REDACTAREA ÎN FORMAT ELECTRONIC A DLR. CÂMPUL „DEFINIŢIE” – POSIBILITĂŢI 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, definiţie, informatizarea, dicţionar.
Key-words: e-lexicography, definition, digitalisation, dictionary.

Dezvoltarea fără precedent a ştiinţei şi a tehnologiei a dus la apariţia unor noi subdomenii ale lingvisticii, cum ar fi neurolingvistica, dar şi la noi direcţii de cercetare în cadrul disciplinelor lingvistice deja consacrate. Nici lexicografia nu a scăpat de „virusul informatizării”: de la dicţionare explicative unilingve şi/sau bilingve la dicţionare tezaur, toate tind să aibă cel puţin o variantă electronică, astfel încât acest tip de redactare a dicţionarelor este o practică deja verificată de tradiţia lexicografică occidentală.

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

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

Redactarea în format XML presupune un efort suplimentar de organizare a informaţiei, pentru ca varianta electronică să permită utilizatorului mai multe tipuri de căutare şi, deci, mai multe modalităţi de cercetare. Acest efort de organizare a informaţiei îi este util şi redactorului din motivele arătate mai jos, sintetizate într-o prezentare succintă a procesului de redactare în această noua 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 priveşte modul de organizare a câmpului „definiţie" în acest nou format.

Plecând de la tipurile de definiţii existente în forma tradiţională a DLR, am sugerat moduri în care utilizarea programelor lexicografice ar putea uşura munca de redactare, eliminând, pe cât posibil, eroarea umană. Am arătat cum restricţiile pe care aceste programe le impun pot fi folosite drept avantaj şi nu ca elemente care să îngreuneze redactarea. Am avut ca exemple modul în care alte dicţionare în format electronic au ales să organizeze câmpul „definiţie" şi, păstrând specificul DLR, am testat dacă este posibilă şi o sistematizare (şi ridigizare) a tipului de definiţii întâlnite în acest dicţionar şi, apoi, dacă vreuna dintre soluţiile oferite de lexicografii englezi sau francezi poate fi profitabilă pentru a doua lucrare importantă a departamentului de Lexicografie, anume *Dicţionarul explicativ al limbii române* (DEX).

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

Informatica s-a insinuat în redactarea dicţionarelor încă din anii 1960, prin informatizarea dicţionarelor Webster (site-ul Webster); în acelaşi timp a început şi 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, strămoşul actualului FRANTEXT, a permis extragerea automată a 430 000 de citate (TLFi 2004: 2). Dimensiunea bazei de date obţinute prin despuierea 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 construcţii sintactice.

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

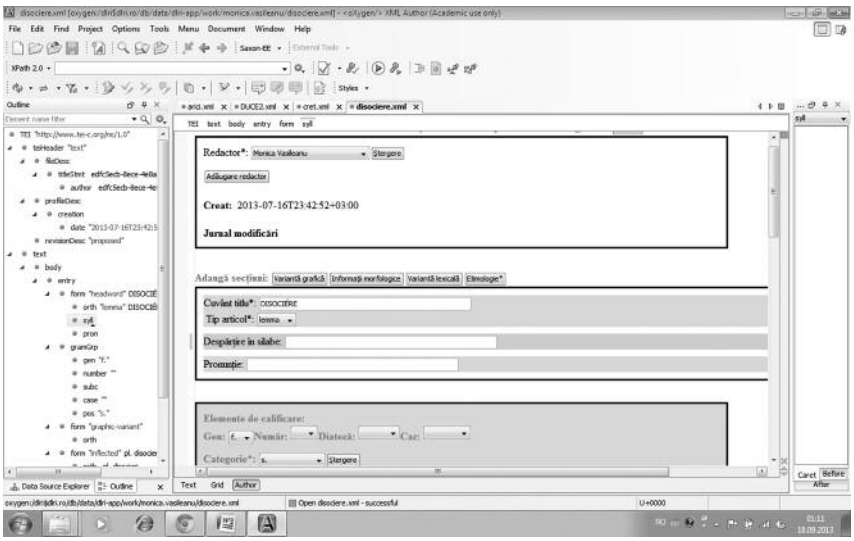
majuscule etc.) fiecărui tip de informaţie ar trebui aplicate manual, ci transformarea lui într-un document structurat (TLFi 2004: 10; Queens, Reker-Hamm 2003: 1). Pentru a transforma un document electronic simplu într-un document structurat este nevoie: a) de o segmentare a documentului, b) de adnotarea 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 dicţionar sunt purtătoare de semnificaţii (ex. cuvântul-titlu se marchează în DLR prin majuscule aldine, prin urmare cititorul ştie că acele cuvinte scrise cu majuscule aldine sunt cuvinte-titlu, aşadar în funcţie de aceste stiluri se cataloghează informaţia). Pentru ca aceste adnotări să poată fi procesate de mai multe programe, a fost necesară standardizarea lor şi deci crearea unei liste de asemenea adnotări. Întrucât aceste adnotări transformă documentul într-un fel de hartă, marcând şi delimitând tipurile de informaţie, au mai fost numite şi balize între lexicografi.

Standardizarea lor s-a petrecut în anii 1980, prin recunoaşterea de către ISO (International Organization for Standrdization) a SGML (=Standard General Markup Language) drept un standard în tehnologia informatică (Renear 2004: 226). Din această tehnologie standard s-au dezvoltat limbajele HTML şi XML, utilizate pe scară largă astăzi mai ales în crearea paginilor de internet. Limbajul XML şi dezvoltat de TEI (Text Encoding Initiative), un consorţiu care menţine şi 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 şi în lexicografie. Comunitatea TEI organizează conferinţe şi ateliere, punând la dispoziţia utilizatorilor şi programele care pot procesa balizele-standard acceptate de TEI (www.tei-c.org). O prezentare a acestor balize şi a ierarhiei lor există pe site-ul TEI, sub forma unui ghid foarte cuprinzător şi... stufos.

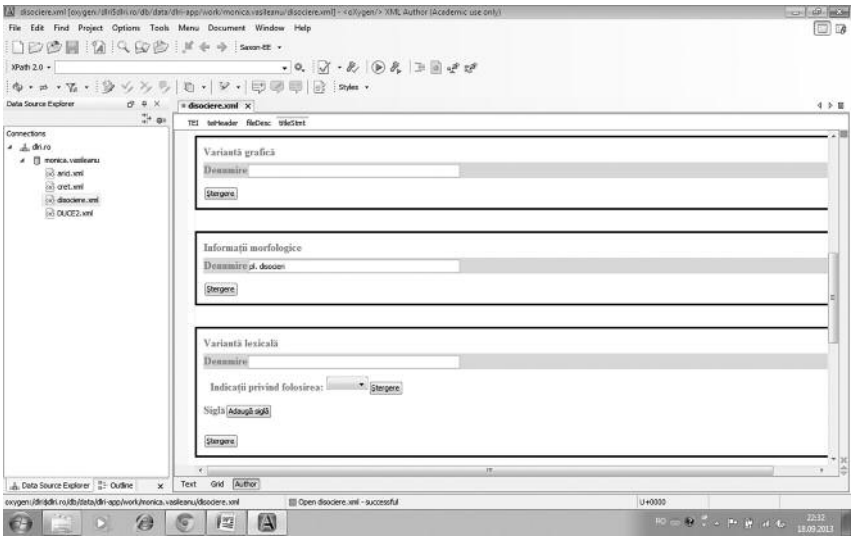
În informatizarea DLR se foloseşte limbajul XML, procesat prin programul Oxygen. Astfel, balizele folosite pentru transformarea articolelor de dicţionar în documente structurate sunt cele din ghidul TEI. Întrucât DLR este un dicţionar complex şi conţine multe tipuri de informaţie, informaticianul Claudius Teodorescu a pregătit o interfaţă care să ascundă etichetele şi să faciliteze astfel munca redactorilor. Prezentăm mai jos câteva capturi de ecran, aşa cum arătau ele în luna noiembrie 2013. De atunci, interfaţa a suferit şi va mai suferi modificări datorate specificului *DLR-ului*.

(1) Există posibilitatea de a vedea şi partea tehnică, în formt pur XML, dar lexicografului, lingvist de profesie şi nu informatician, îi este mai util să lucrez în interfaţa „Author”. Pentru a avea o evidenţă riguroasă a muncii fiecărui redactor, primele câmpuri conţin metadatele: Se înregistrează

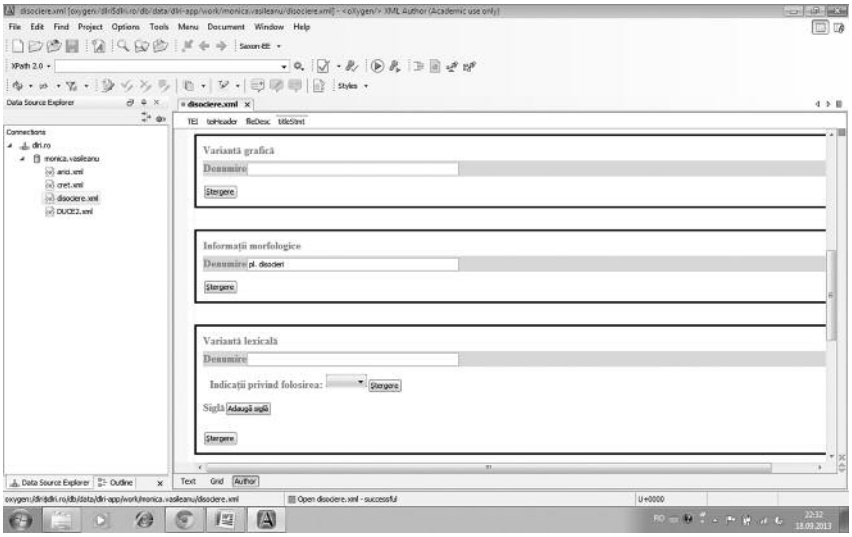
numele redactorului (într-o fază ulterioară al revizorului), data creării, data modificării etc.



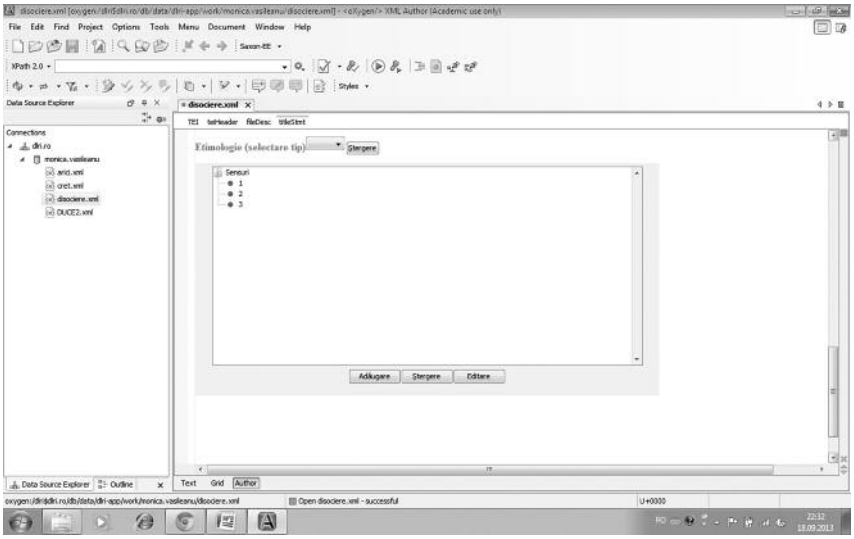
(2) Urmează câmpurile specifice dicţionarului propriu-zis: cuvântul-titlu, tipul articolului (lemă sau variantă); astfel, variantele ar putea fi inserate automat la ordinea alfabetică, iar corelarea cu cuvântul-titlu s-ar face şi ea automat. Sunt prezente şi câteva elemente de ortoepie, cum ar fi despărţirea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



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

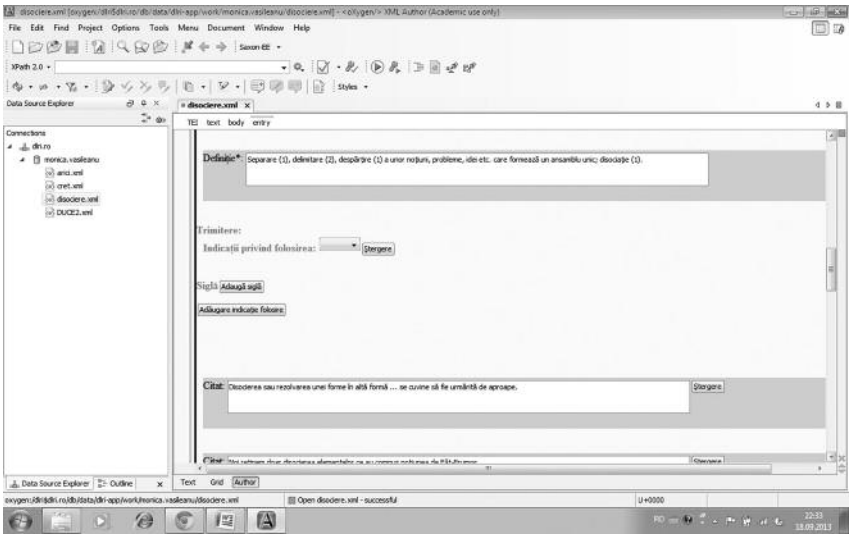


(4) Informaţia semantică are aspectul unui arbore de sensuri. Butoanele permit adăugarea de subsensuri şi subsubsensuri – numărul de trepte este probabil infinit, însă în DLR există maximum 5 trepte.



(5) Definiţia propriu-zisă este introdusă într-un câmp unic, nesegmentat. Fiecărei definiţii îi urmează o trimitere (opţională) către alt cuvânt-intrare înrudit, indicaţii privind folosirea (mărci diastratice, indicaţii de domeniu, restricţii combinatorice etc.) – în forma tipărită, acestea se menţionează înaintea definiţiei) şi citatele ilustrative. Siglele citatelor vor fi introduse automat, prin selectare dintr-un meniu derulant.

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



2. Câmpul definiţie: posibilităţi de structurare

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

Am văzut, în capturile de ecran de mai sus, că definiţia este redactată într-un câmp unic, nesegmentat, nerestricţionat. Ne-am gândit la două posibilităţi de structurare a acestui câmp, posibilităţi exploatate de alte dicţionare europene sau aflate încă în stadiul de cercetare, cu aplicaţii multiple:

a. Segmentarea câmpului „definiţie" în mai multe subcâmpuri, precum „gen proxim" + „diferenţe specifice". Această idee presupune segmentarea

şi balizarea informaţiei semantice, operaţiune dificilă pentru redactor, dar care deschide noi perspective de cercetare;

b. Restricţionarea câmpului „definiţie" prin stabilirea unui număr finit de cuvinte care să poată fi folosite în definiţii. Lexicograful presupune că aceste cuvinte folosite în „definiţia controlată" sunt cunoscute deja de utilizator. Această practică este specifică dicţionarelorcu scop didactic – *learners' dictionaries* (Geeraerts 2003: 91).

2.1 Segmentarea câmpului definiţie

Întrucât dicţionarele de referinţă din lexicografia europeană au apărut mai întâi în format tipărit şi abia apoi în format electronic, definiţiile au fost redactate fără ajutorul uneltelor electronice. Redactorii au aplicat „şabloane de definire", redactând definiţii-tip de fiecare dată când a fost posibil, însă acesta a fost un proces efectuat de oameni în mod tradiţional, nu automatizat. Asemenea definiţii-tip au fost aplicate numai unui număr finit de cuvinte. În ce măsură se poate extinde această practică şi 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 şi Definiens – două modele

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

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

Lexicografia franceză (şi nu numai) parcurge un drum invers nu de redactare a definiţiei după un şablon, ci de extragere a şablonului din definiţii deja redactate. Un grup de cercetători ai celebrului Institut de linguistique française se ocupă, în cadrul programului Definiens, cu segmentarea şi adnotarea definiţiilor din *Trésor de la langue française informatisé*, fără a rescrie sau a revizui aceste definiţii (cel puţin pentru moment), cu scopul de a forma o bază de date cu definiţii structurate, de mare importanţă pentru toate cercetările lexicale şi 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 şi 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, Pologùere 2010: 5).

Ghidul de adnotare (Barque, Pologùere 2012) cuprinde instrucţiuni pentru lingviştii implicaţi în proiect şi o listă a balizelor folosite. Segmentarea definiţiilor se petrece doar în cazul definiţiilor analitice, definiţiile prin sinonime sau prin serie sinonimică fiind balizate ca atare:

```
FLAGADA =
  <DEFI><PARAPH_SYNOS>Fatigué, avachi, sans force</PARAPH_SYNOS>.</DEFI>
```

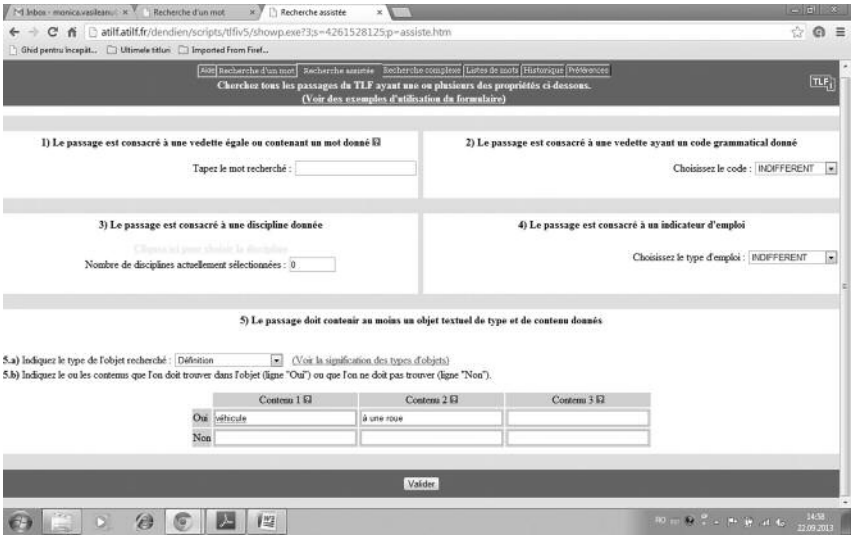
(Barque, Pologùere 2012: 5)

Segmentarea definiţiilor analitice are mai multe etape. Într-o primă etapă, lingvistul trebuie să detaşeze „componenta centrală" de „componentele periferice", adică „genul proxim" de „diferenţele 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, Pologùere 2010: 3)

Alegerea componentei centrale se dovedeşte o acţiune dificilă întrucât presupune o bună cunoaştere a dicţionarului. Componenta centrală nu se stabileşte doar pe criterii sintactice (să fie centrul de grup al parafrazei-definiţie), ci prin comparaţie cu alte cuvinte din acelaşi câmp lexico-semantic. Spre exemplu, în cazul cuvântului *brouette*, lingvistul oscilează între „véhicule" şi „véhicule à une roue". Dacă „véhicule à une roue" ar mai fi găsită ca centru al altei definiţii, atunci întreaga sintagmă ar fi acceptată ca gen proxim. Verificarea se poate face prin căutarea sintagmei în câmpul definiţiilor din TLF:



Întrucât singurul articol găsit în urma acestei căutări a fost *brouette*, genul proxim indicat rămâne „véhicule". Însă şi această căutare poate fi înşelătoare, întrucât adeseori acelaşi conţinut 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 aceeaşi componentă centrală, exprimtă prin aceleaşi cuvinte, poate fi discontinuă într-un caz şi continuă în altul:

RÉGLETTE =
<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 charactéristiques**>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 étiq=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), obţinute în urma segmentării unui eşantion mare de definiţii; cu siguranţă, lista nu este exhaustivă şi până la balizarea tuturor definiţiilor analitice din TLFi vor mai fi adăugate şi altele.

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

Prin această balizare se obţine o bază de date cu definiţii 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 ţări cu o tradiţie lexicografică mai îndelungată, cu bugete mai mari alocate cercetării lingvistice şi cu o mai mare experienţă în procesarea automată a limbii respective. În ce măsură se pot aplica aceste idei în lexicografia românească? Vom discuta cele două dicţionare principale elaborate de departamentul de Lexicologie şi lexicografie, DLR şi DEX. Începem cu DLR, deoarece redactarea în format XML a acestui dicţionar este acum tema principală a departamentului de Lexicologie şi lexicografie. Complexitatea ridicată a DLR face ca soluţiile găsite pentru automatizarea acestui dicţionar să se aplice fără probleme şi la DEX.

DLR este dicţionarul-tezaur al limbii române, aşadar este cea mai cuprinzătoare operă lexicografică românească, incluzând nu doar cuvintele curente, ale limbii literare contemporane, ci şi cuvinte ieşite din uz, regionale, populare, cuvinte efemere, creaţii ale unor scriitori, termeni meşteşugăreşti, termeni tehnico-ştiinţifici (cu anumite restricţii) şi chiar cuvinte cu sens neprecizat. Limitele largi ale dicţionarului înseamnă şi o complexitate semantică greu de formalizat.

Normele de redactare a DLR enumeră trei tipuri principale de definiţii. Vom vedea, acestea se aplică numai cuvintelor pline semantic şi cu anumite restricţii:

– definiţia analitică;
– definiţia prin serie sinonimică;
– definiţia prin sinonim unic.

Definiţia analitică este procedeul de bază al descrierii semantice (Norme 121). A fost concepută încă de la început ca o definiţie structurată, în funcţie de mai multe criterii. Astfel, fiece parte de vorbire urmează un anumit tipar sintactic al definiţiei. Definiţia analitică a verbelor începe cu „a", marca infinitivului, cea a adjectivelor, cu „care".

Întrucât numai definiţia analitică poate fi segmentată, acesta va fi singurul tip de definiţie discutat în prezenta secţiune. Excludem, aşadar, definiţiile prin serie sinonimică; definiţiile prin sinonim unic, aplicate în special cuvintelor neliterare (Norme 155), pot fi incluse în discuţie într-o fază ulterioară, preluând definiţia analitică a sinonimului literar.

Normele DLR recomandă structurarea definiţiei în gen proxim şi diferenţe specifice (Norme 127-129). Motivul este de natură ontologică: este considerată definiţia 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 nuanţe sensibile într-un sens nu vor putea fi subordonate unui gen proxim" (Norme 129).

Anumite serii de cuvinte, în majoritate substantive şi verbe, primesc definiţii-tip: zilele săptămânii, numele lunilor, gradele de rudenie, gradele militare, numele de plante şi animale, majoritatea abstractelor etc. (Fascicula 0: XI). Cu toate acestea, diferenţa mare de timp între momentele redactării definiţiilor unor cuvinte aflate în aceeaşi serie face ca şablonul să sufere mici modificări. Exemplificăm întâi cu zilele săptămânii, o serie finită şi uşor de uniformizat:

LUNI: Prima zi a săptămânii, care urmează după duminică.
MIERCURI: Ziua a treia a săptămânii, care urmează după marţi.
MARŢI: Ziua a doua a săptămânii, care urmează după luni.
SÂMBĂTĂ: Ziua a şasea a săptămânii, care urmează după vineri. V. sabat[1] (2).
DUMINICĂ: Ultima (sau prima) zi a săptămânii, considerată zi de odihnă, în civilizaţiile creştine.

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

Cu aceste observaţii ajungem la primul avantaj al existenţei unui câmp-definiţie segmentat: posibilitatea observării neregularităţilor existente în vederea revizuirii, proces mult mai simplu şi mult mai puţin costisitor în era electronică.

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

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

– dacă selectează o definiţie analitică, redactorul va găsi două câmpuri, de gen proxim şi diferenţe specifice;

– în cazul cuvintelor din aceeaşi serie, cum a fost cazul zilelor săptămânii, indicarea genului proxim să acţioneze ca o constrângere asupra câmpului rămas liber pentru diferenţele specifice. Spre exemplu, odată ce am introdus genul proxim „zi a săptămânii", câmpul diferenţelor specifice să se segmenteze automat în mai multe subcâmpuri: unul, etichetat „ordinea", să fie completat cu numeralul ordinal („prima", „a doua", „a şasea" etc.), al doilea, etichetat „precedent", să conţină formularea „care urmează după ...", redactorul urmând să introducă manual numele zilei precedente, şi un al treilea câmp, lăsat liber, pentru a putea adăuga alte note semantice specifice unui singur membru al seriei – cazul substantivului „duminică", în definiţia căruia găsim „considerată zi de odihnă în civilizaţiile creştine";

– asemenea şabloane am putea avea în cazul plantelor: odată introdus genul proxim „plantă", câmpul diferenţelor să se segmenteze automat în „familie", „detalii descriptive", „utilitate" şi un câmp liber, lăsat pentru detaliile specifice doar unuia dintre membrii seriei.

Cu siguranţă, este nevoie de o segmentare a câmpului definiţie în cazul definiţiilor 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 afirmaţii etc.; (concretizat) semn, mărturie, argument etc. în sprijinul (sau împotriva) cuiva sau a ceva ori pentru confirmarea unui fapt, a unei afirmaţii, a unei presupuneri etc.

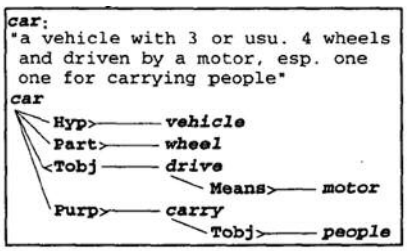
Avantajul actualei echipe de redactori ai DLR este că pot reformula definiţiile. Segmentarea câmpului, în loc de constrângere, poate deveni un ajutor în uniformizarea formulărilor. Existenţa acestor subcâmpuri structurează datele pe care redactorul trebuie să le introducă în definiţie şi le ierarhizează.

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

2.1.3 Utilitatea segmentării definiţiilor – direcţii 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 şi, mai ales, prin etichetarea lor. Am văzut deja un proces anevoios întreprins de o mică echipă de lingvişti francezi asupra definiţiilor din TLFi, proces care urmăreşte a) segmentarea definiţiilor şi b) etichetarea componentelor semantice, pentru a face un inventar al componentelor semantice.

Acelaşi proces de segmentare şi etichetare, ba chiar cu câţiva paşi în plus, a fost efectuat în câteva interesante proiecte recente asupra limbii engleze, prin folosirea unor programe speciale, numite *parsers*. Încă din anii 1990, definiţii din diverse dicţionare (preferate fiind *learners' dictionaries*, vom vedea mai jos de ce) au fost segmentate cu scopul de a găsi nişte relaţii care să configureze un fel de „gramatică" a definiţiilor. Pe baza acestei gramatici s-ar putea stabili relaţiile semantice existente între toate cuvintele unei limbi. Proiecte internaţionale precum MindNet sau SIMuLLDA au efectuat segmentarea definiţiilor în lanţuri de câte trei termeni (două cuvinte şi indicarea relaţiei). „Gramatica" semantică ar trebui să conţină un număr finit de relaţii. Iată mai jos un exemplu de definiţie segmentată prin MindNet şi tabelul relaţiilor:



(Richardson, Dolan, Vanderwende, 1998: 1099)

Attribute	Goal	Possessor
Cause	Hypernym	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 obţină o reţea semantică pornind de la dicţionare şi enciclopedii. Engleza nu este singura limbă vizată, iar aplicarea acestei „gramatici semantice" asupra altor limbi ar

aduce o eficienţă sporită traducerilor automate şi tuturor programelor de procesare a limbajelor naturale.

SIMuLLDA a început ca o cercetare a unui consorţiu al Comisiei Europene (Martin 2004: 176), în care s-a propus modelul „hub and spoke" („trunchi şi ramuri"). În mod previzibil, scopul proiectului a fost producerea de dicţionare 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 exponenţial – creşte şi numărul de dicţionare bilingve care pot fi elaborate automat. Pentru a putea stabili relaţii de echivalenţă 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, relaţiile dintre limbile-ramuri pot fi stabilite prin calcule, pe baza relaţiei fiecăreia dintre ramuri cu limba-trunchi (Martin 2004: 177). Avantajul este că pot fi asociate, într-un dicţionar bilingv, limbi care n-au fost niciodată studiate constrastiv. Din asocierea a cinci lexicoane rezultă zece posibile dicţionare; prin modelul hub and spoke, numai patru trebuie elaborate de oameni (cele care leagă limba-trunchi de fiecare dintre cele patru ramuri), celelalte şase putând fi generate automat pe baza calculelor (Martin 2004: 179).

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

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

Toate operaţiunile descrise se bazează pe lexicoane monolingve obţinute prin segmentarea şi adnotarea definiţiilor din dicţionarele de referinţă ale fiecărei limbi (Martin 2004: 186). Definiţiile sunt văzute de cercetătorii din proiectul SIMuLLDA ca nişte ansambluri de lanţuri de atribute, asemănător cu lanţurile propuse de MindNet.

Cert este că, pentru ca proiecte similare să poată include şi limba română, definiţia 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 definiţiei, cu evidente „spaţii de rezervă".

2.2 Restricţionarea definiţiilor – „definiţii controlate”

Definiţiile controlate sunt acele definiţii care folosesc un număr limitat de cuvinte, considerate a fi cunoscute în cea mai mare măsură utilizatorilor. Procedeul este foarte des întâlnite în lexicografia britanică, fiind folosit pentru prima dată în *New Method English Dictionary* de Michael West and James Endicott, în anul 1935. La redactarea definiţiilor din acest dicţionar 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 conţinea cele mai frecvente 2000 de cuvinte din limba engleză. Alte dicţionare explicative monolingve (dedicate atât utilizatorilor nativi, cât mai ales celor care învaţă limba engleză sunt: *Macmillan English Dictionary for Advanced Learners*, care utilizează la redactarea definiţiilor aproximativ 2500 de cuvinte, *Oxford Advanced Learner's Dictionary*, cu aproximativ 3000 cuvinte şi *Cambridge International Dictionary of English*, care foloseşte în definiţii chiar mai puţin de 2000 de cuvinte.

Un prim avantaj este acela că lexicografii folosesc la redactarea definiţiilor cuvinte cu o frecvenţă mare, pe care majoritatea utilizatorilor le cunosc, fapt care face ca dicţionarele să fie mai uşor de utilizat atât de către vorbitorii nativi, cât şi de cei nenativi. De aceea, în titlul acestor lexicoane se găseşte adeseori clar menţionat scopul lor didactic în sintagma *learner's dictionary*. Controlarea definiţiilor 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 evitată cuvintele care ar putea fi confundate cu alte cuvinte din limba-ţintă sau din alte limbi străine.

Ca efecte imediate ale utilizării acestui tip de definiţie putem aminti faptul că definiţiile devin sistematice, organizate, scurte şi (mai) concise, clare şi mai puţin criptice. Unele definiţii ar putea fi structurate mai uşor în „gen proxim” şi „diferenţa specifică”.

Obiecţia care li s-ar putea aduce acestor definiţii este posibila inexactitate: din cauza gradului ridicat de control, am putea crede că nu se pot defini cuvinte specilizate din anumite domenii ştiinţifice, tehnice etc. Cu toate acestea, arătăm în exemplele de mai jos că acest lucru este posibil, cuvintele specializate apărând în acest tip de dicţionare. Menţionăm că toate traducerile definiţiilor englezeşti ne aparţin.

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 existenţă de sine stătătoare. Este alcătuită din doi atomi”). Se observă că în această definiţie apare un alt termen specializat, şi anume *atom*, care este şi el definit în dicţionarul mai sus amintit astfel: „*1. SCIENCE the smallest unit of any substance. 2. [usually in negatives] a very small amount of something”* („1. Ştiinţă. Cea mai mică parte a unei substanţe. 2. [de obicei în sens depreciativ] cea mai mică parte din ceva”). Observăm din definiţia de mai sus că folosirea unor definiţii controlate în

explicarea termenilor ştiinţifici este posibilă şi că definiţiile nu îşi pierd din claritate, fiind de fapt pe înţelesul oricărui tip de vorbitor.

Bineînţeles că am încercat să vedem cum poate ajuta o definiţie controlată la structurarea şi organizarea unei definiţii fie ea din DLR sau din alt dicţionar. În acest sens ne vom folosi de exemple luate tot din limba engleză, din *Macmillan Dictionary Online* (M) şi din *The Oxford English Dictionary*, varianta online (OED). Am ales aceste două dicţionare pentru a putea realiza o comparaţie cu tipurile de definţii întâlnite în DEX şi DLR, pentru că Macmillan Dictionary este un dicţionar 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 definiţiile a trei cuvinte, en. *car*, en. *truck* şi en. *bus*.

CAR: *A road vehicle for one driver and a few passengers* (M) („maşină – un vehicul rutier pentru un şofer şi câţiva 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 şofer şi câţiva pasageri şi care are de obicei două roţi în faţă şi 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 roţi folosit pentru trasportul de mărfuri grele; se foloseşte 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 vechicul rutier, de dimensiuni mari, prevăzut cu multe scaune/locuri, pentru a cărui utilizare se plăteşte şi care parcurge distanţe scurte şi are opriri/staţii frecvente”)

A large public vehicle carrying passengers by road, running on a fixed route and typically requiring the payment of a fare (OED); („un vechicul public de dimensiuni mari care transportă pasageri pe şosea, care parcurte o rută fixă şi 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 definiţiei putându-se subsuma diferenţei specifice, care vine să identifice tipul de vehicul. Deşi cotrolate, definiţiile de mai sus sunt concise şi clare, fiind destul de explicative. Remarcăm că definiţiile din Macmillan Dictionary sunt mult mai scurte şi conţin doar elemente definitorii stricte, care vin să diferenţieze un vehicul de altul, în timp ce definiţiile din OED sunt mai complexe, uneori conţinând informaţii parantetice de tipul en. *usually an internal-combustion engine* („de obicei un motor cu combustie internă"), informaţii enciclopedice. Acestă diferenţă este dată de tipul de dicţionar, adică Macmillan Dictionary este doar un dicţionar explicativ, care se adresează non-nativilor sau nativilor care doresc să înţeleagă un cuvânt, însă nu doresc să dobândească informaţii enciclopedice despre acel cuvânt şi OED este un dicţionar tezaur, care încearcă prin definiţiile sale să ofere o istorie a cuvântului şi informaţii cât mai complexe care să definească referentul lexemului în cauză.

Mai mult, folosirea definiţiilor controlate duce la stabilirea unor relaţii de hiper- şi 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 vechicul rutier, de dimensiuni mari, prevăzut cu multe scaune/locuri, pentru a cărui utilizare se plăteşte şi care parcurge distanţe scurte şi are opriri/staţii 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 foloseşte pentru deplasare de cabluri electrice fixate deasupra şoselei)*.

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

Definiţiile din *DLR* se apropie ca structură de cele in *OED*, ambele dicţionare fiind de acelaşi fel. Astfel, ne-am întrebat dacă definiţiile controlate ar putea fi folosite în *DLR.* Conform normelor de redactare ale DLR, definiţia trebuie să conţină cuvinte cu sensul lor propriu, nu cu sensul lor figurat (Norme 127), nu trebuie să conţină cuvinte învechite, regional, iar definiţia unui derivat trebuie să conţină mereu cuvântul-bază (Norme 133). Se poate deci observa cu uşurinţă că multe dintre principiile care stau la baza definiţiilor controlate se verifică (cel puţin teoretic) şi în cazul *DLR-ului*.

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

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

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

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

Am încercat să găsim răspuns şi la întrebarea de mai sus. Am considerat că ţinând cont de specificul acestui dicţionar – un dicţionar explicativ – folosirea definiţiilor controlate ar fi binevenită pentru că unele definiţii din DEX nu respectă întrutotul criteriile de organizare.

Totuşi, pentru a putea folosi definiţiile controlate în DEX ar fi nevoie de alcătuirea unui lexicon al definiţiei, proces care ar fi şi el necesar, pentru că ar duce la sistematizarea informaţiei în aşa fel încât s-ar putea evita mai uşor definiţiile circulare (acolo unde acest lucru ar fi un avantaj pentru utilizator). În plus, acest lexicon ar putea ajuta la organizarea materialului lexical în funcţie de relaţiile 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 definiţiile de mai jos. Bineînţeles că orice vorbitor de română ar spune că troleibuzul este un autobuz cu troleu.

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

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

Bineînţeles că un nativ nu întâmpină prea multe probleme în înţelegerea definiţiilor de mai sus, dar considerăm că definţiile de tipul celor discutate pentru limba engleză sunt mult mai clare şi mai concise, dând acces la informaţie şi utilizatorilor mai puţin experimentaţi.

3. Concluzii

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

În articolul de faţă am încercat să aducem câteva argumente care să susţină ipoteza noastră conform căreia aplicarea unor constrângeri asupra definiţiilor, prin metode automate, fie în sensul limitării cuvintelor folosite, fie pe baza unor relaţii semantice („gen proxim" şi „diferenţă 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 şi la redactarea DLR. În primul rând, folosirea unei definiţii organizate în subcâmpuri va duce la organizarea muncii lexicografului şi sistematizarea produsului finit (toţi redactorii vor fi „constrânşi" să redacteze un anumit tip de definiţie într-un anumit fel – zilele săptămânii, tipuri de vehicule, tipuri de flori etc.), ceea ce va face ca dicţionarul să fie omogen, iar utilizatorii să îl folosească mai uşor. Într-o etapă ulterioară, definiţiile bazate pe formalizarea relaţiilor semantice în câmpuri şi subcâmpuri ar putea permite ataşarea limbii române în alte proiecte internaţionale 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", conferinţă 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).

Dicţionarul explicativ al limbii române, ediţia a II-a, Bucureşti, Univers Enciclopedic, 1998 (DEX).

Dicţionarul limbii române. Seria nouă, Bucureşti, 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", Bucureşti