Recenzie

BĂNCI DE ARBORI

Treebanks

A. Abeillé (ed.)

Kluwer Academic Publishers, Dordrecht, 2003, 405 p. + xxvi

În domeniul prelucrării limbajului natural, lingvistica corpusului are rădăcini din ce în ce mai puternice. Corpusuri adnotate morfologic şi analizate sintactic (uneori chiar adnotate semantic şi/sau pragmatic) se construiesc pentru un număr în creştere de limbi naturale.

Volumul editat de A. Abeillé reuneşte 21 de articole grupate în două părţi: prima cuprinde lucrări a căror temă este prezentarea modalităţilor de construire a unei bănci de arbori de derivare; în partea a doua, sunt înfăţişate diverse întrebuinţări ale acestor bănci.

Lucrările din prima parte au, în linii mari, aceeaşi structură, autorii prezentând caracteristicile corpusurilor cu care au lucrat, schemele de adnotare şi metodologia folosite şi chiar utilitatea „post-proiect" a rezultatelor obţinute.

Limbile asupra cărora se concentrează articolele sunt: engleză, germană, cehă, poloneză, spaniolă, franceză, italiană, portugheză, chineză, japoneză, turcă.

Alegerea corpusului de adnotat depinde de scopul urmărit. În majoritatea cazurilor, corpusul este format din articole de ziar. Pentru portugheză, însă, el este reprezentat de fragmente din texte medievale, disponibile în format electronic, pentru că autorii şi-au propus automatizarea parţială a procesului de adnotare a acestora cu ajutorul instrumentelor şi resurselor dezvoltate pentru portugheza modernă. Pentru poloneză, banca se rezumă la un set de propoziţii (aparţinând limbii scrise) ilustrative pentru evaluarea gramaticii folosite (Head - driven Phrase Structure Grammar). În vederea alcătuirii unei bănci cu propoziţii adnotate cu tipurile de greşeli conţinute, corpusul german prezentat de Becker et al. conţine doar mesaje electronice.

Tipul de adnotare ales în fiecare proiect depinde de trei factori: caracteristicile limbii, existenţa studiilor sintactice (într-un anumit cadru gramatical) pentru limba vizată şi obiectivul proiectului. Majoritatea proiectelor fac adnotare morfologică şi analiză sintactică. Adnotarea semantică începe, însă, să se practice pe scară mai largă (vezi banca pentru cehă, italiană, chineză). Există şi adnotări ale elementelor de discurs (adnotări pragmatice) (precum întreruperi, false începuturi, marcatori de discurs, filtre etc. – pentru corpusul transcris de engleză din banca PENN).

Unii cercetători optează pentru adnotări ale elementelor de suprafaţă (vezi Bank of English etc.), alţii fac o analiză de adâncime, a elementelor vide, nelexicalizate, iar alţii combină elemente pe care le găsesc utile din ambele tipuri (vezi corpusul Negra, cel pentru japoneză, chineză, turcă). Situaţiile care ridică probleme sunt, în general, aceleaşi în majoritatea limbilor: elipse, coordonări, apoziţii, ambiguităţi. Autorii articolelor prezintă, pentru fiecare proiect, soluţiile adoptate. Un fenomen gramatical aparte este prezentat pentru japoneză, fără a se propune, însă, o soluţie: stabilirea relaţiei predicat-argumente în propoziţiile subordonate, dată fiind topica destul de liberă înregistrată.

În majoritatea proiectelor, adnotarea se face semiautomat: rezultatele etapei automate sunt verificate de adnotatori umani.

Autorii articolelor prezintă şi instrumentele folosite pe parcursul proiectului, precum şi posibilele utilizări ale resurselor create, anticipând cumva temele lucrărilor prezentate în partea a doua a volumului.

Aceste bănci de propoziţii analizate morfologic, sintactic şi chiar semantic şi pragmatic sunt utile atât cercetărilor lingvistice, cât şi celor de lingvistică computaţională, sociolingvistică, psiholingvistică. În a doua parte a acestui volum, sunt prezentate aplicaţii în lingvistica computaţională: standardizarea formatului acestor bănci (Ide şi Romary), evaluarea analizoarelor sintactice (Carroll et al., Lin), extragerea altor resurse lingvistice din banca de derivare (Bod, Neumann), inducţie gramaticală (Frank et al.).

În domeniul prelucrării limbajului natural, dezvoltarea unei bănci de derivare pentru o limbă naturală devine o etapă indispensabilă în procesul de construire a resurselor lingvistice. După cum ilustrează foarte

bine G. Sampson în articolul său, studiul automat al limbilor naturale a scos la iveală fenomene neinventariate în gramaticile „făcute pe hârtie", dar a şi contrazis, uneori, afirmaţii larg răspândite în lingvistica teoretică.

Volumul editat de Abeillé prezintă experienţele diverselor echipe de lucru în proiectele derulate sau în derulare. Este înfăţişată situaţia actuală în acest domeniu, dar lucrările pot servi şi ca modele pentru cei care sunt la început de drum, în realizarea unei asemenea resurse lingvistice, sau celor care abia îşi propun să o realizeze. Bibliografia aferentă fiecărei lucrări este şi ea utilă acestui scop.

Verginica Barbu Mititelu

Institutul de Cercetări pentru Inteligenţă Artificială,

Academia Română, Bucureşti