Sistemul all-inclusive în reprezentarea cunoştinţelor lexicale

Verginica BARBU MITITELU[*]

Key-words: *semantic relations*, *semantic network*, *wordnet*

1. Introducere

În ultima jumătate de secol, lingvistica a fost dominată de viziunea conform căreia lexicul formează centrul preocupărilor. Teoriile gramaticale dezvoltate în a doua jumătate a secolului trecut (Kaplan şi Bresnan, *Lexical Functional Grammar*, 1982, Gazdar, Klein, Pullum, Sag, *Generalized Phrase Structure Grammar*, 1985, Pollard şi Sag, *Head-driven Phrase Structure Grammar*, 1987 şi 1994) accentuează rolul lexiconului în studiul limbii. Concepute ca teorii alternative la gramatica transformaţională (dezvoltată în tradiţie chomskiană), aceste teorii (în special *Head-driven Phrase Structure Grammar*) dau seama de numeroase fenomene lingvistice la nivelul *lexiconului*, eliminând astfel transformările de la nivelul sintactic. Această evoluţie în lingvistica teoretică a fost dublată de preocupările din lingvistica computaţională de a înţelege sintaxa unei limbi, lucru imposibil fără cunoştinţe de semantică şi vocabular. Concentrându-se iniţial în special asupra traducerii automate, ulterior şi (sau mai ales) asupra dezvoltării unor aplicaţii utile societăţii per ansamblu sau diverselor domenii de lucru, inginerii au căutat cel mai convenabil mod de reprezentare a informaţiilor lexicale din perspectiva prelucrării automate a acestora.

Suntem într-o perioadă în care cercetările interdisciplinare conduc la rezultate ce ajută dezvoltarea disciplinelor individuale, în care deschiderea cercetătorilor spre explorarea din multiple perspective este crucială. Studiul lexicului interesează azi deopotrivă lingvişti, psiholingvişti, sociolingvişti şi inginerii în prelucrarea şi în generarea limbajului natural, sub formă scrisă sau orală.

Cunoştinţele lexicale reprezintă totalitatea informaţiilor despre cuvinte. Aceste informaţii sunt atât lingvistice (fonetice, morfologice, sintactice), cât şi extralingvistice (pragmatice, ontologice, conceptuale). Ele sunt deţinute de om, într-un mod mai mult sau mai puţin conştient. Sunt îmbogăţite permanent, actualizate, atunci când este necesar, şi accesate, de obicei, cu uşurinţă, dovadă a caracterului lor organizat. Un alt aspect important în legătură cu accesul la acestea este modalitatea bidirecţională în care se poate face: şi de la cuvânt la informaţiile despre el, şi de la aceste informaţii la cuvântul pentru care sunt valabile: astfel, vorbitorul este capabil atât să găsească cuvântul potrivit pentru transmiterea unui sens, cât şi să identifice sensul corect al unui cuvânt.

Comutarea interesului asupra lexicului a implicat şi conferirea unui statut privilegiat semanticii, atât celei lexicale, cât şi celei sintactice. Lucrarea de faţă se

[*] Institutul de Cercetări pentru Inteligenţă Artificială „Mihai Drăgănescu”, Academia Română, București, România.

circumscrie semanticii lexicale sincronice, care se ocupă cu studiul semnificaţiei cuvintelor, considerate independent de context (într-o măsură mai mică sau mai mare).

Trei abordări în studiul sensului cuvintelor se remarcă:

- abordarea bazată pe câmpuri semantice: fiecare cuvânt este definit de locul pe care îl ocupă în cadrul câmpului; dovedindu-şi calităţile în lingvistica descriptivă, această abordare nu poate fi totuşi folosită pentru crearea lexicoanelor în vederea adnotării morfo-sintactice a corpusurilor şi nici pentru generarea limbajului;
- analiza componenţială: fiecare cuvânt este definit printr-o mulţime de seme care îl încadrează alături de termenii dintr-un anumit domeniu, precum şi prin alte seme care îi conferă specificitate; această metodă este eficientă în descrierea cuvintelor dintr-o anumită sferă semantică, dar limitările devin evidente când se lucrează cu un număr mai mare de cuvinte din domenii complexe, în care sensurile se suprapun;
- abordarea relaţională: definirea sensurilor unui cuvânt se face prin specificarea relaţiilor pe care le contractează în cadrul domeniului (cf. Evens 1988: 2). Cea mai mare parte a vocabularului se pretează la o descriere de tip relaţional, dar există şi arii care nu pot fi organizate astfel, după cum se va vedea mai jos.

Psiholingvistica, i.e. studiul factorilor psihologici şi neurobiologici care îi ajută pe oameni să achiziţioneze, să înveţe, să folosească, să înţeleagă şi să producă limbajul, nu poate da, deocamdată, un răspuns tranşant la întrebarea „Cum sunt organizate cunoştinţele lexicale în mintea omului?”. S-au desfăşurat, însă, experimente care au dus la formularea ipotezei că lexiconul mintal (totalitatea cunoştinţelor noastre lexicale) are o organizare sistematică, semantica fiind principala coordonată de structurare a materialului lexical. Pentru specialiştii în prelucrarea limbajului realitatea psihologică a acestor relaţii nu are nicio relevanţă; pentru ei contează adecvarea computaţională.

Există câteva modele semantice ale lexiconului mintal, diferenţiate după felul în care este abordată reprezentarea formelor cuvintelor, a sensurilor, felul în care sunt reprezentate trăsăturile semantice, în care se interconectează cuvintele, în care sunt accesate. Cel mai răspândit model este cel al reţelelor semantice de tip wordnet.

Lexicografia de tip wordnet

Lexicografia este ştiinţa alcătuirii dicţionarelor. Se spune despre dicţionare că sunt instituţii sociale „vii” (Leech 1974: 203): este nevoie ca ele să se schimbe pentru a servi mai bine o societate în continuă schimbare. Pentru perioada actuală putem vorbi despre coexistenţa câtorva profiluri de utilizatori ai dicţionarelor:

1. tipul tradiţional – includem aici orice utilizator care preferă să consulte un dicţionar tipărit; poate fi vorba despre un vorbitor nativ, cu un nivel de educaţie cel puţin mediu, sau despre un vorbitor nenativ, care învaţă limba română, în cazul nostru;

2. tipul modern – este vorba despre utilizatorii de calculatoare, de Internet, mai mult sau mai puţin dependenţi de aceste mijloace de comunicare; sunt persoane care consultă dicţionare pe telefon, pe diverse dispozitive de citire a cărţilor în format electronic; pot fi vorbitori nativi sau nenativi;

3. tipul tehnic – aici încadrăm specialiştii în prelucrarea limbajului natural care dezvoltă aplicaţii ce folosesc cunoştinţe lexicale.

Pentru fiecare trebuie gândit un alt tip de produs lexicografic. Astfel, pentru utilizatorul tradiţional se pretează un dicţionar tipărit cu informaţiile obişnuite. Pentru cel modern – un dicţionar în format electronic, conţinând aceleaşi informaţii ca cel tipărit. Un avantaj pe care îl prezintă dicţionarul electronic este faptul că nu există constrângeri asupra dimensiunii. O consecinţă imediată ar fi (şi) absenţa criteriilor riguroase de selectare a materialului de inclus în dicţionar. Pentru „tehnicieni" nu mai discutăm despre dicţionare, ci despre resurse lingvistice organizate astfel încât informaţia conţinută să poată fi prelucrată de calculator, în vederea îmbunătăţirii rezultatelor diverselor aplicaţii create. Dintre tipurile noastre de utilizatori lipsesc specialiştii lingvişti. În funcţie de metoda de lucru adoptată, considerăm că aceştia se distribuie în diverse proporţii între cele 3 tipuri.

Boguraev şi Briscoe (1989: 4–5) identifică cinci tipuri de cunoştinţe necesare în diverse sarcini din prelucrarea limbajului natural:

- cunoştinţe *fonologice* – referitoare la sistemul de sunete şi structura cuvintelor şi a propoziţiilor;
- cunoştinţe *morfologice* – referitoare la structura internă a cuvintelor;
- cunoştinţe *sintactice* – referitoare la organizarea cuvintelor în propoziţii şi în fraze;
- cunoştinţe *semantice* – privitoare la sensul cuvintelor şi la felul în care se formează sensul propoziţiilor din sensurile cuvintelor conţinute;
- cunoştinţe *pragmatice/„enciclopedice"* – referitoare la circumstanţele comunicării.

Este dificil de creat o resursă cu toate aceste tipuri de informaţii, totuşi, mai ales că probabilitatea ca o aplicaţie să aibă nevoie de toate este redusă.

Experienţa din ultimii aproape 30 de ani a demonstrat că reprezentarea cunoştinţelor lexicale sub forma reţelelor semantice de tip wordnet este potrivită pentru utilizarea lor de către calculatoare. De peste două decenii se poate vorbi despre o lexicografie de tip wordnet, creată după modelul Princeton WordNet (PWN) (Miller 1993, Fellbaum 1998). Acesta este realizat manual, de o echipă de lexicografi. Are dimensiuni mari, ultima versiune conţinând 155287 de cuvinte (http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html). Numărul este impresionant, dar trebuie menţionat faptul că în PWN există literali simpli, dar şi numeroase combinaţii relativ stabile de cuvinte. În plus, absenţa restricţiilor de spaţiu şi a criteriilor ferme de filtrare a cuvintelor de inclus în resursă a favorizat reprezentarea a numeroase nume proprii (*Eneida*, *Internaţionala*, *Înălţarea*) şi a multor termeni din diverse domenii (*jurământul_lui_Hippocrate*, *scandalul_Watergate*, *aterizare_forţată*). Prezenţa lor este îndreptăţită de nevoile diverselor aplicaţii: recunoaştere automată a entităţilor numite, recunoaşterea unităţilor lexicale din diverse domenii, calculul frecvenţei termenilor pentru a stabili relevanţa conţinutului unui document pentru un anumit domeniu în vederea extragerii de informaţii etc.

Principii de organizare a reţelelor semantice de tip wordnet

Cuvintele lexicalizează concepte. Un cuvânt lexicalizează atâtea concepte câte sensuri are. Un concept poate fi lexicalizat de un cuvânt (*celulă* cu sensul „element constitutiv fundamental al organismelor vii, alcătuit din membrană, citoplasmă şi nucleu, reprezentând cea mai simplă unitate anatomică"), de mai multe cuvinte (*alegere*, *opţiune* cu sensul „faptul sau dreptul de a alege din două sau mai multe lucruri, posibilităţi etc. pe acela sau pe aceea care îţi convine") sau de nici un cuvânt (pentru sensul „persoană care îi trezeşte pe alţii din somn" nu avem un cuvânt în română). Cuvintele care lexicalizează acelaşi concept sunt în relaţie de sinonimie.

Într-o reţea de tip wordnet există noduri şi arce între ele. Fiecare nod corespunde unui concept; adică în fiecare nod al reţelei găsim o mulţime de cuvinte sinonime. Ele formează un sinset. Cuvintele polisemantice şi omonimele apar în atâtea sinseturi câte sensuri au.

Din perspectivă multilingvă, acelaşi concept poate fi lexicalizat în mai multe limbi prin cuvinte care sunt sinonime interlinguale. Pot fi limbi în care un anumit concept există, dar nu există cuvinte prin care să fie redat, putându-se apela doar la perifraze pentru descrierea sa. Vorbim în aceste cazuri despre goluri lexicale. Teoria specificului semantic al limbilor naturale (Coşeriu 1964) prevede faptul că diverse limbi decupează diferit spaţiul conceptual, ceea ce înseamnă că ierarhiile din limbi diferite nu sunt identice. Cu toate acestea, se pot invoca motive pentru alinierea reţelelor semantice pentru mai multe limbi: din perspectiva lingvisticii comparate, se găsesc cu mare uşurinţă golurile lexicale, deci dovezi în sprijinul teoriei amintite; din punct de vedere computaţionist, în tratamentul paralel al limbilor naturale este esenţial să utilizăm resurse lingvistice de acelaşi tip, cu care să se poată efectua raţionamente identice. Proiectul KYOTO (Fellbaum, Vossen 2012) face un pas în această direcţie propunând o ontologie formală la care să fie conectate toate wordneturile existente şi pe baza căreia să se poată face raţionamente.

Cuvintele unei limbi se grupează în două clase mari: cuvinte-conţinut şi cuvinte funcţionale. Locul celor dintâi este în lexicon, pe când cele din urmă aparţin componentei sintactice a limbii (conform observaţiilor asupra vorbirii subiecţilor afazici, vezi Garrett 1982), deci nu îşi găsesc reprezentare într-un lexicon de tip wordnet (Miller 1993).

Relaţiile din reţea sunt de două feluri: semantice (sau conceptuale) şi lexicale. Cele dintâi se stabilesc între noduri (i.e. corespondentele conceptelor) şi reprezintă arcele reţelei. Celelalte se stabilesc între formele cuvintelor. Dacă primele sunt valabile pentru oricare două cuvinte din sinseturile pe care le unesc, ultimele sunt valabile exclusiv pentru cuvintele între care se stabilesc, nu şi pentru sinonimele acestora, adică pentru cuvintele alături de care acestea apar în sinseturi. Un exemplu pentru primul caz îl reprezintă perechea de sinseturi {*predică*, *cazanie*, *omilie*, *cuvânt*} (cu glosa „cuvântare rostită de un cleric în biserică, în care se explică şi se comentează un text biblic şi se dau credincioşilor îndrumări morale") şi {*cuvântare*, *discurs*} (cu glosa „vorbire în public desfăşurată cu oarecare solemnitate"), primul fiind hiponimul celui de-al doilea; această relaţie stabilită la nivel de sinset este valabilă pentru orice pereche formată dintr-un cuvânt din primul sinset şi un cuvânt din al doilea sinset (i.e. produsul cartezian al celor două sinseturi considerate ca mulţimi). O relaţie lexicală este relaţia derivativă. Date fiind sinseturile {*conduce*,

şofa} („a conduce un automobil") şi {*conducător*, *şofer*} („persoană care conduce un automobil sau un autobuz"), se stabilesc relaţii derivative între *conduce* şi *conducător*, precum şi între *şofa* şi *şofer*, dar nu între *conduce* şi *şofer* şi nici între *şofa* şi *conducător*.

Tipuri de informaţii existente în wordnetul românesc

Wordnetul conţine, în primul rând, cuvinte. Informaţia morfologică despre acestea se limitează la partea de vorbire. Prin cuvinte înţelegem atât unităţi lexicale simple, cât şi îmbinări libere, ce corespund unui concept (*inel_de_logodnă*) sau unui concept creat artificial, pentru uşurinţa ierarhizării (*instrument_muzical_electronic*).

Puţine cuvinte sunt monosemantice. Deci, un literal apare în mai multe sinseturi. Tratarea cuvintelor polisemantice şi a omonimelor lexicale este identică: ele sunt înregistrate în wordnet cu toate sensurile lor, fără a semnaliza în vreun fel distincţia între ele (aşa cum se întâmplă în dicţionare, unde, în cazul omonimelor, se creează intrări diferite şi se notează numere la umărul din dreapta al cuvintelor); importantă este distincţia între sensuri; aproprierea sau depărtarea acestora reiese oricum din poziţia în reţea: sensurile omonime vor fi în arbori diferiţi, sensurile cuvântului polisemantic au o distanţă destul de mică între ele, uneori fiind chiar unite printr-o relaţie semantică (de exemplu, *plantă* cu sensul „nume generic dat organismelor vegetale, cu o organizare mai simplă decât a animalelor și care își extrag hrana prin rădăcini, caracterizându-se prin prezența clorofilei, prin faptul că membrana celulei este formată din celuloză și, în cazul speciilor superioare, prin alcătuirea corpului din rădăcină, tulpină și frunze" este hiperonim (indirect) pentru *plantă* cu sensul „vegetală, mai ales erbacee, cultivată de om sau care crește în mod natural și este utilă omului").

Între cuvinte se stabilesc două tipuri de relaţii: semantice şi lexicale. În terminologia wordnet, ele au interpretările prezentate mai sus.

Relaţiile semantice se stabilesc la nivel conceptual. Sinseturile, fiind lexicalizări ale conceptelor, sunt unite prin relaţii semantice. Experimentele psiholingvistice au găsit dovezi pentru organizarea diferită a cuvintelor-conţinut, în funcţie de partea de vorbire căreia îi aparţin. Astfel, pentru substantive sunt înregistrate:

- hiperonimia – relația dintre un termen mai general, desemnând o clasă de obiecte și un termen mai specific, desemnând o subclasă de obiecte a clasei denumite cu ajutorul hiperonimului (relația inversă se numește hiponimie): *flaut* este hiponimul lui *suflători*, acesta fiind hiperonimul lui *flaut*;
- meronimia – relația dintre un cuvânt desemnând o parte constituentă, o substanță sau un membru și alt cuvânt desemnând întregul (relația inversă se numește holonimie): *arcuşul* este un meronim al *vioarei*, acesta fiind holonim pentru *arcuş* (mero/holonimie de tipul parte-întreg), *ţesutul* este un meronim al lui *organism* (mero/holonimie de tipul substanță-ceea ce o conține), *feminist* este un meronim al lui *feminism* (mero/holonimie de tipul membru-grup);
- relația de instanțiere se stabilește între un substantiv denumind o situație, un eveniment, un fapt, un obiect particular și un substantiv comun desemnând

clasa de obiecte din care face parte acesta: *Turnul_Eiffel* este o instanţă a conceptului *turn*.

Pentru verbe există:

- hiperonimia – relația dintre un verb cu sens mai general și unul cu sens mai specific; relația inversă este hiponimia: *a se mișca* este hiperonim pentru *a se ridica*;
- troponimia – relația dintre două verbe, astfel încât primul exprimă un anumit mod de efectuare a acțiunii exprimate de cel de-al doilea: *a șopti* este un troponim al lui *a vorbi*;
- implicaţia lexicală – un verb v_1 implică un alt verb v_2 dacă acțiunea exprimată de v_1 este condiţionată de efectuarea acțiunii redate de v_2: *a visa* implică *a dormi*;
- relaţia de cauzalitate – de exemplu, relaţia dintre verbul *a ucide* şi verbul *a muri*.

Verbele cu sens asemănător sunt marcate ca aparţinând unui grup: *a se comporta* (cu glosa „a avea o anumită conduită"), *a face* („a-şi lua înfăţişarea de..."), *a se preface* („a crea impresie falsă") formează un grup de verbe.

Sinseturile adjectivale sunt unite de sinseturi substantivale prin relaţia care specifică atributul pentru care adjectivele descriptive exprimă valori. Astfel, substantivul *temperatură* este un atribut cu *fierbinte, cald, răcoros* sau *rece* valori posibile.

Adjectivele relaţionale sunt unite prin relaţia de pertonimie (engl. *pertainymy*) de substantivele de la care s-au format: *chimic* este un pertonim al lui *chimie*.

Între adjectivele cu sensuri asemănătoare există o legătură *see_also* (vezi_şi): *răcoros* este în relaţie *see_also* cu *rece*.

Adjectivele descriptive se organizează în jurul unui centru. Între centru şi sateliţii săi există relaţia de similitudine: *nou-născut, adolescent, puber, copilăresc* se numără printre sateliţii lui *tânăr*.

Pentru adverbe nu există relaţii semantice care să le organizeze.

Relaţiile lexicale. Principala relaţie lexicală din wordnet este sinonimia: ea organizează cuvintele în sinseturi, în funcţie de sensurile pe care le au. De exemplu, *doctorie* (cu sensul 1) şi *medicament* (cu sensul 1) sunt sinonime atunci când denumesc un „preparat, [o] substanţă care se foloseşte pentru vindecarea, ameliorarea sau prevenirea unei boli".

Antonimia (directă) se stabileşte între literali de aceeaşi parte de vorbire, cu sensuri opuse. Astfel, *a ierna* este antonim al lui *a văra*, *dispariţie* al lui *apariţie* etc.

Antonimia este o relaţie între forme, căreia îi corespunde o opoziţie în plan conceptual (deci între sinseturi), nemarcată, însă, în reţea. Astfel, *perfecţiune* şi *imperfecţiune* sunt antonime, dar *perfecţiune* este sinonim (i.e. este în acelaşi sinset) cu *desăvârşire* şi *sublim*. Putem spune că între *imperfecţiune*, pe de o parte, şi *desăvârşire* şi *sublim*, pe de altă parte, există o relaţie de opoziţie conceptuală.

Antonimia indirectă se stabileşte între un literal adjectival, centru de grup, şi un literal dintr-un sinset satelit al antonimului primului literal. De exemplu, *copilăresc* este un antonim indirect al lui *bătrân*, deoarece *bătrân* şi *tânăr* sunt antonime directe, iar *copilăresc* este satelit al lui *tânăr*.

Relaţiile între cuvinte derivate şi bazele lor de derivare sunt tot de natură lexicală. Se stabilesc între cuvinte de diverse părţi de vorbire.

Informaţiile de natură *stilistică* se referă la domeniul căruia îi aparţin cuvintele, regiunea în care sunt folosite, registrul în care sunt utilizate.

Trebuie remarcat faptul că relaţiile conceptuale sunt transferabile dintr-o limbă în alta, iar cele lexicale şi stilistice sunt specifice limbii pentru care au fost precizate. Se poate întâmpla să regăsim echivalentele acestor relaţii şi în alte limbi. În cazul antonimiei situaţia este foarte frecventă, probabil rezultat al existenţei opoziţiei conceptuale. Şi în cazul relaţiilor derivative sunt multe astfel de potriviri, dar ele nu ne dau dreptul de a le transfera dintr-o limbă în alta.

All-inclusive

Despre wordnet putem spune că reprezintă mai multe resurse într-una singură, afimaţie pe care o argumentăm în cele ce urmează.

Dicţionar explicativ. Fiecare sinset are asociată o glosă; de obicei, în wordnetul românesc, ea este preluată din Dicţionarul Explicativ al Limbii Române (DEX 1996). Alteori, fie este vorba despre o adaptare a acesteia, fie despre o traducere a glosei pentru sinsetul corespunzător din wordnetul american (PWN, Miller 1993, Fellbaum 1998).

Orice tip de dicţionar semantic poate fi extras din wordnet. Exemplificăm mai jos discutând despre două tipuri: dicţionarul de sinonime şi cel de antonime.

Dicţionar de sinonime. Fiecare sinset este o posibilă intrare dintr-un dicţionar de sinonime. Dacă într-un dicţionar obişnuit de sinonime sunt enumerate cuvintele sinonime fără precizarea sensului pentru care sunt valabile, cititorul fiind nevoit să deducă acest sens (dacă îl ştie) sau să apeleze la un dicţionar explicativ pentru aflarea lui, în cazul wordnetului, distincţia semantică este deja operată, iar sinonimele sunt indicate pentru fiecare sens în parte; dacă sinonimia între două cuvinte este valabilă pentru mai multe dintre sensurile lor, atunci cele două cuvinte vor apărea împreună într-un număr de sinseturi egal cu numărul de sensuri pe care le au în comun; în plus, sensul exprimat de cuvintele din fiecare sinset este lămurit prin glosa ataşată acestuia.

Dicţionar de antonime. În wordnet există două tipuri de antonimie în cazul adjectivelor: directă şi indirectă. Antonimia directă se stabileşte între literali cu sensuri total opuse. Antonimia indirectă se stabileşte între un adjectiv a_1 şi oricare dintre adjectivele satelit ale adjectivului aflat în relaţie de antonimie cu a_1. În plus, putem extinde relaţia de antonimie, considerând şi opoziţia dintre cuvintele din acelaşi sinset cu cele în antonimie directă. Vezi discuţia despre sinseturile {*perfecţiune*, *desăvârşire*, *sublim*} şi {*imperfecţiune*} de mai sus. În consecinţă, precizia tipului de antonimie este mult mai mare.

În dicţionarele de antonime perechile sunt inventariate cel mai adesea fără precizarea sensului. În cazul (cel mai frecvent al) cuvintelor polisemantice, nu reiese clar pentru care dintre sensuri sunt antonime cele două cuvinte.

Testele psiholingvistice în care li se cere subiecţilor să indice cuvântul cu sens opus cuvântului dat dovedesc, prin răspunsurile acestora, existenţa relaţiei de opoziţie conceptuală: vezi Sîrbu (1977: 82–83) unde se prezintă rezultatele unei anchete psiholingvistice în cadrul căreia subiecţii au indicat ca antonim pentru

fericire atât *nefericire*, cât şi *tristeţe*; sunt mai multe exemple de acest fel printre rezultate.

Dicţionar de cuvinte derivate. Marcarea relaţiilor derivative între cuvintele reţelei semantice asigură posibilitatea extragerii unui dicţionar de cuvinte derivate din aceasta. Pentru limba română, unirea literalilor derivaţi cu baza lor s-a făcut în sensul unui tratament unificator al diverselor tipuri de situaţii, chiar cu preţul încălcării „adevărului" etimologic. Pe de o parte, este vorba despre tratarea unitară a cuvintelor în care se poate recunoaşte acelaşi afix, chiar dacă nu toate sunt derivate pe teren românesc, unele fiind împrumutate, din ele decupându-se ulterior afixul şi folosit la derivarea pe teren românesc. Astfel, *veselie* este un împrumut din slavă, iar *hărnicie* un derivat românesc de la *harnic* (de origine slavă); existenţa lui *vesel* (împrumut din slavă) în română ne permite tratarea similară a perechilor *harnic – hărnicie*, *vesel – veselie*: în structura morfematică a substantivului se recunoaşte adjectivul şi sufixul *-ie*.

Pe de altă parte, este vorba despre tratamentul similar al cazurilor de derivare regresivă (de exemplu, *şofa* de la *şofer*), al celor de derivare progresivă (*pădurar* de la *pădure*) şi al celor de substituţie de afix (*deschinga* de la *închinga*). Strategia care ne-a permis aceasta a fost stabilirea relaţiei între bază şi derivat, dar fără a-i preciza o direcţie.

Spre deosebire de un dicţionar de derivate obişnuit, în wordnet relaţiile de derivare sunt marcate doar între anumite sensuri ale cuvintelor, nu sunt valabile pentru toate sensurile cuvintelor ţintă. Explicaţia rezidă în faptul că derivarea este şi un fenomen semantic, nu doar formal. Între derivat şi baza sa se pot stabili diverse relaţii semantice. În wordnet, acestea apar ca etichete semantice, fiind deci valabile la nivelul sinseturilor. De exemplu, în cazul sinseturilor {*conduce*, *şofa*} („a conduce un automobil") şi {*conducător*, *şofer*} („persoană care conduce un automobil sau un autobuz") am marcat relaţii derivative între *conduce* şi *conducător* şi între *şofa* şi *şofer*; cele două relaţii au ataşată eticheta semantică AGENT, ceea ce înseamnă că derivatul este agent al verbului de la care s-a format; această etichetă este valabilă la nivelul sinseturilor: dacă întalnim o propoziţie precum

> Un şofer conducea maşina cu o alcoolemie de 1,59 mg/l (http://www.observatorcl.info/un-sofer-conducea-masina-cu-o-alcoolemie-de-159-mg l accesat la 13 septembrie 2012)

şofer trebuie interpretat ca Agent al verbului *conducea*.

Uneori, relaţiile derivative stabilite între diverse sensuri ale aceloraşi cuvinte poartă etichete semantice diferite. Între *sătean* cu sensul „persoană care locuieşte într-un sat" şi *sat* „aşezare rurală a cărei populaţie se ocupă în cea mai mare parte cu agricultura" există o relaţie derivativă căreia i se poate ataşa eticheta semantică OF ORIGIN (i.e. derivatul desemnează o persoană cu originea în locul indicat de bază), iar între *sătean* cu acelaşi sens şi *sat* „locuitorii dintr-un sat" există o relaţie derivativă pentru care eticheta semantică potrivită este MEMBER MERONYM (i.e. derivatul este un meronim al bazei).

În măsura în care reţeaua semantică include relaţii către cuvintele derivate şi cele de la care acestea s-au format, putem spune că ea cuprinde şi un dicţionar etimologic.

Dicţionar multilingv. Existenţa wordneturilor pentru limbi diferite aliniate între ele la nivel de sinset şi păstrătoare ale aceloraşi ierarhii semantice permite stabilirea corespondenţelor lexico-semantice între acele limbi. Reiese clar ce cuvinte exprimă un anumit concept în diverse limbi, deci sunt sinonime interlinguale, precum şi pentru care dintre sensurile lor.

Dicţionar sintatic. Pentru peste 600 de sinseturi verbale, în wordnetul românesc există cadre de subcategorizare. O noutate în modul de formulare a acestor cadre o constituie indicarea, pentru fiecare rol semantic al unui verb, a literalului cu sensul aferent acestuia din wordnet cu care poate participa la satisfacerea respectivei valenţe a verbului. Iată cadrul de valenţă pentru verbul *a mânca*:

{*mânca*} nom*AG(persoană:1|animal:1)=acc*SUBSTANCE(hrană:1)

Subiectul acestui verb are rolul semantic Agent, iar el poate fi satisfăcut de *animal* cu sensul 1 din wordnet – de *persoană* cu sensul 1 din wordnet – sau orice substantiv la cazul nominativ (nom), care apare în wordnet ca hiponim al lui *persoană:1* sau al lui *animal:1*. Obiectul direct, în cazul acuzativ (acc), are rolul semantic Substanţă şi poate fi satisfăcut de substantivul *hrană* cu sensul 1 sau de oricare dintre hiponimele sale din wordnet.

Concluzii

Nevoile diversificate şi tipurile diferite de utilizatori au condus la noi perspective în evoluţia lexicografiei, ajungându-se chiar la crearea termenului *e-lexicografie*.

Perioada contemporană se caracterizează prin continuitatea tradiţiei în reprezentarea cunoştinţelor lexicale, dar şi prin inovaţii în ceea ce priveşte modalităţile de reprezentare şi organizare a acestora. Evoluţiile tehnice au favorizat sau au impus majoritatea acestor inovaţii.

Mulţumiri

Dezvoltarea şi distribuirea wordnetului românesc se fac acum în cadrul proiectului european METANET4U (ICT PSP grant #270893). Marcarea relaţiilor derivative şi a etichetelor semantice corespunzătoare se realizează în cadrul proiectului „Valorificarea identităţilor culturale în procesele globale”, cofinanţat de Uniunea Europeană şi Guvernul României din Fondul Social European prin Programul Operaţional Sectorial Dezvoltarea Resurselor Umane 2007–2013, contractul de finanţare nr. POSDRU/89/1.5/S/59758.

Bibliografie

Boguraev, Briscoe 1989: B. Boguraev, E. Briscoe, *Introduction*, în B. Boguraev, T. Briscoe (Eds.), *Computational Lexicography for Natural language Processing*, London, New York, Longman, p. 1–40.

Coşeriu 1964: E. Coşeriu, *Pour une sémantique diachronique structurale*, în *Travaux de linguistique et de littérature*, vol I, Strasbourg.

DEX 1996: *Dicţionar explicativ al limbii române* (DEX), ediţia a doua, Bucureşti, Editura Univers Enciclopedic.

Evens 1988: Martha Walton Evens (ed.), *Relational Models of the Lexicon*, Cambridge University Press.

Fellbaum 1998: Christiane Fellbaum (ed.), *WordNet: An electronic lexical database*, Cambridge, MA, MIT Press.

Fellbaum, Vossen 2012: Christiane Fellbaum, Piek Vossen, *Challenges for a multilingual wordnet*, în *Language Resources and Evaluation*, publicat online.

Garrett 1982: M.F. Garrett, *Production of Speech: Observations from Normal and Pathological Language Use*, în A. Ellis (ed.), *Normality and Pathology in Cognitive Functions*, London, Academic Press.

Gazdar, Klein *et alii* 1985: G. Gazdar, E. H. Klein, G. K. Pullum, I. A. Sag, *Generalized Phrase Structure Grammar*, Oxford, Blackwell şi Cambridge, MA, Harvard University Press.

Kaplan, Bresnan 1982: R.M. Kaplan, J. Bresnan, *Lexical-functional grammar. A formal system for grammatical representation*, în J. Bresnan (ed.), *The mintal representation of grammatical relations*, Cambridge, MIT Press, p. 173–281.

Leech 1974: G. Leech, *Semantics*, Harmondsworth, Penguin.

Miller 1993: G. Miller et al., *Five Papers in WordNet. Technical Report CSL Report 43*, Cognitive Science Laboratory, Princeton University.

Pollard, Sag 1987: C. Pollard, I.A. Sag, *Information-based Syntax and Semantics*, volume 1. *Fundamentals*, Stanford, CSLI Publications.

Pollard şi Sag 1994: C. Pollard, I.A. Sag, *Head-driven Phrase Structure Grammar*, Chicago, University of Chicago Press.

Sîrbu 1977: Richard Sîrbu, *Antonimia lexicală în limba română*, [Timişoara], Editura Facla.

The All-inclusive System in the Representation of Lexical Knowledge

Lexicography is a domain of activity that must keep pace with the needs of the users. Nowadays, when there is a diversified range of users, from the traditional kind (looking up a printed dictionary) to the engineer specialized in the natural language processing, a new way of representing lexical knowledge has emerged: the semantic network, the most widely known type of it being the wordnet. Its advantage is that it makes such knowledge accessible both to humans and to machines. In a wordnet content words are organized in synsets according to their meanings: each word occurs as many times as many meanings it has. Synsets are interlinked by semantic relations of various types (hypo/hyperonymy, meronymy/holonymy, troponymy, cause, etc.). Word forms can be further linked by lexical relations (synonymy, antonymy, derivational relations). Semantic relations are conceptual, thus have cross-lingual validity, while lexical relations are language specific. A language resource such as a wordnet can be seen as a repository of more resources: as synsets are associated a gloss, a wordnet can be looked up as an explanatory dictionary; from it one can extract various semantic relations dictionaries or even a derivational dictionary. Once such resources are created for various languages and they are aligned (i.e. the corresponding synsets in the different languages are clearly marked and the semantic relations between them are considered to be the same), one can extract from them multilingual dictionaries in which the interlingual correspondences are marked at the word sense level, not at the word level. A wordnet is extremely valuable for natural language processing: it is like a thesaurus, from which the semantic and lexical relations between semantically disambiguated words can be easily exploited in various tasks (such as question answering, information retrieval, and others) involving the expansion of the key words inserted by a user or calculating the semantic similarity between words.