate în cercetarea aplicată englezei (Barque, Nasr, Pologuère 2010: 7).

2.1.2 Aplicabilitate în lexicografia românească

Ceea ce am văzut mai sus s-a aplicat sau se află în curs de aplicare în țări cu o tradiție lexicografică mai îndelungată, cu bugete mai mari alocate cercetării lingvistice și cu o mai mare experiență în procesarea automată a limbii respective. În ce măsură se pot aplica aceste idei în lexicografia românească? Vom discuta cele două dicționare principale elaborate de departamentul de Lexicologie și lexicografie, DLR și DEX. Începem cu DLR, deoarece redactarea în format XML a acestui dicționar este acum tema principală a departamentului de Lexicologie și lexicografie. Complexitatea ridicată a DLR face ca soluțiile găsite pentru automatizarea acestui dicționar să se aplice fără probleme și la DEX.

DLR este dicționarul-tezaur al limbii române, așadar este cea mai cuprinzătoare operă lexicografică românească, incluzând nu doar cuvintele curente, ale limbii literare contemporane, ci și cuvinte ieșite din uz, regionale, populare, cuvinte efemere, creații ale unor scriitori, termeni meșteșugărești, termeni tehnico-științifici (cu anumite restricții) și chiar cuvinte cu sens neprecizat. Limitele largi ale dicționarului înseamnă și o complexitate semantică greu de formalizat.

Normele de redactare a DLR enumeră trei tipuri principale de definiții. Vom vedea, acestea se aplică numai cuvintelor pline semantic și cu anumite restricții:

- definiția analitică;
- definiția prin serie sinonimică;
- definiția prin sinonim unic.

Definiția analitică este procedeul de bază al descrierii semantice (Norme 121). A fost concepută încă de la început ca o definiție structurată, în funcție de mai multe criterii. Astfel, fiecă parte de vorbire urmează un anumit tipar sintactic al definiției. Definiția analitică a verbelor începe cu „a”, marca infinitivului, cea a adjectivelor, cu „care”.

Întrucât numai definiția analitică poate fi segmentată, acesta va fi singurul tip de definiție discutat în prezenta secțiune. Excludem, așadar, definițiile prin serie sinonimică; definițiile prin sinonim unic, aplicate în special cuvintelor neliterare (Norme 155), pot fi incluse în discuție într-o fază ulterioară, preluând definiția analitică a sinonimului literar.

Normele DLR recomandă structurarea definiției în gen proxim și diferențe specifice (Norme 127-129). Motivul este de natură ontologică: este considerată definiția care reflectă cel mai bine realitatea, constituind o descriere succintă, care permite identificarea obiectelor (Fascicula 0: XI). Cu toate acestea, avertizează autorii Normelor, „de cele mai multe ori însă, diferitele nuanțe sensibile într-un sens nu vor putea fi subordonate unui gen proxim” (Norme 129).

Anumite serii de cuvinte, în majoritate substantive și verbe, primesc definiții-tip: zilele săptămânii, numele lunilor, gradele de rudenie, gradele militare, numele de plante și animale, majoritatea abstractelor etc. (Fascicula 0: XI). Cu toate acestea, diferența mare de timp între momentele redactării definițiilor unor cuvinte aflate în aceeași serie face ca șablonul să sufere mici modificări. Exemplificăm întâi cu zilele săptămânii, o serie finită și ușor de uniformizat:

LUNI: Prima zi a săptămânii, care urmează după duminică.

MIERCURI: Ziua a treia a săptămânii, care urmează după marți.

MARȚI: Ziua a doua a săptămânii, care urmează după luni.

SÂMBĂȚĂ: Ziua a șasea a săptămânii, care urmează după vineri. V. sabat¹ (2).

DUMINICĂ: Ultima (sau prima) zi a săptămânii, considerată zi de odihnă, în civilizațiile creștine.

Șablonul stabilit în primele definiții ar fi obligat să se menționeze numele zilei precedente în fiecare definiție, ceea ce nu se mai respectă în articolul duminică; în plus, paranteza din definiție creează o ambiguitate, întrucât dacă duminică e socotită prima zi a săptămânii, nu se poate ca și luni să aibă același statut... Această paranteză din definiția cuvântului „duminică” s-a introdus, cu siguranță, prin raportare la sisteme de numărare a zilelor săptămânii inaccesibile redactorilor DLR în momentul redactării literei M: calendarele anglo-americane încep săptămâna cu duminică și, mai nou, și calendarele bisericesti, viziuni nepermise de regimul politic din perioada redactării literei M.

Cu aceste observații ajungem la primul avantaj al existenței unui câmp-definiție segmentat: posibilitatea observării neregularităților existente în vederea revizuirii, proces mult mai simplu și mult mai puțin costisitor în era electronică.

Propunerea noastră concretă vizează segmentarea câmpului definiție, câmp unitar, nestructurat, segmentare care poate fi făcută la mai multe niveluri:

- la început, redactorul poate opta pentru un câmp al definiției analitice (structurat) sau pentru un câmp liber, în care să introducă serii sinonimice, sinonime unice sau alte tipuri excepționale de definiție (definiții metalingvistice pentru instrumentele gramaticale, definiții-citat preluate din glosare sau comunicări, pentru cuvintele obscure, pentru care nu s-a putut formula o definiție proprie în lipsa unor izvoare edificatoare – vezi Norme 138, 141);

- dacă selectează o definiție analitică, redactorul va găsi două câmpuri, de gen proxim și diferențe specifice;

- în cazul cuvintelor din aceeași serie, cum a fost cazul zilelor săptămânii, indicarea genului proxim să acționeze ca o constrângere asupra câmpului rămas liber pentru diferențele specifice. Spre exemplu, odată ce am introdus genul proxim „zi a săptămânii”, câmpul diferențelor specifice să se segmenteze automat în mai multe subcâmpuri: unul, etichetat „ordinea”, să fie completat cu numeralul ordinal („prima”, „a doua”, „a șasea” etc.), al doilea, etichetat „precedent”, să conțină formularea „care urmează după ...”, redactorul urmând să introducă manual numele zilei precedente, și un al treilea câmp, lăsat liber, pentru a putea adăuga alte note semantice specifice unui singur membru al seriei – cazul substantivului „duminică”, în definiția căruia găsim „considerată zi de odihnă în civilizațiile creștine”;

- asemenea șabloane am putea avea în cazul plantelor: odată introdus genul proxim „plantă”, câmpul diferențelor să se segmenteze automat în „familie”, „detalii descriptive”, „utilitate” și un câmp liber, lăsat pentru detaliile specifice doar unuia dintre membrii seriei.

Cu siguranță, este nevoie de o segmentare a câmpului definiție în cazul definițiilor lărgite (Norme 145), care reunesc de fapt două sensuri diferite, dar apropiate, înrudite, precum în exemplul următor:

DOVADĂ 1. Confirmare, demonstrare, probare a unui adevăr, a unui fapt, a unei afirmații etc.; (concretizat) semn, mărturie, argument etc. în sprijinul (sau împotriva) cuiva sau a ceva ori pentru confirmarea unui fapt, a unei afirmații, a unei presupunerii etc.

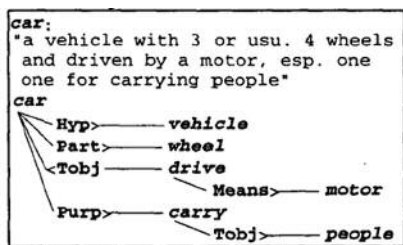
Avantajul actualei echipe de redactori ai DLR este că pot reformula definițiile. Segmentarea câmpului, în loc de constrângere, poate deveni un ajutor în uniformizarea formulărilor. Existența acestor subcâmpuri structurează datele pe care redactorul trebuie să le introducă în definiție și le ierarhizează.

Însă această segmentare mai aduce ceva în plus: permite formalizarea informației semantice. Astfel, dacă redactorul ar trebui – eventual într-o etapă ulterioară – să adauge fiecărei diferențe specifice și o etichetă, acest proces ar deschide direcții noi în cercetarea lexicului limbii române.

2.1.3 Utilitatea segmentării definițiilor – direcții noi în cercetarea lingvistică

Segmentarea semantică, idee dezvoltată de lexicologia structuralistă franceză, care viza descompunerea sensului în seme, ar putea fi formalizată prin aceste câmpuri și, mai ales, prin etichetarea lor. Am văzut deja un proces anevoios întreprins de o mică echipă de lingviști francezi asupra definițiilor din TLFi, proces care urmărește a) segmentarea definițiilor și b) etichetarea componentelor semantice, pentru a face un inventar al componentelor semantice.

Același proces de segmentare și etichetare, ba chiar cu câțiva pași în plus, a fost efectuat în câteva interesante proiecte recente asupra limbii engleze, prin folosirea unor programe speciale, numite *parsers*. Încă din anii 1990, definiții din diverse dicționare (preferate fiind *learners' dictionaries*, vom vedea mai jos de ce) au fost segmentate cu scopul de a găsi niște relații care să configureze un fel de „gramatică” a definițiilor. Pe baza acestei gramatici s-ar putea stabili relațiile semantice existente între toate cuvintele unei limbi. Proiecte internaționale precum MindNet sau SIMuLLDA au efectuat segmentarea definițiilor în lanțuri de câte trei termeni (două cuvinte și indicarea relației). „Gramatica” semantică ar trebui să conțină un număr finit de relații. Iată mai jos un exemplu de definiție segmentată prin MindNet și tabelul relațiilor:



(Richardson, Dolan,
Vanderwende, 1998: 1099)

Attribute	Goal	Possessor
Cause	Hypervnm	Purpose
Co-Agent	Location	Size
Color	Manner	Source
Deep Object	Material	Subclass
Deep Subject	Means	Synonym
Domain	Modifier	Time
Equivalent	Part	User

Table 1. Current set of semantic relation types in MindNet

Richardson, Dolan, Vanderwende,
1998: 1099)

Scopul proiectului MindNet, elaborat de Microsoft, este să obțină o rețea semantică pornind de la dicționare și enciclopedii. Engleza nu este singura limbă vizată, iar aplicarea acestei „gramatici semantice” asupra altor limbi ar

aduce o eficiență sporită traducerilor automate și tuturor programelor de procesare a limbajelor naturale.

SIMuLLDA a început ca o cercetare a unui consorțiu al Comisiei Europene (Martin 2004: 176), în care s-a propus modelul „hub and spoke” („trunchi și ramuri”). În mod previzibil, scopul proiectului a fost producerea de dicționare bilingve cu un ajutor cât mai mare din partea calculatoarelor. În SIMuLLDA s-a urmărit asocierea mai multor lexicoane monolingve pentru crearea de lexicoane bilingve; cu cât mai mare numărul de lexicoane monolingve asociate, cu atât mai mare – chiar exponențial – crește și numărul de dicționare bilingve care pot fi elaborate automat. Pentru a putea stabili relații de echivalență lexicală cât mai exacte între cât mai multe limbi, creatorii proiectului au preferat să ofere uneia dintre limbi statutul de hub („trunchi”), iar celelalte să constituie „ramurile”. Astfel, relațiile dintre limbile-ramuri pot fi stabilite prin calcule, pe baza relației fiecăreia dintre ramuri cu limba-trunchi (Martin 2004: 177). Avantajul este că pot fi asociate, într-un dicționar bilingv, limbi care n-au fost niciodată studiate contrastiv. Din asocierea a cinci lexicoane rezultă zece posibile dicționare; prin modelul hub and spoke, numai patru trebuie elaborate de oameni (cele care leagă limba-trunchi de fiecare dintre cele patru ramuri), celelalte șase putând fi generate automat pe baza calculelor (Martin 2004: 179).

Pentru fiecare limbă a trebuit să se stabilească un inventar de „unități de formă” și „unități de sens”, urmându-se asocierea lor (Martin 2004: 177). Asocierea unei unități de formă cu o unitate de sens diferă de la o limbă la alta, aceste diferențe de asociere constituind una dintre mărcile de specificitate ale fiecărei limbi. Legătura dintre formă și sens trebuia să fie dublată și de alte informații, precum registrul stilistic (Martin 2004: 179). Fără aceste informații, nu se pot stabili echivalențe lexicale satisfăcătoare.

Bineînțeles, aceste dicționare bilingve obținute prin SIMuLLDA trebuie verificate manual, întrucât programul se folosește de etichetele formalizate ale unor relații semantice, dar multe cuvinte au și note semantice individuale, care nu pot fi formalizate. De asemenea, decupajul lexical dintre limbi fiind diferit, rămân adeseori goluri lexicale – anumitor unități de sens nu le sunt asociate, în unele limbi, și unități de formă, în vreme ce alte limbi pot face această asociere (Martin 2004: 184).

Toate operațiunile descrise se bazează pe lexicoane monolingve obținute prin segmentarea și adnotarea definițiilor din dicționarele de referință ale fiecărei limbi (Martin 2004: 186). Definițiile sunt văzute de cercetătorii din proiectul SIMuLLDA ca niște ansambluri de lanțuri de attribute, asemănător cu lanțurile propuse de MindNet.

Cert este că, pentru ca proiecte similare să poată include și limba română, definiția nu mai poate fi văzută ca un câmp indivizibil, iar redactorul trebuie să aibă mereu în minte sau, mai bine, sub ochi, oferită de calculator, o structură a definiției, cu evidente „spații de rezervă”.

2.2 Restricționarea definițiilor – „definiții controlate”

Definițiile controlate sunt acele definiții care folosesc un număr limitat de cuvinte, considerate a fi cunoscute în cea mai mare măsură utilizatorilor. Procedul este foarte des întâlnit în lexicografia britanică, fiind folosit pentru prima dată în *New Method English Dictionary* de Michael West and James Endicott, în anul 1935. La redactarea definițiilor din acest dicționar autorii au folosit doar 1490 de cuvinte, considerate cunoscute de orice utilizator. În 1978, la redactarea *Longman Dictionary of Contemporary English* s-au folosit 2000 de cuvinte, selectate pe baza unei liste (General Service List) creată tot de Michael West (1953), listă care conținea cele mai frecvente 2000 de cuvinte din limba engleză. Alte dicționare explicative monolingve (dedicate atât utilizatorilor nativi, cât mai ales celor care învață limba engleză sunt: *Macmillan English Dictionary for Advanced Learners*, care utilizează la redactarea definițiilor aproximativ 2500 de cuvinte, *Oxford Advanced Learner's Dictionary*, cu aproximativ 3000 cuvinte și *Cambridge International Dictionary of English*, care folosește în definiții chiar mai puțin de 2000 de cuvinte.

Un prim avantaj este acela că lexicografii folosesc la redactarea definițiilor cuvinte cu o frecvență mare, pe care majoritatea utilizatorilor le cunosc, fapt care face ca dicționarele să fie mai ușor de utilizat atât de către vorbitorii nativi, cât și de cei nenativi. De aceea, în titlul acestor lexicoane se găsește adeseori clar menționat scopul lor didactic în sintagma *learner's dictionary*. Controlarea definițiilor exclude folosirea arhaismelor sau a regionalismelor. În general, se folosesc sensurile cele mai frecvente, sensurile de bază ale cuvintelor, evitându-se sensurile figurate. De asemenea, redactorii evită cuvintele care ar putea fi confundate cu alte cuvinte din limba-țintă sau din alte limbi străine.

Ca efecte imediate ale utilizării acestui tip de definiție putem aminti faptul că definițiile devin sistematice, organizate, scurte și (mai) concise, clare și mai puțin criptice. Unele definiții ar putea fi structurate mai ușor în „gen proxim” și „diferența specifică”.

Obiecția care li s-ar putea aduce acestor definiții este posibila inexactitate: din cauza gradului ridicat de control, am putea crede că nu se pot defini cuvinte specializate din anumite domenii științifice, tehnice etc. Cu toate acestea, arătăm în exemplele de mai jos că acest lucru este posibil, cuvintele specializate apărând în acest tip de dicționare. Menționăm că toate traducerile definițiilor englezești ne aparțin.

Moleculă (en. *molecule*) este definit în *Macmillan Dictionary Online* ca „*the smallest part of an element or compound that is capable of independent existence. It consists of two or more atoms*” („cea mai mică parte a unui element chimic sau a unui compus care are existență de sine stătătoare. Este alcătuită din doi atomi”). Se observă că în această definiție apare un alt termen specializat, și anume *atom*, care este și el definit în dicționarul mai sus amintit astfel: „*1. SCIENCE the smallest unit of any substance. 2. [usually in negatives] a very small amount of something*” („1. Știință. Cea mai mică parte a unei substanțe. 2. [de obicei în sens depreciativ] cea mai mică parte din ceva”). Observăm din definiția de mai sus că folosirea unor definiții controlate în

explicarea termenilor științifici este posibilă și că definițiile nu își pierd din claritate, fiind de fapt pe înțelesul oricărui tip de vorbitor.

Bineînțeles că am încercat să vedem cum poate ajuta o definiție controlată la structurarea și organizarea unei definiții fie ea din DLR sau din alt dicționar. În acest sens ne vom folosi de exemple luate tot din limba engleză, din *Macmillan Dictionary Online* (M) și din *The Oxford English Dictionary*, varianta online (OED). Am ales aceste două dicționare pentru a putea realiza o comparație cu tipurile de definiții întâlnite în DEX și DLR, pentru că *Macmillan Dictionary* este un dicționar explicativ, iar OED-ul este tezaurul limbii engleze, echivalentul DLR-ului.

Ne vom opri la câmpul semantic al tipurilor de vehicule. Am ales să analizăm definițiile a trei cuvinte, en. *car*, en. *truck* și en. *bus*.

CAR: *A road vehicle for one driver and a few passengers* (M) („mașină – un vehicul rutier pentru un șofer și câțiva pasageri”.)

A road vehicle powered by a motor (usually an internal-combustion engine), designed to carry a driver and a small number of passengers, and usually having two front and two rear wheels, esp. for private, commercial, or leisure use; an automobile; (OED) („un vehicul rutier prevăzut cu un motor (de obicei un motor cu combustie internă), menit să poarte un șofer și câțiva pasageri și care are de obicei două roți în față și două în spate; este folosit mai ales de persoane fizice, în scop comercial sau pentru uz personal; un automobil)

TRUCK: *1. a large road vehicle used for carrying goods; 2. BE. British a railway vehicle used for carrying goods* (M) (camion „1. Un vehicul rutier, de dimensiuni mari, utilizat la transportul de mărfuri 2. (Engleză britanică) Un vehicul feroviar, utilizat la transportul de mărfuri”)

A wheeled vehicle for carrying heavy weights; variously applied. (OED) („Un vehicul cu roți folosit pentru transportul de mărfuri grele; se folosește în mai multe domenii”)

BUS: *A large road vehicle with a lot of seats that you pay to travel on, especially one that takes you fairly short distances and stops frequently* (M) (autobuz „un vehicul rutier, de dimensiuni mari, prevăzut cu multe scaune/locuri, pentru a cărui utilizare se plătește și care parcurge distanțe scurte și are opriri/stații frecvente”)

A large public vehicle carrying passengers by road, running on a fixed route and typically requiring the payment of a fare (OED); („un vehicul public de dimensiuni mari care transportă pasageri pe șosea, care parcurge o rută fixă și pentru a cărei utilizare trebuie plătit un tarif”).

Din exemplele de mai sus se poate observa că toate mijloacele de transport, fie ele personale sau publice, destinate transportului de călători sau transportului de mărfuri, sunt definite prin termenul supraordonat en. *[road] vehicle* („vehicul [rutier]”) ca gen proxim, restul definiției putându-se subsuma diferenței specifice, care vine să identifice tipul de vehicul. Deși cotate, definițiile de mai sus sunt concise și clare, fiind destul de explicative. Remarcăm că definițiile din Macmillan Dictionary sunt mult mai scurte și conțin doar elemente definitorii stricte, care vin să diferențieze un vehicul de altul, în timp ce definițiile din OED sunt mai complexe, uneori conținând informații parantetice de tipul en. *usually an internal-combustion engine* („de obicei un motor cu combustie internă”), informații enciclopedice. Această diferență este dată de tipul de dicționar, adică Macmillan Dictionary este doar un dicționar explicativ, care se adresează non-nativilor sau nativilor care doresc să înțeleagă un cuvânt, însă nu doresc să dobândească informații enciclopedice despre acel cuvânt și OED este un dicționar tezaur, care încearcă prin definițiile sale să ofere o istorie a cuvântului și informații cât mai complexe care să definească referențialul lexemului în cauză.

Mai mult, folosirea definițiilor cotate duce la stabilirea unor relații de hiper- și hiponimie între cuvinte. De exemplu, în Macmillan Dictionary autobuzul este definit *bus – a large road vehicle with a lot of seats that you pay to travel on, especially one that takes you fairly short distances and stops frequently (M) autobuz – un vehicul rutier, de dimensiuni mari, prevăzut cu multe scaune/locuri, pentru a cărui utilizare se plătește și care parcurge distanțe scurte și are opriri/stații frecvente; traducerea noastră*), în timp ce troleibuzul este definit ca en. *trolleybus – a bus that operates using electric power from wires fixed above the road; (troleibuz – un autobuz care se folosește pentru deplasare de cabluri electrice fixate deasupra șoselei)*.

2.2.1. Se poate utiliza o astfel de definiție în DLR?

Definițiile din DLR se apropie ca structură de cele în OED, ambele dicționare fiind de același fel. Astfel, ne-am întrebat dacă definițiile cotate ar putea fi folosite în DLR. Conform normelor de redactare ale DLR, definiția trebuie să conțină cuvinte cu sensul lor propriu, nu cu sensul lor figurat (Norme 127), nu trebuie să conțină cuvinte învechite, regional, iar definiția unui derivat trebuie să conțină mereu cuvântul-bază (Norme 133). Se poate deci observa cu ușurință că multe dintre principiile care stau la baza definițiilor cotate se verifică (cel puțin teoretic) și în cazul DLR-ului.

Cu toate acestea, există niște probleme care ne împiedică să folosim acest tip de definiție în DLR, cea mai importantă fiind lipsa unui vocabular specific definițiilor lexicografice. Crearea acestuia ar fi destul de dificilă ținând cont că nu există încă baze de date cu ajutorul cărora să stabilim frecvența cuvintelor

din limba română. O altă problemă ar fi o normă din *DLR* conform căreia trebuie preluate ca atare definițiile din glosare și comunicări (Norme, p. 135, 138, 141).

Plecând de la tiparul prezent în *Macmillan Dictionary* și *OED*, o posibilă soluție (în faza actuală) ar fi structurarea anumitor tipuri de definiții în așa fel încât să se ajungă la o organizare și sistematizare a unor leme care pot fi subsumate unui anumit câmp lexical (a unei familii lexicale, etc.) și care ar permite și o ierarhizare de tipul hiper/hiponimie arătat anterior. Ținând cont de diversitatea cuvintelor din *DLR*, de existența unui număr mare de arhaisme, regionalisme etc. este greu de crezut că o astfel de sistematizare ar putea fi aplicată la toate definițiile; mai mult, încarcerarea prea acută a redactorului ar putea duce la opacizarea definiției.

2.2.2. Se poate utiliza o astfel de definiție în DEX?

Am încercat să găsim răspuns și la întrebarea de mai sus. Am considerat că ținând cont de specificul acestui dicționar – un dicționar explicativ – folosirea definițiilor controlate ar fi binevenită pentru că unele definiții din DEX nu respectă întrutotul criteriile de organizare.

Totuși, pentru a putea folosi definițiile controlate în DEX ar fi nevoie de alcătuirea unui lexicon al definiției, proces care ar fi și el necesar, pentru că ar duce la sistematizarea informației în așa fel încât s-ar putea evita mai ușor definițiile circulare (acolo unde acest lucru ar fi un avantaj pentru utilizator). În plus, acest lexicon ar putea ajuta la organizarea materialului lexical în funcție de relațiile existente între cuvinte.

Plecând de la exemplul din engleză *bus-trolleybus*, am observat că troleibuzul nu este definit prin autobuz în DEX, ci doar este asemănător cu autobuzul, după cum se poate observa din definițiile de mai jos. Bineînțeles că orice vorbitor de română ar spune că troleibuzul este un autobuz cu troleu.

AUTOBÚZ, *autobuze*, s. n. Automobil cu caroseria închisă sau parțial decapotabilă, folosit la transportul în comun al unui număr mare de persoane.

TROLEIBÚZ, *troleibuze*, s. n. Vehicul rutier de transport în comun, cu tracțiune electrică, asemănător cu autobuzul, prevăzut cu troleu.

Bineînțeles că un nativ nu întâmpină prea multe probleme în înțelegerea definițiilor de mai sus, dar considerăm că definițiile de tipul celor discutate pentru limba engleză sunt mult mai clare și mai concise, dând acces la informație și utilizatorilor mai puțin experimentați.

3. Concluzii

Anul 2013 reprezintă un moment de răscruce în lexicografia românească pentru că informatizarea DLR deschide un drum nou în redactarea dicționarilor, dar și în lectura lor, și în cercetările lexico-semantice.

În articolul de față am încercat să aducem câteva argumente care să susțină ipoteza noastră conform căreia aplicarea unor constrângeri asupra definițiilor, prin metode automate, fie în sensul limitării cuvintelor folosite, fie pe baza unor relații semantice („gen proxim” și „diferență specifică”, hiper/hiponimie) poate prezenta o serie de avantaje. Dintre cele două metode discutate, prima ar fi mai degrabă aplicabilă DEX, în vreme ce a doua ar putea ajuta și la redactarea DLR. În primul rând, folosirea unei definiții organizate în subcâmpuri va duce la organizarea muncii lexicografului și sistematizarea produsului finit (toți redactorii vor fi „constrânși” să redacteze un anumit tip de definiție într-un anumit fel – zilele săptămânii, tipuri de vehicule, tipuri de flori etc.), ceea ce va face ca dicționarul să fie omogen, iar utilizatorii să îl folosească mai ușor. Într-o etapă ulterioară, definițiile bazate pe formalizarea relațiilor semantice în câmpuri și subcâmpuri ar putea permite atașarea limbii române în alte proiecte internaționale de procesare a limbajului natural.

BIBLIOGRAFIE

- Barque, Lucie, Alain Polguère, 2012, „Guide des annotateurs pour le balisage des définitions du TLFi. Projet Definiens”, http://www.atilf.fr/IMG/pdf/guide_anno_tlfi_2012.pdf
- Barque, Lucie, Alexis Nasr, Alain Polguère, 2010, „From the Definitions of the *Trésor de la Langue Française* to a Semantic Database of the French Language”, conferință prezentată la *The XIVth Euralex International Congress, Leeuwarden: Netherlands 2010*, <http://www.atilf.fr/IMG/pdf/BarqueNasrPolguere-2010.pdf>
- Definiens site - <http://www.atilf.fr/spip.php?article3780> (site-ul proiectului).
- Dicționarul explicativ al limbii române*, ediția a II-a, București, Univers Enciclopedic, 1998 (DEX).
- Dicționarul limbii române*. Seria nouă, București, Editura Academiei, 1965-2010.
- Geeraerts, Dirk, 2003, „Meaning and Definition”, în van Sterkenburg, P. G. J. (ed.), *A Practical Guide to Lexicography*, Amsterdam/Philadelphia, John Benjamins, p. 84-93.
- Norme DLR: *Normele tehnice de redactare*, copie dactilografiată.
- Queens, Frank, Uter Reker-Hamm, 2003, „A Net-based Toolkit for Collaborative Editing and Publishing of Dictionaries”. http://mhdwb.uni-trier.de/TARes/ACH_ALLC2003.pdf
- Renear, Allen, 2004, „Text Encoding”, în Susan Schreibman, Ray Siemens, John Unsworth (eds.), *A Companion to Digital Humanities*, Blackwell Publishing, p. 218-239.

- Richardson, Stephen D., William B. Dolan, Lucy Vanderwende, 1998, „MindNet: acquiring and structuring semantic information from text”, http://delivery.acm.org/10.1145/990000/980749/p1098-richardson.pdf?ip=88.25.252.234&id=980749&acc=OPEN&key=BF13D071DEA4D3F3B0AA4BA89B4BCA5B&CFID=247737684&CFTOKEN=47894148&__acm__=1379797496_509f93460aef2abda1ffb852e6f93cdf
- Rundell, Michael, 2012, „The road to automatic lexicography: An editor’s viewpoint”, în Sylviane Granger, Maguali Paquot, *Electronic Lexicography*, Oxford University Press, p. 15-30.
- Trésor de la langue française informatisé* (TLFi), CNRS éditions, Paris, 2004.
Webster site: www.merriam-webster.com
- Willy Martin, 2004, „SIMuLLDA, the Hub-and-Spoke Model and Frames or How to Make the Best of Three Worlds?”, în *International Journal of Lexicography*, vol. 17, no. 2, p. 175-187.

Monica VASILEANU, Anabella-Gloria NICULESCU-GORPIN
Institutul de Lingvistică al Academiei Române
„Iorgu Iordan – Al. Rosetti”, București