

Sistemul all-inclusive în reprezentarea cunoștințelor 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 și Bresnan, *Lexical Functional Grammar*, 1982, Gazdar, Klein, Pullum, Sag, *Generalized Phrase Structure Grammar*, 1985, Pollard și Sag, *Head-driven Phrase Structure Grammar*, 1987 și 1994) accentuează rolul lexiconului în studiul limbii. Concepute ca teorii alternative la gramatica transformațională (dezvoltată în tradiție 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 evoluție în lingvistica teoretică a fost dublată de preocupările din lingvistica computațională de a înțelege sintaxa unei limbi, lucru imposibil fără cunoștințe de semantică și vocabular. Concentrându-se inițial în special asupra traducerii automate, ulterior și (sau mai ales) asupra dezvoltării unor aplicații utile societății per ansamblu sau diverselor domenii de lucru, inginerii au căutat cel mai convenabil mod de reprezentare a informațiilor 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 lingviștii, psiholingviștii, sociolingviștii și inginerii în prelucrarea și în generarea limbajului natural, sub formă scrisă sau orală.

Cunoștințele lexicale reprezintă totalitatea informațiilor despre cuvinte. Aceste informații sunt atât lingvistice (fonetice, morfologice, sintactice), cât și extralingvistice (pragmatice, ontologice, conceptuale). Ele sunt deținute de om, într-un mod mai mult sau mai puțin conștient. Sunt îmbogățite permanent, actualizate, atunci când este necesar, și accesate, de obicei, cu ușurință, dovadă a caracterului lor organizat. Un alt aspect important în legătură cu accesul la acestea este modalitatea bidirecțională în care se poate face: și de la cuvânt la informațiile despre el, și de la aceste informații la cuvântul pentru care sunt valabile: astfel, vorbitorul este capabil atât să găsească cuvântul potrivit pentru transmiterea unui sens, cât și să identifice sensul corect al unui cuvânt.

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

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

circumscrie semanticii lexicale sincronice, care se ocupă cu studiul semnificației 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-și calitățile în lingvistica descriptivă, această abordare nu poate fi totuși folosită pentru crearea lexicoanelor în vederea adnotării morfo-sintactice a corpusurilor și nici pentru generarea limbajului;

- analiza componențială: fiecare cuvânt este definit printr-o mulțime de seme care îl încadrează alături de termenii dintr-un anumit domeniu, precum și 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 relațională: definirea sensurilor unui cuvânt se face prin specificarea relațiilor pe care le contractează în cadrul domeniului (cf. Evens 1988: 2). Cea mai mare parte a vocabularului se pretează la o descriere de tip relațional, dar există și arii care nu pot fi organizate astfel, după cum se va vedea mai jos.

Psiholingvistica, i.e. studiul factorilor psihologici și neurobiologici care îi ajută pe oameni să achiziționeze, să învețe, să folosească, să înțeleagă și să producă limbajul, nu poate da, deocamdată, un răspuns tranșant la întrebarea „Cum sunt organizate cunoștințele lexicale în mintea omului?”. S-au desfășurat, însă, experimente care au dus la formularea ipotezei că lexiconul mintal (totalitatea cunoștințelor noastre lexicale) are o organizare sistematică, semantica fiind principala coordonată de structurare a materialului lexical. Pentru specialiștii în prelucrarea limbajului realitatea psihologică a acestor relații nu are nicio relevanță; pentru ei contează adecvarea computațională.

Există câteva modele semantice ale lexiconului mintal, diferențiate 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 rețelelor semantice de tip wordnet.

Lexicografia de tip wordnet

Lexicografia este știința alcătuirii dicționarelor. Se spune despre dicționare că sunt instituții 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 coexistența câtorva profiluri de utilizatori ai dicționarelor:

1. tipul tradițional – includem aici orice utilizator care preferă să consulte un dicționar tipărit; poate fi vorba despre un vorbitor nativ, cu un nivel de educație cel puțin mediu, sau despre un vorbitor nenativ, care învață limba română, în cazul nostru;

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

3. tipul tehnic – aici încadrăm specialiștii în prelucrarea limbajului natural care dezvoltă aplicații ce folosesc cunoștințe lexicale.

Pentru fiecare trebuie gândit un alt tip de produs lexicografic. Astfel, pentru utilizatorul tradițional se pretează un dicționar tipărit cu informațiile obișnuite. Pentru cel modern – un dicționar în format electronic, conținând aceleași informații ca cel tipărit. Un avantaj pe care îl prezintă dicționarul electronic este faptul că nu există constrângeri asupra dimensiunii. O consecință imediată ar fi (și) absența criteriilor riguroase de selectare a materialului de inclus în dicționar. Pentru „tehnicieni” nu mai discutăm despre dicționare, ci despre resurse lingvistice organizate astfel încât informația conținută să poată fi prelucrată de calculator, în vederea îmbunătățirii rezultatelor diverselor aplicații create. Dintre tipurile noastre de utilizatori lipsesc specialiștii lingviști. În funcție de metoda de lucru adoptată, considerăm că aceștia se distribuie în diverse proporții între cele 3 tipuri.

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

- cunoștințe *fonologice* – referitoare la sistemul de sunete și structura cuvintelor și a propozițiilor;
- cunoștințe *morfologice* – referitoare la structura internă a cuvintelor;
- cunoștințe *sintactice* – referitoare la organizarea cuvintelor în propoziții și în fraze;
- cunoștințe *semantice* – privitoare la sensul cuvintelor și la felul în care se formează sensul propozițiilor din sensurile cuvintelor conținute;
- cunoștințe *pragmatice/„enciclopedice”* – referitoare la circumstanțele comunicării.

Este dificil de creat o resursă cu toate aceste tipuri de informații, totuși, mai ales că probabilitatea ca o aplicație să aibă nevoie de toate este redusă.

Experiența din ultimii aproape 30 de ani a demonstrat că reprezentarea cunoștințelor lexicale sub forma rețelelor 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 conținând 155287 de cuvinte (<http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html>). Numărul este impresionant, dar trebuie menționat faptul că în PWN există literalii simpli, dar și numeroase combinații relativ stabile de cuvinte. În plus, absența restricțiilor de spațiu și a criteriilor ferme de filtrare a cuvintelor de inclus în resursă a favorizat reprezentarea a numeroase nume proprii (*Eneida, Internaționala, Înălțarea*) și a multor termeni din diverse domenii (*jurământul_lui_Hippocrate, scandalul_Watergate, aterizare_forțată*). Prezența lor este îndreptățită de nevoile diverselor aplicații: recunoaștere automată a entităților numite, recunoașterea unităților lexicale din diverse domenii, calculul frecvenței termenilor pentru a stabili relevanța conținutului unui document pentru un anumit domeniu în vederea extragerii de informații etc.

Principii de organizare a rețelelor 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 și nucleu, reprezentând cea mai simplă unitate anatomică”), de mai multe cuvinte (*alegere, opțiune* cu sensul „faptul sau dreptul de a alege din două sau mai multe lucruri, posibilități etc. pe acela sau pe aceea care îți convine”) sau de nici un cuvânt (pentru sensul „persoană care îi trezește pe alții din somn” nu avem un cuvânt în română). Cuvintele care lexicalizează același concept sunt în relație de sinonimie.

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

Din perspectivă multilingvă, același 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 (Coșeriu 1964) prevede faptul că diverse limbi decupează diferit spațiul conceptual, ceea ce înseamnă că ierarhiile din limbi diferite nu sunt identice. Cu toate acestea, se pot invoca motive pentru alinierea rețelelor semantice pentru mai multe limbi: din perspectiva lingvisticii comparate, se găsesc cu mare ușurință golurile lexicale, deci dovezi în sprijinul teoriei amintite; din punct de vedere computaționist, în tratamentul paralel al limbilor naturale este esențial să utilizăm resurse lingvistice de același tip, cu care să se poată efectua raționamente identice. Proiectul KYOTO (Fellbaum, Vossen 2012) face un pas în această direcție propunând o ontologie formală la care să fie conectate toate wordneturile existente și pe baza căreia să se poată face raționamente.

Cuvintele unei limbi se grupează în două clase mari: cuvinte-conținut și cuvinte funcționale. Locul celor dintâi este în lexicon, pe când cele din urmă aparțin componentei sintactice a limbii (conform observațiilor asupra vorbirii subiecților afazici, vezi Garrett 1982), deci nu își găsesc reprezentare într-un lexicon de tip wordnet (Miller 1993).

Relațiile din rețea sunt de două feluri: semantice (sau conceptuale) și lexicale. Cele dintâi se stabilesc între noduri (i.e. corespondentele conceptelor) și reprezintă arcele rețelei. 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 și 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 și se comentează un text biblic și se dau credincioșilor îndrumări morale”) și {*cuvântare, discurs*} (cu glosa „vorbire în public desfășurată cu oarecare solemnitate”), primul fiind hiponimul celui de-al doilea; această relație stabilită la nivel de sinset este valabilă pentru orice pereche formată dintr-un cuvânt din primul sinset și un cuvânt din al doilea sinset (i.e. produsul cartezian al celor două sinseturi considerate ca mulțimi). O relație lexicală este relația derivativă. Date fiind sinseturile {*conduce,*

șofa } („a conduce un automobil”) și {*conducător, șofer*} („persoană care conduce un automobil sau un autobuz”), se stabilesc relații derivative între *conduce* și *conducător*, precum și între *șofa* și *șofer*, dar nu între *conduce* și *șofer* și nici între *șofa* și *conducător*.

Tipuri de informații existente în wordnetul românesc

Wordnetul conține, în primul rând, cuvinte. Informația morfologică despre acestea se limitează la partea de vorbire. Prin cuvinte înțelegem atât unități lexicale simple, cât și îmbinări libere, ce corespund unui concept (*inel_de_logodnă*) sau unui concept creat artificial, pentru ușurința ierarhizării (*instrument_muzical_electronic*).

Puține cuvinte sunt monosemantice. Deci, un literal apare în mai multe sinseturi. Tratarea cuvintelor polisemantice și a omonimelor lexicale este identică: ele sunt înregistrate în wordnet cu toate sensurile lor, fără a semnaliza în vreun fel distincția între ele (așa cum se întâmplă în dicționare, unde, în cazul omonimelor, se creează intrări diferite și se notează numere la umărul din dreapta al cuvintelor); importantă este distincția între sensuri; apropierea sau depărtarea acestora reiese oricum din poziția în rețea: sensurile omonime vor fi în arbori diferiți, sensurile cuvântului polisemantic au o distanță destul de mică între ele, uneori fiind chiar unite printr-o relație 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 relații: semantice și lexicale. În terminologia wordnet, ele au interpretările prezentate mai sus.

Relațiile semantice se stabilesc la nivel conceptual. Sinseturile, fiind lexicalizări ale conceptelor, sunt unite prin relații semantice. Experimentele psiholingvistice au găsit dovezi pentru organizarea diferită a cuvintelor-conținut, în funcție de partea de vorbire căreia îi aparțin. 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 *sufletori*, acesta fiind hiperonimul lui *flaut*;
- meronimia – relația dintre un cuvânt desemnând o parte constituantă, o substanță sau un membru și alt cuvânt desemnând întregul (relația inversă se numește holonimie): *arcușul* este un meronim al *vioarei*, acesta fiind holonim pentru *arcuș* (mero/holonimie de tipul parte-întreg), *țesutul* 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 instanță 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*;
- implicația lexicală – un verb v_1 implică un alt verb v_2 dacă acțiunea exprimată de v_1 este condiționată de efectuarea acțiunii redată de v_2 : *a visa* implică *a dormi*;
- relația de cauzalitate – de exemplu, relația dintre verbul *a ucide* și verbul *a muri*.

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

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

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

Între adjectivalele cu sensuri asemănătoare există o legătură *see_also* (vezi_și): *răcoros* este în relație *see_also* cu *rece*.

Adjectivalele descriptive se organizează în jurul unui centru. Între centru și sateliții săi există relația de similitudine: *nou-născut*, *adolescent*, *puber*, *copilăresc* se numără printre sateliții lui *tânăr*.

Pentru adverbe nu există relații semantice care să le organizeze.

Relațiile lexicale. Principala relație lexicală din wordnet este sinonimia: ea organizează cuvintele în sinseturi, în funcție de sensurile pe care le au. De exemplu, *doctorie* (cu sensul 1) și *medicament* (cu sensul 1) sunt sinonime atunci când denumesc un „preparat, [o] substanță care se folosește pentru vindecarea, ameliorarea sau prevenirea unei boli”.

Antonimia (directă) se stabilește între literalii de aceeași parte de vorbire, cu sensuri opuse. Astfel, *a ierna* este antonim al lui *a văra*, *dispariție* al lui *aparitie* etc.

Antonimia este o relație între forme, căreia îi corespunde o opoziție în plan conceptual (deci între sinseturi), nemarcată, însă, în rețea. Astfel, *perfectiune* și *imperfectiune* sunt antonime, dar *perfectiune* este sinonim (i.e. este în același sinset) cu *desăvârșire* și *sublim*. Putem spune că între *imperfectiune*, pe de o parte, și *desăvârșire* și *sublim*, pe de altă parte, există o relație de opoziție conceptuală.

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

Relațiile între cuvinte derivate și bazele lor de derivare sunt tot de natură lexicală. Se stabilesc între cuvinte de diverse părți de vorbire.

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

Trebuie remarcat faptul că relațiile conceptuale sunt transferabile dintr-o limbă în alta, iar cele lexicale și stilistice sunt specifice limbii pentru care au fost precizate. Se poate întâmpla să regăsim echivalentele acestor relații și în alte limbi. În cazul antonimiei situația este foarte frecventă, probabil rezultat al existenței opoziției conceptuale. Și în cazul relațiilor 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, afirmație pe care o argumentăm în cele ce urmează.

Dicționar explicativ. Fiecare sinset are asociată o glosă; de obicei, în wordnetul românesc, ea este preluată din Dicționarul 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 dicționar semantic poate fi extras din wordnet. Exemplificăm mai jos discutând despre două tipuri: dicționarul de sinonime și cel de antonime.

Dicționar de sinonime. Fiecare sinset este o posibilă intrare dintr-un dicționar de sinonime. Dacă într-un dicționar obișnuit de sinonime sunt enumerate cuvintele sinonime fără precizarea sensului pentru care sunt valabile, cititorul fiind nevoit să deducă acest sens (dacă îl știe) sau să apeleze la un dicționar explicativ pentru aflarea lui, în cazul wordnetului, distincția 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 atașată acestuia.

Dicționar de antonime. În wordnet există două tipuri de antonimie în cazul adjectivelor: directă și indirectă. Antonimia directă se stabilește între literalii cu sensuri total opuse. Antonimia indirectă se stabilește între un adjectiv a_1 și oricare dintre adjectivele satelit ale adjectivului aflat în relație de antonimie cu a_1 . În plus, putem extinde relația de antonimie, considerând și opoziția dintre cuvintele din același sinset cu cele în antonimie directă. Vezi discuția despre sinseturile {*perfectiune*, *desăvârșire*, *sublim*} și {*imperfectiune*} de mai sus. În consecință, precizia tipului de antonimie este mult mai mare.

În dicționarele 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 subiecților să indice cuvântul cu sens opus cuvântului dat dovedesc, prin răspunsurile acestora, existența relației de opoziție conceptuală: vezi Sîrbu (1977: 82–83) unde se prezintă rezultatele unei anchete psiholingvistice în cadrul căreia subiecții au indicat ca antonim pentru

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

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

Pe de altă parte, este vorba despre tratamentul similar al cazurilor de derivare regresivă (de exemplu, *șofa* de la *șofer*), al celor de derivare progresivă (*pădurar* de la *pădure*) și al celor de substituție de afix (*deschinga* de la *închinga*). Strategia care ne-a permis aceasta a fost stabilirea relației între bază și derivat, dar fără a-i preciza o direcție.

Spre deosebire de un dicționar de derivate obișnuit, în wordnet relațiile de derivare sunt marcate doar între anumite sensuri ale cuvintelor, nu sunt valabile pentru toate sensurile cuvintelor țintă. Explicația rezidă în faptul că derivarea este și un fenomen semantic, nu doar formal. Între derivat și baza sa se pot stabili diverse relații semantice. În wordnet, acestea apar ca etichete semantice, fiind deci valabile la nivelul sinseturilor. De exemplu, în cazul sinseturilor {*conduce*, *șofa*} („a conduce un automobil”) și {*conducător*, *șofer*} („persoană care conduce un automobil sau un autobuz”) am marcat relații derivative între *conduce* și *conducător* și între *șofa* și *șofer*; cele două relații au atașat 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 întâlnim o propoziție precum

Un șofer conducea mașina 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)

șofer trebuie interpretat ca Agent al verbului *conducea*.

Uneori, relațiile derivative stabilite între diverse sensuri ale acelorași cuvinte poartă etichete semantice diferite. Între *sătean* cu sensul „persoană care locuiește într-un sat” și *sat* „așezare rurală a cărei populație se ocupă în cea mai mare parte cu agricultura” există o relație derivativă căreia i se poate atașa eticheta semantică OF ORIGIN (i.e. derivatul desemnează o persoană cu originea în locul indicat de bază), iar între *sătean* cu același sens și *sat* „locuitorii dintr-un sat” există o relație derivativă pentru care eticheta semantică potrivită este MEMBER MERONYM (i.e. derivatul este un meronim al bazei).

În măsura în care rețeaua semantică include relații către cuvintele derivate și cele de la care acestea s-au format, putem spune că ea cuprinde și un dicționar etimologic.

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

Dicționar sintactic. 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 valențe a verbului. Iată cadrul de valență 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 Substanță și poate fi satisfăcut de substantivul *hrană* cu sensul 1 sau de oricare dintre hiponimele sale din wordnet.

Concluzii

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

Perioada contemporană se caracterizează prin continuitatea tradiției în reprezentarea cunoștințelor lexicale, dar și prin inovații în ceea ce privește modalitățile de reprezentare și organizare a acestora. Evoluțiile tehnice au favorizat sau au impus majoritatea acestor inovații.

Mulțumiri

Dezvoltarea și distribuirea wordnetului românesc se fac acum în cadrul proiectului european METANET4U (ICT PSP grant #270893). Marcarea relațiilor derivative și a etichetelor semantice corespunzătoare se realizează în cadrul proiectului „Valorificarea identităților culturale în procesele globale”, cofinanțat de Uniunea Europeană și Guvernul României din Fondul Social European prin Programul Operațional Sectorial Dezvoltarea Resurselor Umane 2007–2013, contractul de finanțare 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.
- Coșeriu 1964: E. Coșeriu, *Pour une sémantique diachronique structurale*, în *Travaux de linguistique et de littérature*, vol I, Strasbourg.
- DEX 1996: *Dicționar explicativ al limbii române* (DEX), ediția a doua, București, 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 și 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 mental 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 și 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*, [Timișoara], 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.