

Introducere

Credem că limba română este răsplătită de eforturile de analiză, documentare, păstrare și publicare ale institutelor de lingvistică și universităților în mai bine de 100 de ani de cercetare (pentru a remarca numai perioada inaugurată de Hașdeu prin activitatea la dicționarul tezur). În acești ani s-au elaborat și tipărit dicționare, s-au emis și dezbătut teorii, s-au constituit puncte de vedere oficiale și personale și a fost suficient timp chiar și pentru contestarea unora dintre ele și perpetuarea unor dispute.

Între timp, limba română nu a stat nici ea pe loc, iar mijloacele de a studia limba s-au schimbat de asemenea. Dacă, pentru a-i studia evoluția sau pentru a găsi filioane lingvistice încă nedescoperite, atenția cercetătorului rămâne captată în continuare de aspecte de fonologie, sintaxă, semantică, lexicologie, terminologie etc., randamentul și precizia observațiilor lui crește dacă face apel la metode de investigare informatică a limbii. De câțiva timp accesul la o carte se poate face și altfel decât ținând-o în mână și deschizând-o. Și nu mai e nevoie ca ea să existe în biblioteca de lângă noi ca s-o putem citi. Dintr-o dată a devenit posibil să ne uităm la o carte și altfel decât parcurgând-o în secvența ei liniară. Rafturile cu fișe de ocurențe ale lexicografilor, care luau ani pentru a fi completate, sunt acum generate automat prin metode de indexare de către programe și regăsirea unui context se face cât ai clipi...

Dar domeniile lingvisticii computaționale și ale tehnologiilor limbajului uman au repercusiuni și de altă natură decât ca metode de cercetare asupra unei limbi. Aplicații de prelucrare a limbajului natural care să deschidă un nou tip de acces la informații pot fi acum concepute. Textul, chiar și în format electronic, începe să fie privit și altfel decât ca un șir de caractere sau de cuvinte. Au început să apară metode de a pătrunde în structura lui sintactică și semantică încât structura și înțelesul textului să poată fi relevate mașinii și ea să poată opera cu ele așa cum operează cu numere, de când a fost ea inventată. Începem să știm cum să facem mașinile noastre să execute un alt tip de „calcul”, mai apropiat de modul nostru de gândire, și care-și găsește originea în text...

Limba română trebuie să ajungă la nivelul de tehnologizare de care se pot mândri astăzi alte limbi intens studiate. Rostul acestei cărți, pe care o dorim prima dintr-o serie, trebuie atașat acestei ambiții. Ea este scrisă de lingviști și informaticieni români care, spre marea noastră bucurie, încep să se înțeleagă din ce în ce mai bine. Este exact ceea ce a urmărit acel grup de constituire a Comisiei de Informatizare pentru Limba Română, când, în martie 2001, s-a reunit pentru prima dată în sediul de pe Calea Victoriei al Academiei Române. Ulterior, această întâlnire a devenit o tradiție prin organizarea anual în București, Iași și Chișinău a unor sesiuni de lucru ale unui grup largit, care, din acest motiv s-a numit Consorțiu. De doi ani am dorit să invităm la aceste întâlniri și cercetători aflați la mai mare distanță de noi. Ca urmare, ultimele două întâlniri au căpătat caracterul de ateliere de lucru și au fost organizate în regim de teleconferință. Am putut asculta astfel glasuri de români care lucrează în universități din America, Germania, Italia, Franța și Anglia, după cum și ei ne-au putut urmări pe noi.

Întâlnirea din 3 noiembrie 2006 a Atelierului, a fost găzduită de Biblioteca Facultății de Informatică a Universității „Al.I.Cuza” din Iași și a beneficiat de implicarea MEC în finanțare. Această generoasă contribuție bănească ne-a permis să-i îmbunătățim organizarea, dar mai ales, să tipărim această carte. Îi suntem recunoscători pentru acest ajutor, cu precădere d-nei Veronica Bubulete. Mulțumim totodată participanților la atelier, aflați în sală sau conectați prin Internet, cât și colectivului de recenzori care ne-au ajutat să îmbunătățim calitatea lucrărilor.

Cuprins

Introducere

Capitolul 1. Resurse lingvistice pentru prelucrarea vorbirii1

Situl ‘Limba Română Vorbită’ Horia-Nicolai Teodorescu, Monica Feraru, Diana. Trandabăț.....	3
Schemă XML de adnotare a intonației în cadrul corpusurilor de text Vasile Apopei, Doina Jitcă	9

Capitolul 2. Dicționare și corpusuri adnotate pentru prelucrarea textelor.....15

Noi dezvoltări ale wordnet-ului românesc Dan Tufiș, Verginica Barbu Mititelu, Alexandru Ceaușu, Luigi Bozianu, Cătălin Mihăilă, Margareta Manu Magda	17
Framenet român: tentativă de elaborare Victoria Bobicev, Victoria Maxim, Tatiana Zidrașco, Alina Iaciurinschi.....	23
DEI Multimedia: evoluții, perspective Dumitru Todoroi, Adrian Chiorescu	29
Maparea cuvintelor dintr-un lexicon pe ontologie Natalia Burciu, Antonina Bîrlădeanu	35
Crearea resurselor lingvistice cu ajutorul unui limbaj specializat Ștefan Diaconescu	39
Resurse lingvistice românești în format electronic. <i>Biblia 1688</i> Bogdan-Mihai Aldea, Gabriela Haja	45
Resurse românești în cadrul proiectului LT4eL Diana Trandabăț, Adrian Iftene, Ionuț Pistol, Corina Forăscu, Dan Cristea.....	51
Tehnici de validare și corecție focalizată a adnotării morfo-sintactice în corpusuri de mari dimensiuni Dan Tufiș, Elena Irimia	57
RoGER – un corpus paralel aliniat Monica Gavrilă, Natalia Elița.....	63
TimeBank 1.2: O versiune adnotată în limba română Corina Forăscu, Radu Ion.....	69
Resurse lingvistice reutilizabile Constantin Ciubotaru, Svetlana Cojocar, Elena Boian, Alexandru Colesnicov, Ludmila Malahova, Valentina Demidov, Oleg Burlaca.....	75
Capitolul 3. Aplicații ale tehnologiilor lingvistice81	
Sisteme de Întrebare Răspuns pentru limba română Adrian Iftene, Ionuț Pistol, Diana Trandabăț, Georgiana Pușcașu, Corina Forăscu, Dan Cristea.....	83
Identificarea și extragerea automată a cologațiilor din texte Dan Ștefănescu, Dan Tufiș, Elena Irimia	89

Spre o extragere automată a cologațiilor: cazul verbului “a face” Amalia Todirașcu	95
Rezoluția anaforei pentru limba română Gabriela Pavel, Oana Postolache, Ionuț Pistol, Dan Cristea	101
Instrumente pentru consultarea Atlasului Lingvistic și editarea textelor dialectale Silviu Bejinariu, Vasile Apopei, Ramona Luca, Luminița Botoșineanu, Florin Olariu ...	107
Generare de concordanțe pentru dicționarul limbajului poetic eminescian Mihaela Brut, Dumitru Irimia, Oana Panait	113
Crearea unui generator morfologic pentru verbele din limba română Antonina Bîrlădeanu, Natalia Burciu	119
Parsarea predicatului (verbal / nominal) și a clauzei (finite / nefinite) în limba română. Aplicare la parsarea FDG Alex Moruz, Neculai Curteanu, Diana Trandabăț, Iustin Dornescu, Cecilia Bolea	123
Prelucrarea resurselor românești în cadrul proiectului LT4eL Ionuț Pistol, Adrian Iftene, Diana Trandabăț, Dan Cristea, Corina Forăscu	129
Sistem de instruire asistată de calculator pentru morfologia limbii române Elena Boian, Constantin Ciubotaru, Svetlana Cojocar, Galina Magariu, Tatiana Verlan, Iuri Rogojin	135
Capitolul 4. Modelare lingvistică	141
Structura grupului verbal, predicția lexicală și reprezentarea logică a predicatului în limba română Neculai Curteanu, Diana Trandabăț, Mihai Moruz	143
Perspective semantice din nou: cum și sub ce formă avansăm lexicologic spre DLRI Cristina Florescu	149
Modelarea relațiilor semantice într-un dicționar de simboluri Cristina Ciocârlău, Mihaela Brut	155
Dreptul de publicare pe web Noemi Bomher	161
Modelare cu ontologii și adnotări Radu Cibotaru	165
Cadre pentru o implementare PC-PATR a verbelor tranzitive din limba română Nadia Luiza Huțuliac	171
Index de autori	
177	