Lucrările atelierului

Resurse lingvistice şi instrumente pentru prelucrarea limbii române

Iaşi, 3 noiembrie 2006

Editura Universității Alexandru Ioan Cuza Iaşi

Volum apărut cu sprijinul Ministerului Educaţiei şi Cercetării,
prin Autoritatea Naţională pentru Cercetare Ştiinţifică

Descrierea CIP a Bibliotecii Naţionale a României
Lucrările atelierului Resurse lingvistice şi instrumente pentru prelucrarea limbii române / Corina Forăscu, Dan Tufiş, Dan Cristea (editori) –
Iaşi, Editura Universităţii „Alexandru Ioan Cuza", 2006

ISBN: 978-973-703-208-9

I. Forăscu, Corina

II. Tufiş, Dan

III. Cristea, Dan

Lucrările atelierului
Resurse lingvistice şi instrumente pentru prelucrarea limbii române
Iaşi, 3 noiembrie 2006

Editori:

Corina Forăscu

Dan Tufiş

Dan Cristea

Organizatori:

Facultatea de Informatică,
Universitatea Al. I. Cuza – Iaşi

Institutul de Cercetări pentru Inteligență Artificială
Academia Română – Bucureşti

Institutul de Informatică Teoretică
Academia Română, Filiala Iaşi

Comitetul de Program

Corneliu Burileanu, Facultatea de Elecronică, Universitatea Politehnica Bucureşti şi Institutul de Cercetări pentru Inteligenţă Artificială, A.R., Bucureşti, România
Constantin Ciubotaru, Institutul de Matematică şi Informatică, Academia de Ştiinţe a Moldovei, Chişinău, R. Moldova
Svetlana Cojocaru, Institutul de Matematică şi Informatică, Academia de Ştiinţe a Moldovei, Chişinău, R. Moldova
Dan Cristea, Facultatea de Informatică, Universitatea "Al. I. Cuza" şi Institutul de Informatică Teoretică, A.R., Iaşi, România
Nicolae Curteanu, Institutul de Informatică Teoretică, A.R., Iaşi, România
Cristina Florescu, Institutul de Filologie Română "Al. Philippide", A.R., Iaşi, România
Corina Forăscu, Facultatea de Informatică, Universitatea "Al. I. Cuza", Iaşi şi Institutul de Cercetări pentru Inteligenţă Artificială, A.R., Bucureşti, România
Gabriela Haja, Institutul de Filologie Română "Al. Philippide", A.R., Iaşi, România
Cătălina Hallett, Open University, UK
Radu Ion, Institutul de Cercetări pentru Inteligenţă Artificială, A.R., Bucureşti, România
Rada Mihalcea, Universitatea North Texas, SUA
Constantin Orăsan, Universitatea Wolverhampton, Anglia
Oana Postolache, ISI - Universitatea California, SUA
Irina Prodanoff, ILC-Pisa şi Universitatea Pavia, Italia
Georgiana Puşcaşu, Universitatea Wolverhampton, Anglia
Valentin Tablan, Universitatea Sheffield, Anglia
Amalia Todiraşcu, Universitatea Marc Bloch, Strasbourg, Franţa
Dumitru Todoroi, Academia de Studii Economice, Chişinău, R.Moldova
Doina Tătar, Universitatea „Babeş-Bolyai", Cluj-Napoca, Romania
Horia-Nicolai Teodorescu, Institutul de Informatică Teoretică, A.R. şi Universitatea Tehnică, Iaşi, România
Dan Tufiş, Institutul de Cercetări pentru Inteligenţă Artificială, A.R., Bucureşti şi Universitatea "Al. I. Cuza", Iaşi, România
Adriana Vlad, Facultatea de Elecronică, Universitatea Politehnica Bucureşti şi Institutul de Cercetări pentru Inteligenţă Artificială, A.R., Bucureşti, România

Comitetul de Organizare

Alexandru Ceauşu, ICIA-AR (alceasu@racai.ro)
Dan Cristea, FII-UAIC şi IIT-AR (dcristea@info.uaic.ro)
Corina Forăscu, FII-UAIC şi ICIA-AR (corinfor@info.uaic.ro)
Adrian Iftene, FII-UAIC (adiftene@info.uaic.ro)
Elena Irimia, ICIA-AR (elena@racai.ro)
Ionuţ Pistol, FII-UAIC (ipistol@info.uaic.ro)
Dan Ştefănescu ICIA-AR (danstef@racai.ro)
Horia-Nicolai Teodorescu, IIT-AR şi UT Iaşi (hteodor@etc.tuiasi.ro)
Diana Trandabăţ, FII-UAIC şi IIT-AR (dtrandabat@info.uaic.ro)
Dan Tufiş, ICIA-AR şi FII-UAIC (tufis@racai.ro)

Introducere

Credem că limba română este răsplătită de eforturile de analiză, documentare, păstrare şi publicare ale institutelor de lingvistică şi universităților în mai bine de 100 de ani de cercetare (pentru a remarca numai perioada inaugurată de Haşdeu prin activitatea la dicționarul tezaur). În aceşti ani s-au elaborat şi tipărit dicționare, s-au emis şi dezbătut teorii, s-au constituit puncte de vedere oficiale şi personale şi a fost suficient timp chiar şi pentru contestarea unora dintre ele şi 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 filoane lingvistice încă nedescoperite, atenția cercetătorului rămâne captată în continuare de aspecte de fonologie, sintaxă, semantică, lexicologie, terminologie etc., randamentul şi precizia observațiilor lui creşte dacă face apel la metode de investigare informatică a limbii. De câtva timp accesul la o carte se poate face şi altfel decât ținând-o în mână şi deschizând-o. Şi 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 şi altfel decât parcurgând-o în secvența ei liniară. Rafturile cu fişe de ocurențe ale lexicografilor, care luau ani pentru a fi completate, sunt acum generate automat prin metode de indexare de către programe şi regăsirea unui context se face cât ai clipi...

Dar domeniile lingvisticii computaționale şi ale tehnologiilor limbajului uman au repercusiuni şi 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 şi în format electronic, începe să fie privit şi altfel decât ca un şir de caractere sau de cuvinte. Au început să apară metode de a pătrunde în structura lui sintactică şi semantică încât structura şi înțelesul textului să poată fi relevate maşinii şi ea să poată opera cu ele aşa cum operează cu numere, de când a fost ea inventată. Începem să ştim cum să facem maşinile noastre să execute un alt tip de „calcul", mai apropiat de modul nostru de gândire, şi care-şi găseşte 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 ataşat acestei ambiții. Ea este scrisă de lingvişti şi 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 Bucureşti, Iaşi şi Chişinău a unor sesiuni de lucru ale unui grup lărgit, care, din acest motiv s-a numit Consorțiu. De doi ani am dorit să invităm la aceste întâlniri şi cercetători aflați la mai mare distanță de noi. Ca urmare, ultimele două întâlniri au căpătat caracterul de ateliere de lucru şi au fost organizate în regim de teleconferință. Am putut asculta astfel glasuri de români care lucrează în universități din America, Germania, Italia, Franța şi Anglia, după cum şi 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 Iaşi şi 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 şi colectivului de recenzori care ne-au ajutat să îmbunătățim calitatea lucrărilor.

Editorii Iaşi, decembrie 2006

Cuprins

Capitolul 1

Resurse lingvistice pentru prelucrarea vorbirii

SITUL 'LIMBA ROMÂNĂ VORBITĂ'

HORIA-NICOLAI TEODORESCU[1, 2], MONICA FERARU[2], DIANA TRANDABĂȚ[1,3]

[1]*Institutul de Informatică Teoretică, Academia Română, Filiala Iaşi*

[2]*Universitatea Tehnică „Gh. Asachi", Iaşi*

[3] *Facultatea de Informatică, Universitatea "Al.I.Cuza", Iaşi*

{hteodor, mferaru}@etc.tuiasi.ro, dtrandabat@info.uaic.ro

Rezumat

Inițiativa construirii unei arhive publice a sunetelor a fost determinată de lipsa unei asemenea resurse pentru limba română, lipsă resimțită atât în cercetare, cât şi în învățământ. Situl include peste 600 de înregistrări, în diverse formate de precizie şi codare, ale sunetelor limbii române vorbite.

1. Introducere

Situl (http://www.etc.tuiasi.ro/sibm/romanian_spoken_language/index.htm) a fost creat prin colaborarea dintre Institutul de Informatică Teoretică al Academiei Române - Grupul de Prelucrarea Vorbirii, Universitatea "Al. I. Cuza" Iaşi - Facultatea de Informatică şi Universitatea Tehnică "Gh. Asachi" Iaşi - Centrul de Excelență în cercetare "CERFS" (coordonat de primul autor) în ideea realizării unui suport pentru un "dicționar al sunetelor" limbii române. Situl cuprinde, pe lângă sunetele propriu-zise (vocale, consoane, diftongi, scurte fraze), informații despre fonetica limbii române, protocoale de documentare şi de înregistrare, instrumente de analiză, trimiteri la lucrări referitoare la prelucrarea limbii vorbite etc.

Scopul acestei inițiative a fost realizarea unei arhive pentru sunetele limbii române cu următoarele caracteristici:

- Bază de date cu voci atât profesionale (pronunții "perfecte"), cât şi ne-profesionale ("vocea omului de pe stradă"), din zona Iaşi, iar apoi, pe cât posibil, cu pronunții (sunete) specifice diverselor regiuni.
- Pe baza acestor date, realizarea unui studiu statistic amplu al sunetelor limbii române, care să includă, de exemplu "triunghiul formanților limbii române", "caracteristici statistice ale pronunțiilor regionale" etc.
- Corectarea unor sisteme de sinteză ne-concatenativă, pe baza datelor din arhivă.
- Îmbunătățirea unor sisteme de recunoaştere acustică, pe baza datelor din arhivă. (Arhiva ar putea deveni, sperăm, un banc de probă pentru asemenea sisteme).

- Realizarea unei baze de date (sunete) a limbii vorbite pentru persoane cu diverse patologii (neurologice, laringiene, nazale, bucale, respiratorii) (Teodorescu et al., 2006a). Această bază de sunete produse de voci patologice va fi parţial utilă şi în cercetările pentru un grant CEEX.
- Realizarea unei baze de date de tip silabe şi cuvinte (sursă pentru sintetizoare concatenative şi banc de probă pentru sisteme de recunoaştere de cuvinte).
- Un dicţionar electronic al pronunţiilor din limba română. Corelarea cu Atlasele limbii române.
- Pagina de referinţe, care să prezinte toate titlurile de volume şi lucrări ce au ca obiect sunetele limbii române (fonetică, sinteză, recunoaştere etc.).

S-au efectuat peste 600 de înregistrări, în diverse formate de precizie şi codare (Teodorescu et al., 2006b). Fiecare vorbitor a rostit de trei ori fiecare frază, propoziţie, cuvânt, vocală, consoană, diftong, triftong, hiat precum şi grupuri de sunete specifice limbii române (ex. *ce*, *ci*, *che*, *chi* etc). Înregistrările au fost efectuate folosind programul Goldwave 5.0 la o frecvenţă de eşantionare de 22050Hz, codate pe 16 şi 24 biţi, mono.

Fişierele sunt grupate în clase, după cum urmează:

A. Sunete de bază: i) fişiere de vocale; ii) fişiere de consoane, înregistrate conform standardului IPA, în forma VCV, unde V este vocala *a*; iii) fişiere de diftongi, triftongi şi hiatusuri; iv) fişiere de sunete specifice, care în scrierea în limba română corespund grupurilor *ce*, *ci*, *che* (ke), *chi* (ki), *ge*, *gi*, *ghe*, *ghi*. Subiecţii sunt atât bărbaţi, cât şi femei, persoane cu vârsta cuprinsă între 26-31 ani, proveniţi (născuţi şi educaţi) din zona Moldovei de mijloc (judeţele Iaşi, Vaslui, Bacău), cu educaţie superioară şi fără patologii manifestate. Vocalele sunt înregistrate atât în varianta scurtă (pronunţie uzuală), cât şi în varianta „susţinută".

B. Scurte propoziţii sau segmente de fraze, cu încărcătură emoţională diferită

Pe lângă sunete simple au fost înregistrate şi fraze scurte: *Cine a făcut asta*, *Vine mama*, *Aseară*. Fiecărui subiect i s-a cerut să rostească fiecare frază simulând următoarele emoţii: fericire, tristeţe, bucurie, ură, optimism, pesimism, ton exclamativ, ton interogativ, ton plat şi starea de supărare. Ulterior, aceste stări au fost reduse la patru: fericire, supărare, furie şi ton neutru.

2. Metodologia de înregistrare

Analiza zgomotului de fond

Zgomotul de fond este un semnal aleator, arareori staţionar, mixat cu unele perturbaţii deterministe de tipul perturbaţiilor de frecvenţa ale reţelei. Pentru a asigura calitatea înregistrării, amplitudinea zgomotului trebuie să fie mult mai mică decât amplitudinea semnalului. Înregistrările au fost efectuate într-un laborator cu zgomot redus, dar nu s-a dispus de un spaţiu total izolat fonic. O înregistrare de bună calitate are un zgomot în al cărui spectru nu se depăşeşte valoarea de -80dB pentru nici o componentă spectrală, în timp ce formantul F1 are un nivel cu cel puţin 30dB mai mare, iar formanţii superiori au

tipic amplitudini cu peste 15-20dB peste nivelul zgomotului (Teodorescu, 2006a). Fişierele sunt în curs de verificare, urmând a fi eliminate cele care nu satisfac nivelul de calitate pe care ni l-am impus.

Alegerea microfonului

Microfonul este primul element în convertirea sunetelor şi are un rol esenţial în calitatea înregistrărilor. Un microfon de bandă limitată, de sensibilitate redusă sau cu zgomot mare poate compromite înregistrarea. Un microfon de calitate are zgomot redus şi raportul semnal / zgomot bun. Caracteristica omnidirecţională conduce la o sensibilitate mare la zgomotele ambientale şi nu este de dorit. Sunetele au fost înregistrate folosind căşti cu microfon SONIC Stereo Dinamic Headphones HP-259 cu caracteristicile: frecvenţa de răspuns: 20-20.000 Hz; impedanţa microfon: U=3V, R=1,5K Ω; impedanţa căşti: 32 Ω; sensibilitate microfon: -58dB±2; sensibilitate căşti: 100dB/mw; putere: 100mW.

Poziţionarea microfonului

O atenţie specială trebuie acordată poziţiei microfonului, deoarece apar zgomote sau distorsiuni introduse prin poziţionarea deficitară. Ţinerea microfonului prea aproape de gură poate duce la efectul de saturare a amplificatorului, cu rezultat de puternică distorsionare a semnalului. Se recomandă menţinerea microfonului mai jos de gură, aproximativ în dreptul bărbiei, la câţiva centimetri de aceasta. Distanţa de la bărbie trebuie să fie aproximativ egală cu distanţa până la buze.

Placa de sunet şi driverele corespunzătoare

Majoritatea calculatoarelor actuale conţin pe placa de bază circuite de preluare a semnalelor de la microfon şi de generare de semnale audio la căşti sau difuzoare (calitatea acestor circuite diferă substanţial de la o placă la alta). Placa de bază a calculatorului pe care au fost efectuate înregistrările este MB FOXCONN 760 GXK8MC-S, având încorporată o placă de sunet Sound MAX Digital Audio produsă de Analog Devices cu caracteristicile: procesor de semnal SiS964, standard AC '97.

3. Adnotări

Sunetele şi frazele disponibile pe sit au fost adnotate la diferite niveluri cu scopul de a avea un corpus pentru analiza statistică a datelor. Adnotarea s-a realizat folosind utilitarul Praat™ (www.praat.org), ales datorită eficienţei, recunoaşterii internaţionale şi uşurinţei în folosire. Primul pas a fost segmentarea la nivel de fonem. Ulterior, au fost grupate fonemele pentru a se realiza o segmentare în silabe, cuvinte şi propoziţii. Principala problemă a fost faptul că este dificil de stabilit întotdeauna cu exactitate unde se află graniţa dintre foneme. În figura 1 este prezentat un exemplu de adnotare pentru propoziţia *Vine mama*! . În viitor, adnotările vor fi validate prin analiza efectuată de mai mulţi adnotatori.

Informaţia de la nivel fonologic va fi completată cu informaţii prozodice (ton, intensitate, durată etc.), deoarece unul dintre obiectivele noastre este de a detecta parametri prozodici care fac diferenţa dintre vorbirea umană şi cea sintetică.

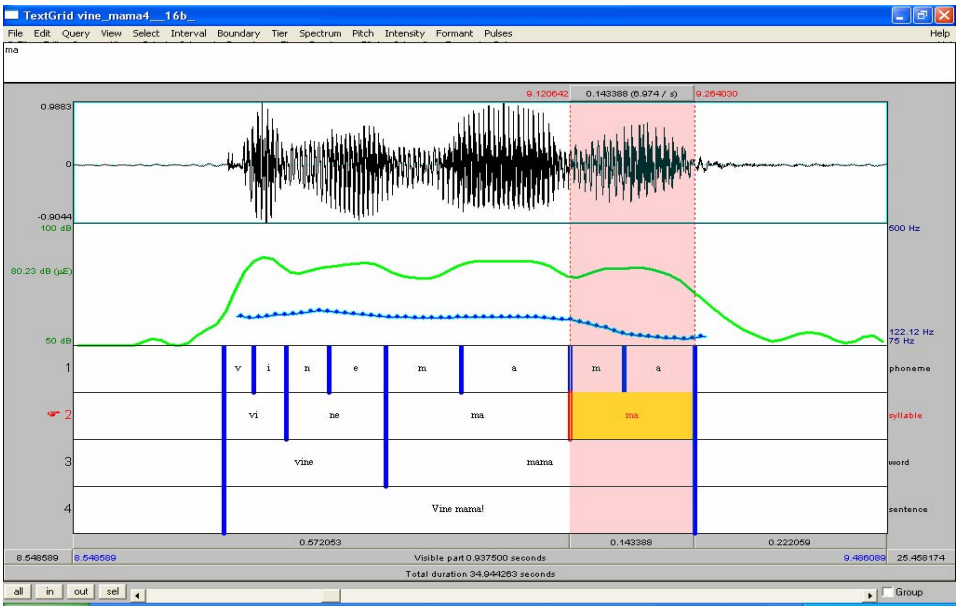


Figura 1: Exemplu de adnotare a propoziţiei *Vine mama!*

4. Alte elemente ale resursei: documentaţii, instrumente etc.

Subiecţii au fost informaţi anterior înregistrărilor despre obiectivele proiectului, fiind asiguraţi de confidenţialitatea datelor personale. Subiecţii au semnat un consimţământ informat în conformitate cu „Protocolul de Protecţie a Subiecţilor Umani" al *U.S. Food and Drug Administration* (http://www.fda.gov/cdrh/devadvice/ide/informed_consent.shtml) şi cu „Principiile etice ale Asociaţiei Acustice Americane privind cercetările care implică fiinţa umană" (http://asa.aip.org/ethical.html). Vorbitorilor li s-au explicat în prealabil condiţiile de înregistrare: poziţia microfonului, susţinerea vocalelor pe o durată cât mai lungă, dar fără a se realiza vreun efort etc. De asemenea, fiecare subiect a completat o fişă personală, care include date despre vârsta, sexul, limba, educaţia vorbitorului, patologii, precum şi evaluarea calităţii subiective a vocii. Pe lângă sunetele propriu-zise şi fişele vorbitorilor, situl mai conţine şi instrumente de analiză a semnalului vocal.

5. Metode suplimentare de analiză: măsurări accelerometrice

Printre elementele specifice introduse de grupul nostru pentru analiza procesului vorbirii, se afla metoda accelerometrică. Metoda constă în determinarea acceleraţiilor mandibulei în timpul vorbirii şi corelarea mişcărilor cu tipul sunetelor pronunţate, cu duratele sunetelor şi cu energia sonoră. Deşi gradul de corelare constatat până în prezent este relativ redus, sperăm că metoda poate ajuta la segmentarea automată a semnalului vocal şi la evidenţierea unor corelaţii între caracteristicile pronunţiei cu mişcările

fonatorii. Fără să fie absolut nouă (există un număr mic de lucrări cu abordări oarecum similare, dar nu în scopul segmentării), abordarea sperăm să aducă elemente suplimentare în explicarea proceselor vorbirii. Rezultatele preliminare obținute (Teodorescu 2006a, b) evidențiază caracteristici ale tranzițiilor sunet nazal - vocală, *o-a* în diftongul *oa*, precum şi consoană plozivă-vocală.

6. Concluzii şi direcții viitoare

Considerăm că resursa este utilă în prezent ca mijloc educațional, iar în viitorul apropiat ca suport în cercetarea lingvistică şi în realizarea de aplicații informatice (sinteză şi recunoaştere). Credem că resursa impune şi noi standarde de calitate în realizarea unor resurse similare.

Arhiva sunetelor limbii române va fi dezvoltată prin adăugarea de noi înregistrări şi de prelucrări statistice ale sunetelor. Ulterior, se va urmări adăugarea unor înregistrări cu uşoare patologii, cum ar fi tremurul vocii (de natura emoțională sau patologică), adnotate şi prelucrate (Teodorescu et al., 2006c).

Referințe bibliografice

Teodorescu H.N. (2006a). Gnatofonia şi Gnatosonia. Analiza semnalelor vocale, Capitolul 2, *Ed. Performatica*, Iaşi, România, pag. 29-40.

Teodorescu H.N. (2006b). Gnatophonetics – A New Discipline Analyzing Relations between Speech and the Stomato-Gnathic System. *Zilele Academice Iesene*, *Simp Inventica. Simpozionul național "Bazele performanței şi inventică"* organizat în cadrul "Zilelor Academice Ieşene" ISBN 973-730-244-3, 978-973-730-244-1, 9 September 2006.

Teodorescu H.N., Zbancioc M., Mihăilescu E. (2006a), Speech Technology and Bio-Medical Engineering Teaching Based on the Web – A New Tool and Case Study. *Conference ICL 2006*, Villach, 27 -29 September 2006, Proceedings CD 2005 Ambient and Mobile Learning, Kasset University Press, Editors Michael Auer, Ursula Auer and R. Mittermeir, ISBN 3-89958-136-9.

Teodorescu H.N., Tandabăț D., Feraru M., Zbancioc M., Luca R.(2006b). A corpus of the sounds in the Romanian spoken language for language-related education. *International Conference on Human and Material Resources in Foreign Language Learning – RFLL 2006*, Murcia, Spania, 12-14 iulie 2006.

Teodorescu H.N., Feraru M., Tandabat D. (2006c), Nonlinear Assessment of Professional Voice 'Pleasantness', Conference *BIOSIGNAL 2006,* ISBN 80-214-3152-0, Brno, 28-30 June 2006, pag. 63-66.

Voiced Sounds of Romanian Language Project. [http://iit.iit.tuiasi.ro/ romanain_spoken_language/index.htm].

SCHEMĂ XML DE ADNOTARE A INTONAȚIEI ÎN CADRUL CORPUSURILOR DE TEXT

VASILE APOPEI, DOINA JITCĂ

Institutul de Informatică Teoretică, Academia Română, Filiala Iaşi

{vapopei, jdoina}@ iit.tuiasi.ro

Rezumat

În lucrarea se propune o schemă de adnotare a unui corpus de text în format XML, cu informație prozodică rezultată din analiza rostirilor respectivului text. În secțiunea 2 se prezintă ierarhia unităților intonaționale pe care s-a bazat structurarea textului şi setul de etichete folosit în adnotarea evenimentelor tonale din conturul F0. În secțiunea 3 se prezintă schema XML de adnotare a intonației în cadrul unui corpus de text, prin prezentarea tagurilor şi a atributelor acestora. În secțiunea 4 se prezintă un exemplu ce ilustrează corespondența între evenimentele de intonație marcate pe conturul F0 şi structurarea XML a aceleaşi informații.

1. Introducere

Lucrarea prezintă o modalitate de introducere a nivelului intonațional în adnotarea XML a corpusurile de text, structurate deja la nivel morfologic şi sintactic. O structurare multinivel a unui corpus de text este realizată în cadrul proiectului MULI la care adnotarea s-a efectuat pe 3 nivele: sintactic, prozodic, discurs, în vederea intercorelării trăsăturilor corespunzătoare acestora (Baumann et al. 2004). În adnotarea intonației autorii au avut în vedere trei unități de structurare: unitate intonațională, unitate intonațională intermediară şi cuvântul, care se regăsesc şi în adnotarea noastră. Evenimentele intonaționale din cadrul acestora, la fel ca şi în aplicația noastră, s-au adnotat folosind o variantă a sistemului de etichete ToBI (GToBI).

Analiza intonației asupra corpusului de voce s-a efectuat din perspectiva modelului autosegmental-metric concretizat într-o structură ierarhică a unităților intonaționale (prezentată în secțiunea 2) stabilită pe baza structurilor folosite de autori precum: Selkirk (1984), şi Di Cristo (2004).

Pentru marcarea evenimentelor intonaționale, de pe conturul frecvenței fundamentale F0, din cadrul unităților intonaționale, am folosit etichete ale sistemului ToBI de adnotare a intonației: (Beckman, Ayers, 1997).

În afară evenimentelor luate în considerație de sistemul ToBI s-au marcat şi alte tonuri semnificative din conturul F0 folosind etichetele H+ şi L+. Aceste tonuri se pot afla fie pe silaba anterioară unei silabe accentuate, fie pe silaba următoare.

Proiectarea schemei XML de structurare a unui text din punct de vedere al intonației s-a bazat pe analize anterioare referitoare la structurarea diferitelor rostiri în unități

intonaționale, în diverse contexte de accentuare a cuvintelor, atât în propoziții afirmative cât şi interogative (Apopei et al., 2006).

Textul folosit în realizarea corpusului a fost extras din romanul "1984" al autorului George Orwell şi rostit de doi vorbitori. Adnotarea rostirilor s-a realizat cu un program dezvoltat la Institutul de Informatică Teoretică Iaşi iar informația rezultată a fost apoi convertită în format XML, folosind schema de adnotare prezentată în secțiunea 2.

2. Prezentarea ierarhiei unităților intonaționale

Adnotarea prozodică pe care o propunem se bazează pe ierarhia unităților intonaționale prezentată în figura 1. În cadrul acestei ierarhii, cea mai mică unitate căreia i se poate asocia un eveniment din conturul frecvenței F0 este silaba. Silabele constituie părți componente ale cuvintelor. Cuvintele sunt purtătoare ale accentelor sintactice sau gramaticale. În general, unitățile de accentuare cuprind un cuvânt cu accent şi unul sau mai multe cuvinte clitice. Există situații în care unitățile de accentuare pot include pe lângă cuvântul accentuat, un alt cuvânt neclitic, dar care şi-a pierdut complet accentul în vecinătatea acestuia.

O unitate de accentuare purtătoare de accent puternic se grupează în cadrul acestei ierarhii cu alte unități care includ cuvinte purtătoare de accente mai slabe, formând unități ritmice (Di Cristo, 2004). Grupările cuvintelor realizate de unitățile ritmice corespund la nivel semantic, sintagmelor.

Una sau mai multe unități ritmice compun o frază intonațională (intonational phrase, in limba engleză, şi notată, IP) sau o frază intonațională intermediară (intermediate phrase, în limba engleză şi notată, ip).

Diferența dintre unitățile IP şi cele notate ip o constituie durata tonurilor finale şi a pauzei de după acestea. În delimitarea unităților IP/ip se ține cont de modificările tendințelor în evoluția tonurilor țintă asociate evenimentelor din cadrul conturului frecvența F0. Tendințele pot fi descrescătoare (downsteping) sau crescătoare (upsteping). De exemplu, când frecvența F0 părăseşte tendința de coborâre începută, înainte de a atinge un final al frazei intonaționale se consideră că s-a încheiat o frază intermediară .

Evenimentele intonaționale din conturul frecvenței F0 avute în vedere în adnotare sunt următoarele: accentele de pitch, produse pe durata silabelor accentuate (în engleză „Pitch Accent"); tonurile de sfârşit ale frazelor intonaționale intermediare; tonurile graniță ale frazelor intonaționale; alte tonuri semnificative din conturul F0 (în engleză „Target Ton"), care se pot afla fie pe silaba anterioară silabei accentuate, fie pe silaba următoare. Pentru marcarea primelor trei tipuri de evenimente s-au folosit etichetele sistemului de adnotare ToBI, iar pentru ultima categorie s-au adăugat două etichete, H+ şi L+, care au fost folosite şi în alte aplicații de adnotare prozodică (Baumann et al., 2004).

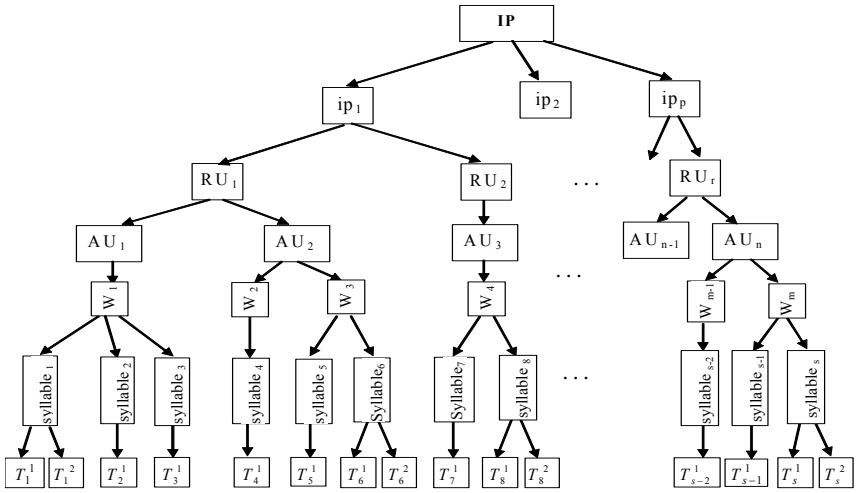


Figura 1: Ierarhia structurii intonaționale

3. Schema XML de adnotare a intonației

În stabilirea tag-urilor pentru adnotarea intonației în format XML s-au avut în vedere toate unitățile din ierarhia prezentată în figura 1, creând câte un tag pentru marcarea unităților de pe fiecare nivel.

Atributele tag-urilor conțin pe lângă secvențe de etichete referitoare la tipul de evenimente şi informații cantitative legate de nivelul tonurilor țintă asociate acestora. În acest scop s-a împărțit gama de variație a frecvenței F0, în cadrul rostirii de adnotat, în semitonuri şi s-a realizat o scală de măsurare a tonurilor cu baza la nivelul tonului celui mai înalt. Astfel atributele de tip *ToneValue*s se exprimă prin numere întregi cuprinse între 0 şi 20 corespunzătoare celui mai înalt nivel de ton şi respectiv, celui mai scăzut din cadrul rostirii.

Tag-urile şi atributele corespunzătoare sunt prezentate în tabelul 1. Unele atribute reprezintă etichete de ton, cum ar fi: cele de început şi respectiv, sfârşit ale unităților IP (*BeginToneLabel*, *BoundaryToneLabel*), cele de sfârşit ale unităților intermediare (*PhraseToneLabel*) sau cele legate de accentele cuvintelor şi a tonurilor adiacente (*TonalGroupLabel*- când sunt marcate la nivelul cuvintelor şi *Accent, TargetTone* –când sunt marcate la nivelul silabelor).

Secvența de numere întregi, separate prin virgulă, ce constituie valori pentru atributele de tip ToneValues corespunde tonurilor marcate prin etichete în cadrul atributelor aceloraşi unități. Sincronizarea secvenței de etichete stabilită prin atributul *TonalGroupLabe*l cu secvența de silabe a cuvântului, se realizează prin asocierea etichetei de accent de pitch cu silaba accentuată. În cazul cuvintelor compuse cu mai multe accente se impune marcarea separată a fiecărei silabe din cadrul acestora, cu tag-ul <Sillable>. În cadrul unui astfel de cuvânt vor exista silabe marcate cu atributele *Accent* şi *ToneValues*, silabe marcate cu atributele *TargetTone* şi *ToneValues*, şi silabe fără nici o indicație de ton. Oricare silabă poate fi caracterizată prin nivelul de energie, sau printr-o măsură a duratei acesteia folosind atributele *Energy* şi respectiv, *Length*.

Tag-ul <RU> nu are atribute pentru că unităților ritmice nu conțin tonuri semnificative diferite de cele ale subunităților componente (unități de accentuare, cuvinte, silabe).

Tabel 1: Atributele şi valorile tag-urilor de adnotare prozodică

Tag	Atribut	Valoare	Unitate intonațională
<IP>	*BeginToneLabel*	*%L, %H*	frază intonațională
	BoundaryToneLabel	*L%, H%*	
	ToneValues	*Sir de numere întregi*	
<ip>	*PhraseToneLabel*	*L-, H-*	frază intonaționale intermediare
	PhraseToneValue	*Număr întreg*	
<RU>			unitate ritmică
<AU>	*Break*	*No, Short, Large*	unitate de accentuare
	PunctSign	*/, /:/ ; /./ !/ ?/*	
<W>	*TonalGroupLabel*	*Secvențe de etichete*	Cuvânt
	ToneValues	*Sir de numere intregi*	
	ID_W	*Numeric*	
<Sillable>	*Accent*	*H*, L*, L+H*, H+!H* , H+L**	Silabă
	TargetTone	*H+, L+*	
	ToneValues	*Sir de numere intregi.*	
	Length	*Small, Medium, Large*	
	Energy	*Small, Medium, Large*	

Atributul "*Break*" indică absența sau prezența pauzei după unitatea de accentuare. Prezența pauzei este marcată în termenii *Short* sau *Large* după cum aceasta este de durată mai scurtă sau mai lungă. Valoarea implicită este *No*, adică absența pauzei.

Atributul "*PunctSign*" indică prezența unui semn de punctuație în text după ultimul cuvânt din unitatea de accentuare.

Atributul ID_W al tag-ului <W> poate fi folosit pentru a face legătura cu alte tipuri de structuri ale aceluiaşi text cum ar fi cele care marchează categoriile morfologice, sintactice sau de discurs. Folosind Tag-urile prezentate în acest paragraf, se poate marca un text cu informație relativă la intonația unei rostiri a acestuia.

4. Exemplu de adnotare a intonației în format XML

Exemplificarea adnotării intonației se va face prin corelarea conturului F0 corespunzător rostirii textului „*Avem de discutat lucruri serioase, zece minute nu-i nevoie să mai faci pe valetul*" (figura 2) cu marcarea textului folosind categoriile XML prezentate în secțiunea 3 (figura 3). Intonația este formată din două fraze intonaționale. În cadrul primei unități *IP* se formează două unități ritmice iar în cadrul celei de a doua unității *IP* se formează două fraze intermediare *ip* ce cuprind cele trei unități ritmice.

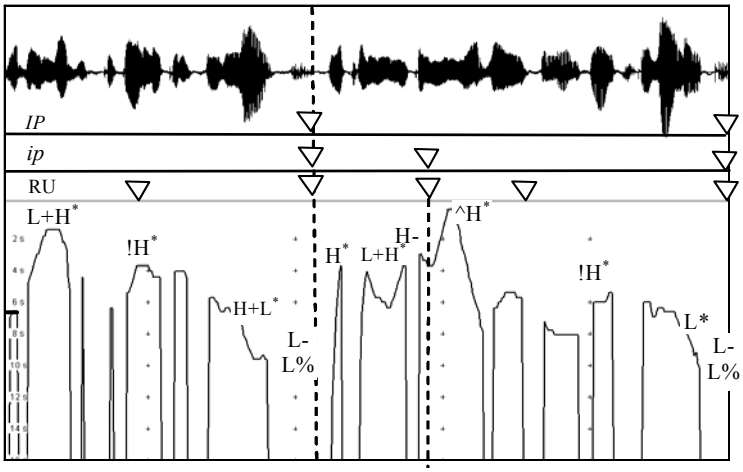


Figura 2. Unda vocală şi conturul F0 al rostirii
"*Avem de discutat lucruri serioase, zece minute nu-i nevoie să mai faci pe valetul* "

Delimitările realizate pe corpusul de voce, corespunzătoare unităților intonaționale ale rostirilor, au fost aplicate textului într-un fişier în format XML generând structurarea acestuia din punct de vedere intonațional.

```
<IP BeginTonLabel="%L" BoundaryTonLabel="L%">
      <RU>
        <AU>
          <W TonalGroupLabel="L+H*" ToneValues="4,1">Avem</W>
        </AU>
        <AU>
          <W>de </W>
          <W TonalGroupLabel="!H*" ToneValues="3">discutat</W>
        </AU>
      </RU>
      <RU>
        <AU>
          <W>lucruri</W>
        </AU>
        <AU Break=”Short” PunctSign=”,”>
          <W TonalGroupLabel="H+L*" ToneValues="8,10">serioase</W>
        </AU>
      </RU>
</IP>
<IP BeginTonLabel="%L" BoundaryTonLabel="L%">
  <ip PhraseTonLabel="H-">
      <RU>
        <AU>
          <W TonalGroupLabel="H*" ToneValues="3">zece</W>
        </AU>
        <AU>
          <W TonalGroupLabel="L+H*" ToneValues="6,3">minute</W>
        </AU>
      </RU>
  </ip>
  <ip PhraseTonLabel="L-">
      <RU>
        <AU>
          <W TonalGroupLabel="^H*" ToneValues="0">nu-i</W>
        </AU>
        <AU>
          <W>nevoie</W>
        </AU>
      </RU>
      <RU>
        <AU>
          <W>să</W>
          <W>mai</W>
          <W TonalGroupLabel="!H*" ToneValues="6">faci</W>
```

```
        </AU>
        <AU Break =”Large” PunctSign=”.”>
          <W>pe</W>
          <W TonalGroupLabel="L*" ToneValues="9">valetul</W>
        </AU>
      </RU>
  </ip>
</IP>
```

Figura 3. Adnotarea intonației rostirii textului
„*Avem de discutat lucruri serioase, zece minute nu-i nevoie să mai faci pe valetul*”

5. Concluzii

Schema de adnotare prezentată în această lucrare, a fost dezvoltată în scopul de a realiza un corpus de text în limba română cu informație relativă la intonația rezultată din rostirea acestuia. Alegerea unor fragmente din romanul “1984” al autorului George Orwell este justificată de existența unor adnotări la nivel morfologic şi sintactic pentru acest text. Folosind identificatorul de cuvânt <ID> aceste resurse pot fi alinate şi pe baza lor se pot îmbunătăți sau dezvolta noi aplicații lingvistice pentru limba română.

Corpusurile de text adnotate la nivel intonațional sunt utile în deducerea de reguli ce vizează implementarea intonației în sinteza vocală în limba română.

Referințe bibliografice

Albert Di Cristo (2004). La prosodie au carrefour de la phonétique, de la Phonologie et de l’articulation formes-functions, *Travaux Interdisciplinaires du Laboratoire Parole et Langage*,vol. 23, p. 67-211

Beckman M., Ayers G. (1997). *Guidelines for ToBI Labelling* (version 3), [http://www.ling.ohio-state.edu/~tobi/ame_tobi/labelling_guide_v3.pdf].

Baumann S, Brinckmann C., et all (2004). Multi-dimensional annotation of linguistic corpora for investigating information structure, *In: Proceedings Grontiers in Corpus Annotation Workshop at HLT/NAACL,* Boston, USA, p. 39-46.

Apopei V, Jitcă D, Turculeț A. (2006). Intonational structures in Romanian Yes-No Questions, *Jurnal of Compuer Science of Moldavia*, pp. 113-137, vol 14, no 1(40), Chişinău

Selkirk, E.O. (1984). *Phonology and Syntax: The Relation between Sound and Structure*. Cambridge, MA: MIT Press.

Capitolul 2

Dicționare şi corpusuri adnotate pentru prelucrarea textelor

NOI DEZVOLTĂRI ALE WORDNET-ULUI ROMÂNESC

DAN TUFIŞ, VERGINICA BARBU MITITELU, ALEXANDRU CEAUŞU, LUIGI BOZIANU, CĂTĂLIN MIHĂILĂ, MARGARETA MANU MAGDA

Institutul de Cercetări pentru Inteligență Artificială, Academia Română, Bucureşti

{tufis, vergi, alceausu, bozi, cata}@racai.ro; mmagda@csb.ro

Rezumat

Ontologiile lexicale de tip wordnet sunt dintre cele mai importante resurse lexicale folosite în aplicațiile de prelucrare a limbajului natural. O astfel de ontologie este diponibilă şi pentru limba română. Lucrarea descrie o parte din activitățile de îmbunătățire cantitativă şi calitativă a wordnet-ului românesc.

1. Dezvoltarea wordnet-ului românesc

Una dintre cele mai importante resurse lingvistice computaționale pentru limba română este, fără îndoială, ontologia lexicală de tip WordNet (Fellbaum 1998), a cărei dezvoltare a început, în colaborare cu Facultatea de Informatică a Universității Al. I Cuza din Iaşi, în anul 2001, în cadrul proiectului BalkaNet[1]. Dezvoltarea wordnet-ului românesc a continuat la Institutul de Cercetări pentru Inteligență Artificială şi după anul 2004 (Tufiş et al., 2006), când proiectul BalkaNet s-a încheiat.

Experiența acumulată în timpul proiectului BalkaNet, precum şi o serie de noi instrumente de achiziție lexicală au făcut ca productivitatea echipei de lingvişti să crească substanțial, astfel încât, actualmente, wordnet-ul românesc are 33422 de sinseturi (i.e. serii sinonimice), conținând un număr de 31246 de literali unici.

Tabel 1: Date statistice despre wordnet-ul românesc

33421 serii sinonimice (1289 nelexicalizate) • 24640 substantive • 7096 verbe • 851 adjective • 834 adverbe 53160 literali (31246 literali unici) 163 domenii 1773 categorii SUMO/MILO	Relații semantice: • hypernym 32041 • holo_part 2096 • holo_member 1029 • holo_portion 199 • category_domain 1861 • also_see 508 • similar_to 899 Relații lexicale: • near_antonym 1976 • be_in_state 566 • verb_group 1196 • causes 148 • subevent 264

Principala strategie de dezvoltare a wordnet-ului românesc constă în implementarea în limba română a seriilor sinonimice din wordnet-ul englezesc. Pentru selectarea

[1] http://www.ceid.upatras.gr/Balkanet/

sinseturilor ce urmau a fi implementate au fost urmărite criteriile ce ţin de acoperirea cu măcar un sens pe literal a cuvintelor din corpusul Acquis-ului comunitar. Ca modalităţi de stabilire a relevanţei unui sinset au fost considerate: (i) domeniul şi caracterizarea SUMO a seriei sinonimice (fiind avantajate domenii ca politic, legislativ etc., specifice Acquis-ului); (ii) numărul de apariţii ale literalilor în corpus; iar în cazul literalilor monosemantici, (iii) rangul de ocurenţă întors de motorul de căutare Google.

Folosind metoda de dezambiguizare automată descrisă în (Ion, Tufiş, 2004), dar, de data aceasta, pe un bitext diferit (SemCor), s-au depistat circa 7.000 de literali absenţi din sinseturile deja implementate. Actualizarea acestor sinseturi incomplete este în curs de realizare (până la data scrierii acestui articol au fost completate 600 de sinseturi), în paralel cu adăugarea de noi sinseturi.

2. Alinierea wordnet-ului românesc la versiunea PWN 2.1

În timpul scurs de la încheierea proiectului BalkaNet, cercetătorii de la Princeton au dezvoltat versiunea 2.1 a wordnet-ului american, PWN, care a adus o serie de modificări majore faţă de versiunea precedentă. Deşi numărul de sinseturi nu a crescut substanţial, a apărut distincţia între instanţă şi clasă, iar mai multe relaţii au fost redenumite sau rafinate, ceea ce a condus la necesitatea regenerării formatului XML pentru PWN folosit de VisDic (Horak, Smrz, 2004) în cadrul proiectului BalkaNet.

După cum se ştie, wordnet-ul românesc este aliniat cu Princeton WordNet (PWN), cu ontologia SUMO/MILO (Niles, Peace, 2003) precum şi cu taxonomia DOMAINS (Bentivogli et al., 2004). Au fost corectate mai multe erori de mapare între PWN2.0 şi SUMO/MILO şi DOMAINS.

Ontologia WN DOMAINS 2.0 are, faţă de versiunea 1.0, noi etichete de domeniu, iar unele dintre domenii au suferit o schimbare de granularitate. Aceste modificări au fost operate pentru a asigura o distribuţie uniformă a etichetelor de domeniu pentru seriile sinonimice ale wordnet-ului englezesc.

2.1. Generarea în format XML a Princeton WordNet versiunea 2.1

Faţă de versiunea XML a PWN2.0, versiunea 2.1 include şi numărul de ocurenţe ale sensurilor adnotate în corpusul SemCor. În plus, la nivelul relaţiilor lexicale au fost specificaţi literalii între care acestea se stabilesc. Deşi specificarea acestor literali s-a făcut încă de la versiunea 1.5 a Princeton WordNet, ea lipsea din versiunea XML a PWN 2.0.

Interfaţa de vizualizare/editare VisDic (Horak, Smrz, 2004), interfaţă dezvoltată în cadrul proiectului EuroWordNet[2], foloseşte un format XML propriu de reprezentare a wordnet-ului. Versiunea XML a PWN 2.1 are în plus tagurile <FREQ>, <SRCL> şi <TRGL> pentru reprezentarea numărului de ocurenţe din SemCor şi pentru specificarea literalilor între care se stabilesc relaţiile lexicale:

[2] http://www.illc.uva.nl/EuroWordNet/

```
<SYNSET>
  <ID>ENG21-06551177-n</ID>
  <POS>n</POS>
  <SYNONYM>
    <LITERAL>wordnet<SENSE>1</SENSE><FREQ>1</FREQ></LITERAL>
  </SYNONYM>
  <DEF>any of the machine-readable lexical databases modeled
  after the Princeton WordNet</DEF>
  <ILR>ENG21-06550617-n<TYPE>hypernym</TYPE></ILR>
  <DOMAIN>computer_science</DOMAIN>
  <SUMO>ContentBearingObject<TYPE>+</TYPE></SUMO>
</SYNSET>
```

Tabel 2:Date statistice despre Princeton WordNet 2.1

117597 serii sinonimice • 81426 substantive • 13650 verbe • 18877 adjective • 3644 adverbe	155327 literali • 117097 substantive • 11488 verbe • 22141 adjective • 4601 adverbe
Relaţii semantice: • hypernym 88258 • instance 8515 • holo_part 8874 • region_domain 1327 • usage_domain 1258 • category_domain 6534 • holo_portion 793 • holo_member 12262 • also_see 3272 • similar_to 22622	Relaţii lexicale: • derived 8065 • eng_derivative 71914 • near_antonym 8029 • verb_group 1748 • particle 124 • be_in_state 1286 • subevent 409 • causes 219

2.2. Alinierea wordnet-ului românesc la versiunea PWN 2.1

După generarea versiunii XML a PWN2.1, alinierea wordnetului românesc la noua versiune s-a realizat folosind ca resurse WN-Map (Daudé et al., 2000) şi maparea pentru substantive şi verbe disponibilă pe site-ul Princeton WordNet.

WN-Map foloseşte un algoritm iterativ pentru optimizarea unei funcţii bazate pe un set de criterii ce descriu un context local. Criteriile pot fi eticheta morfologică a sinsetului, definiţia sinsetului, locul pe care îl ocupă acesta în ierarhie etc. De asemenea, aplicaţia WN-Map mai foloseşte şi reguli care pot decide asupra compatibilităţii sau incompatibilităţii candidaţilor la aliniere, reguli bazate pe criteriile enunţate anterior.

În cazul sinseturilor pentru care ambiguitatea de mapare nu a putut fi rezolvată automat (destul de puţine, de altfel), dezambiguizarea s-a făcut manual, de către experţii lingvişti implicaţi în proiect.

Problemele de aliniere s-au datorat, în principal, modificărilor operate asupra sinseturilor adjectivale şi adverbiale. O altă problemă de aliniere a provenit din transferul conceptelor asignate indexului interlingual specific ţărilor balcanice (Balkanet Interlingual Index – BILI). Aceste concepte nu au echivalent în wordnet-ul englezesc, dar sunt integrate în ontologia acestuia.

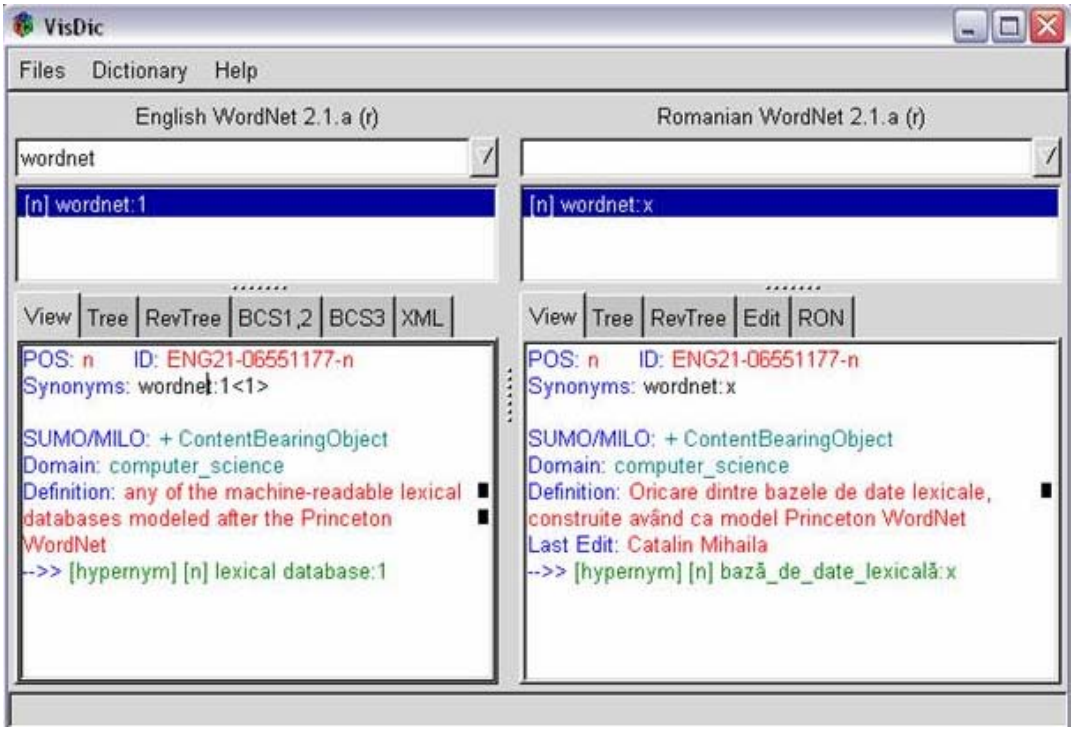


Figura 1: Exemplu de serie sinonimică din PWN 2.1 şi din wordnet-ul românesc, vizualizat în VisDic

3. Perspective

Corpusul JRC-Acquis ce cuprinde setul de legi, dispoziţii, tratate comune tutoror statelor membre ale Uniunii Europene este unul dintre cele mai mari corpusuri paralele disponibile la momentul actual. Mărimea corpusului şi numărul mare de limbi componente îl fac instrumentul perfect de validare a unei ontologii lexicale multiliguale. Astfel, strategia de implementare a seriilor sinonimice din limba engleză în limba română urmăreşte, în principal, acoperirea cu cel puţin un sens per literal a tuturor cuvintelor din JRC-Acquis.

Îmbunătăţirea la nivel cantitativ a wordnet-ului – mai multe serii sinonimice, mai mulţi literali echivalaţi – este de maximă importanţă pentru aplicaţiile de traducere automată din şi în limba română. Nu a fost, însă, ignorată îmbunătăţirea calitativă a wordnet-ului românesc (au fost rezolvate conflictele de asignare a sensurilor, au fost adăugaţi literalii absenţi din unele serii sinonimice, identificaţi cu ajutorul unui algoritm aplicat asupra unui corpus paralel englez-român, dezamabiguizat la nivel semantic).

De asemenea, folosind corpusuri paralele, sunt în curs de implementare sisteme pentru importul automat de sinseturi conținând instanțe şi, respectiv, substantive monosemantice.

Mulțumiri. Autorii sunt recunoscători finanțatorilor proiectului BalkaNet (Comisia Europeană şi Ministerul Educației şi Cercetării), în cadrul căruia a debutat dezvoltarea wordnet-ului românesc, ai proiectului ROTEL (Ministerul Educației şi Cercetării), în cadrul căruia continuă dezvoltarea cantitativă şi calitativă a wordnet-ului românesc, precum şi Academiei Române, pentru finanțarea temelor de plan în cadrul cărora s-a lucrat şi la sporirea calității acestei resurse lingvistice pentru limba română.

Referințe bibliografice

Bentivogli, L., Forner, P., Magnini, B., Pianta, E. (2004). Revising WordNet Domains Hierarchy: Semantics, Coverage, and Balancing. *Proceedings of COLING 2004 Workshop on "Multilingual Linguistic Resources"*, Geneva, Switzerland, 101-108.

Daudé, J., Padró, L., Rigau, G. (2000). Mapping WordNets Using Structural Information. *38th Annual Meeting of the Association for Computational Linguistics* (ACL'2000), Hong Kong, 504-511.

Fellbaum, Ch. (ed.) (1998). *WordNet: AN Electronic Lexical Database*, MIT Press.

Horak, A., Smrz, P. (2004). New Features of Wordnet Editor VisDic. *Romanian Journal of Information Science and Technology*, volume 7, Numbers 1-2, 1-13.

Niles, I., Pease, A. (2003). Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. *Proceedings of the 2003 International Conference on Information and Knowledge Engineering* (IKE '03), Las Vegas, Nevada, June 23-26.

Ion, R., Tufiş, D. (2004). Multilingual Word Sense Disambiguation Using Aligned WordNets. *Romanian Journal of Information Science and Technology*, volume 7, Numbers 1-2, 183-200.

Tufiş, D., Barbu Mititelu, V., Bozianu, L., Mihăilă, C. (2006). Romanian WordNet: New Developments and Applications. *Proceedings of the 3rd Conference of the Global WordNet Association*, Seogwipo, Jeju, Republic of Korea, January 22-26, 337-344

FRAMENET ROMÂNESC: TENTATIVĂ DE ELABORARE

VICTORIA BOBICEV, VICTORIA MAXIM, TATIANA ZIDRAŞCO, ALINA IACIURINSCHI

Universitatea Tehnică din Moldova

vika@rol.md, maxivica@yahoo.com, tzidrashco@mail.md, ialinca@yahoo.com

Rezumat

În lucrarea de faţă sunt prezentate rezultatele lucrului efectuat la crearea variantei româneşti a resurselor multilingve 'Romance FrameNet'. Au fost evidenţiate unele întrebări şi probleme nerezolvate. În concluzie sunt menţionate unele sugestii privind metodologia creării FrameNet-ului multilingv.

1. Introducere

FrameNet (Johnson et al., 2003) este un proiect de cercetare lexicografică iniţiat în cadrul International Computer Science Institute Berkeley, California[1]. Elementele constructive ale bazei de date FrameNet sunt propoziţiile marcate atât semantic, cât şi sintactic, fapt ce permite extragerea automată a cadrelor sintactico-semantice. Aceste cadre vizualizează legătura dintre sens şi structura sintactică prin care acesta este redat. Ca bază pentru marcarea semantică sunt utilizate cadrele semantice (*frames*) – structuri conceptuale ce reprezintă evenimente, obiecte, proprietăţi. Fiecare cadru este dotat cu un set de elemente semantice (*frame elements*) ce caracterizează cadrul dat. Cu fiecare cadru este asociat un număr de cuvinte – unităţi ale lexiconului (*lexicon units*) care evocă sensul reprezentat de cadrul dat.

În legătură cu creşterea interesului în reprezentarea sensului folosind semantica cadrelor (Gildea & Jurafsky, 2002), FrameNet a fost folosit în exerciţiul SensEval-3 (Litkowski, 2004). Sarcina propusă participanţilor a fost marcarea automată a rolurilor semantice într-un set de propoziţii de testare având un subcorpus de propoziţii marcate ale Framenet-ului pentru antrenare. Rezultatele celor 20 sisteme care au participat în exerciţiu sunt destul de bune (precizia medie 80%), ceea ce caracterizează FrameNet-ul ca o resursă semantică excelentă. Astfel, eforturile depuse în scopul creării acestei resurse au un rezultat de mare valoare.

2. FrameNet în alte limbi

Utilizând principii similare cu Framenet-ul englezesc, s-au creat resurse lingvistice pentru alte limbi, şi anume:

- **German FrameNet** este realizat de trei echipe în colaborare:

[1] http://framenet.icsi.berkeley.edu/

  - Scopul proiectului **SALSA (**Erk et al., 2003) este crearea lexiconului limbii germane cu informaţia semantică şi sintactică bazată pe teoria cadrelor semantice şi analiza posibilităţilor de a folosi lexiconul dat pentru procesarea limbii germane. Etapa precedentă **SALSA** I avea ca scop marcarea semantică a unui corpus în limba germană şi cercetarea metodelor de utilizarere a lui în procesarea textului.
  - Altă echipă (Boas, 2004) foloseşte baza de date FrameNet, înlocuind în întregime părţile englezeşti dependente de limbă cu părţile germane.
  - În Stuttgart un grup cercetează analiza şi extragerea datelor din corpusul adnotat, în special colocaţii şi nominalizări (*nominalizations*).
- **Spanish FrameNet (**Subirats & Petruck, 2003) este un proiect naţional la care participă un număr de cercetători din diferite universităţi guvernat de universitatea independentă din Barselona. FrameNet-ul spaniol se creează în colaborare cu echipa framenet-ului englezesc şi se bazează pe semantica cadrelor. Scopul proiectului este adnotarea semantică a propoziţiilor spaniole dintr-un corpus de limbă spaniolă.
- **Japanese FrameNet** – JFN (Ohara et al., 2004) are ca scop crearea unui lexicon care înregistrează descrierea valenţelor cuvintelor japoneze, bazată pe semantica cadrelor. Scopul final al JFN, creat în colaborare cu echipa FrameNet-ului englezesc, este crearea unei baze de date de tip FrameNet pentru cuvintele japoneze.
- **French FrameNet**[2] are ca scop crearea lexiconului semantic francez utilizând metodologia originală FrameNet. La fel este creat un corpus de propoziţii adnotate conform metodologiei FrameNet. În proiectul dat este studiată intens posibilitatea de utilizare repetată a resurselor semantice lexicale franceze. Numai în acest proiect din cele patru descrise se cercetează posibilitatea de traducere a propoziţiilor FrameNet-ului englezesc utilizând resurse valabile de traducere.

După cum se vede din descrierile date, numai în proiectul francez se cercetează posibilitatea de a crea o sursă într-adevăr bilingvă, care să conţină un set de propoziţii paralele adnotate folosind metodologia FrameNet. Astfel o sursă se creează în cadrul proiectului chinez (Fung & Chen, 2004), însă rezultatele încă nu au fost publicate.

3. Romance FrameNet

Cu scopul extinderii ariei de acoperire a proiectului FrameNet s-a propus iniţierea proiectului „Romance FRAMENET”[3], care reprezintă o resursă multilingvă pentru limbile romanice (franceză, spaniolă, italiană, română, portugheză, catalană) bazată pe cadrele semantice FrameNet (Lowe et al., 1997). Această iniţiativă are ca scop crearea unei resurse care ar reprezenta un corpus paralel, aliniat şi adnotat semantic.

Pentru crearea acestei resurse s-a propus o metodologie, asemănătoare metodologiei folosite în crearea lui MultiSemCor (Lupu et al., 2005) care implică câteva etape. Iniţial se traduc în mod manual propoziţiile din limba engleză în limbile sus-numite. Traducerea este efectuată de echipele din ţările corespunzătoare. Apoi în mod automat propoziţiile se aliniază la nivel de cuvânt şi adnotarea semantică se transferă din propoziţiile englezeşti la propoziţiile din altă limbă folosind alinierea efectuată.

[2] http://libresource.inria.fr/projects/framenet/
[3] http://www.icsi.berkeley.edu/~vincenzo/rfn/index.html

Rezultatul va fi o resursă multilingvă aliniată la nivel de cuvânt şi marcată semantic în baza teoriei cadrelor FrameNet-ului englezesc. O astfel de resursă va fi apreciată de cei care studiază limbile implicate, de lingviştii care cercetează diferenţele structurale dintre limbile acestea. În afară de aceasta, resursa dată poate fi folosită în traducerea automată şi în interpretarea semantică multilingvă. Însă crearea resursei date folosind metodologia propusă necesită un volum mare de adnotari manuale, fapt ce duce la încetinirea obţinerii rezultatelor.

4. FrameNet românesc

Echipa noastră participă în crearea părţii româneşti a resursei date. Pentru aprecierea metodologiei propuse iniţial au fost alese 100 de propoziţii din FrameNet-ul englezesc care s-au tradus în limba română şi s-au marcat cu rolurile semantice. Pe parcursul lucrului au fost observate următoarele probleme:

- pot exista mai multe variante de traducere;
- poate să nu fie găsită nici o variantă de traducere;
- calitatea transferului automat al marcajelor semantice nu este suficient de bună;
- diferenţele lexicale şi sintactice dintre limbi cauzează diferite probleme în transferarea marcării semantice.

La următoarea etapă studiul s-a efectuat asupra a 1000 de propoziţii, care au fost traduse de traducători profesionişti. Prin intermediul aliniatorului lexical dezvoltat de RACAI (Tufiş et al., 2006) s-a realizat alinierea automată la nivel de cuvânt. Şi în final marcarea semantică de la propoziţiile englezeşti s-a transferat automat la cele româneşti. Pentru transferarea marcării semantice, două persoane au creat în mod independent două programe de transferare (unul în Perl, altul în C#). Procesul de transferare a fost complicat din cauza semnelor diferite în propoziţiile date, şi anume apostrofuri, semne care reprezintă bani, semne de punctuaţie care nu coincideau şi altele. În multe cazuri au apărut probleme din cauza diferenţei între modul de marcare şi modul de aliniere. Marcarea fragmentelor de propoziţii este executată la nivel de caractere, alinierea este executată la nivel de cuvinte. În afară de aceasta, în cadrul alinierii sunt formate unităţi de traducere care conţin cîteva cuvinte unite cu cu semnul '_'. În unele cazuri unirea cuvintelor în propoziţia marcată şi cea aliniată a fost diferită ce complica procesul de transferare a rolurilor. Rezultatele transferării au fost validate manual. Au fost considerate corecte numai propoziţiile în care nu a fost găsită nici o greşeală în transferarea rolurilor semantice. Primul experiment, efectuat asupra unui număr de propoziţii mai simple, a arătat un rezultat de 59% corectitudine (total propoziţii transferate - 424, corecte - 252). Al doilea exepriment, care a implicat un număr mai mare de propoziţii, a dat rezultatul de 36% (total propoziţii transferate - 600, corecte - 219). O parte din propoziţiile acestea au fost analizate mai detaliat. În primul rînd, în 7% de propoziţii transferarea nu a fost efectuată corect. În restul propoziţiilor a fost controlată alinierea. În 9% din propoziţiile acestea nu era marcat corect cuvîntul de bază (*target*). În 31% a fost marcat greşit numai un cuvînt. În majoritatea cazurilor greşit sunt marcate prepoziţii, pronume sau articole care este problematic de aliniat corect. În multe cazuri un rol în propoziţia englezească conţinea o parte din propoziţie din trei-cinci şi mai multe cuvinte. În astfel de cazuri transferarea rolului pe cuvinte deseori îl întrerupea

şi o parte din cuvintele date nu au fost marcate. Uneori în fragmentul dat toate cuvintele erau marcate în afară de un articol sau un pronume. În 27% de propoziţii au fost observate mai multe erori. Majoritatea din erorile acestea au fost cauzate absenţa alinierii pentru cuvintele englezeşti marcate. Astfel, există necesitatea verificării rezultatului procesării automate.

În scopul creării variantei româneşti a resursei multilingve sus-numite la catedra noastră a fost efectuat un considerabil volum de lucru cu ajutorul studenţilor. Au fost selectate câteva cadre şi repartizate astfel, ca propoziţiile dintr-un cadru să fie traduse şi marcate de studenţii unui grup şi în final toate propoziţiile cadrului să fie traduse. În cele ce urmează sunt descrise datele referitoare la cadrele prelucrate.

Cadrul: **Removing;** unităţile lexicale cu numărul propoziţiilor adnotate: pluck – 38, prised – 3, evacuate – 6, evacuation – 25, purge – 28, remove – 30, extract – 8, oust – 13, expunge – 6, expulsion – 24, evict – 3, eviction – 25, excise – 2, elimination – 13, empty – 16, ejection – 18, eliminate – 21, drain – 2, eject – 19, clear – 15, confiscate – 15; în total – 330 propoziţii.
Cadrul: **Sensation;** unităţile lexicale cu numărul propoziţiilor adnotate: fragrance – 15, sight – 24, taste – 16, bouquet – 8, incense – 6, reek – 14, savour – 8, whiff – 12, scent – 24, sensation – 27, noise – 34, sense – 16, aroma – 6, odour – 16, perception – 18, stink – 13, feeling – 21, flavour – 30, perfume – 12, smell – 39, sound – 50; în total – 409.
Cadrul: **Commerce;** unităţile lexicale cu numărul propoziţiilor adnotate: buyer – 60, purchaser – 31, seller – 37, vendor – 35, retailer – 25; în total -188 propoziţii.
Cadrul: **Change-of-phase;** unităţi lexicale cu numărul propoziţiilor adnotate: condense – 13, thaw – 15, evaporate – 20, defrost – 5, solidify – 12, freeze – 21, vaporize – 5, melt – 27, sublime – 1; în total 119 propoziţii.
În total – 1047 propoziţii.
Apoi propoziţiile traduse s-au marcat în mod manual. Cu scopul facilităţii lucrului s-a creat un produs soft FRAME STUDIO 1.06 care pemite transferarea, marcarea şi corectarea rolurilor semantice. Un avantaj al acestei aplicaţii este o interfaţă foarte simplă în utilizare şi obţinerea rezultatului în format XML. Un exemplu al acestui rezultat este prezentat mai jos:

```
<?xml version="1.0" encoding="utf-16" ?>
   <frames>
<sentence> <frame>Hear</frame>
    <lexunit startpos="22" endpos="25">hear.v</lexunit>
    <markups>
      <markup felement="Hearer" startpos="27" endpos="28" />
      <markup felement="Message" startpos="30" endpos="36" />
    </markups>
    <text>Pat a spus că până să audă ea aceasta, ea nu realizase
faptul cât de anti-feminist este fratele ei.</text>
  </sentence>  . . .
```

În urma lucrului efectuat am constatat că cea mai complicată parte a lucrului este traducerea, care nu poate fi executată automat. Traducerea se complica din cauza faptului că propoziţiile sunt rupte din context şi în unele cazuri nu este clar sensul lor.

Exemplul 1: nu este clar dacă este vorba despre o persoană sau despre luna aprilie.

We love the April approach : seven good-value , gentle products with a luxurious apricot FRAGRANCE.
În unele cazuri nu este clar din ce tip de text propoziţiile sunt extrase. Însă este important de ştiut domeniul şi topica textului, uneori tonalitatea traducerii depinde de tipul textului. În exemplul 2 propoziţia poate fi tradusă în diferite modalităţi:

REMOVE the dumplings with a slotted spoon and serve them the same way as the fried dumplings.

Scoate gogoaşele ...Scoateţi gogoaşele ...Gogoaşele se scot ...

După cum am menţionat deja, metodologia care include traducerea a fost adaptată în baza creării resursei traduse MultiSemCor. Însă diferenţa constă în faptul că MultiSemCor conţinea texte întregi ce permitea înţelegerea lor mai amplă. Sensul fragmentului alcătuit din propoziţii legate este mai clar decât sensul unei propoziţii înafară contextului. Totuşi, propoziţiile din FrameNet au fost scoase din text şi iniţial au fost scrise fără o intenţie să fie percepute înafară contextului lor. Chiar şi traducerea pronumelui „it" creează probleme, fiindcă se referă la ceva în afară propoziţiei date. Astfel, traducerea propoziţiilor devine mai complicată decât traducerea textelor şi este nevoie să fie executată de traducători cu experienţă şi cunoştinţe profunde în limba engleză. După cum am menţionat, nici un proiect de creare a FrameNet-ului descris mai sus nu include traducerea propoziţiilor englezeşti. În cazuri când posibilitatea aceasta este menţionată (BiFrameNet şi FrameNet franţuzesc) rezultatele nu au fost anunţate.

O problemă nerezolvată rămâne întrebarea dacă traducătorii trebuie să fie informaţi despre scopul traducerii sau nu. O condiţie de bază pentru propoziţiile traduse este că propoziţiile trebuie să fie corect alcătuite în limba română. Însă în multe cazuri traducătorii interpretează propoziţiile destul de liber, reformulând sensul redat de propoziţia englezească, şi atunci deseori în traducere nu rămân elementele necesare pentru adnotare. Totuşi, este posibil de tradus în română făcând cât mai mult posibilă similaritatea cu propoziţia englezească.

La catedra noastră traducerile au fost efectuate de către studenţi şi verificate de profesori. Şi studenţii şi profesorii cunoşteau scopul traducerii. Pe parcursul lucrului am observat că există un numar de propoziţii prea complicate. Una din ele este prezentată în exemplul 1. Sensul propoziţiilor acestea poate fi interpretat greşit din cauza lipsei contextului. După părerea noastră, astfel de propoziţii nu trebuiesc traduse. În schimb, pentru fiecare unitate lexicală a fost ales un număr (10-20) de propoziţii româneşti, ce conţin unitatea lexicală respectivă şi au fost marcate cu elementele cadrului respectiv. Nedispunând de un corpus reprezentativ, noi am folosit Internetul pentru obţinerea acestor propoziţii.

5. Concluzii şi discuţii

În urma lucrului efectuat am ajuns la unele concluzii şi întrebări nerezolvate. După părerea noastră, în metodologia adoptată, cea mai complicată parte a lucrului este traducerea. Pentru ameliorarea calităţii rezultatelor obţinute ar fi bine ca traducerea să fie executată de cel puţin trei traducători în mod separat. Problema selectării variantei de traducere mai convenabile pentru marcare rămâne deschisă. E posibil ca această alegere să fie efectuată de către adnotator. O problemă nerezolvată rămâne întrebarea dacă

traducătorii trebuie să fie informaţi despre scopul traducerii sau nu. Propoziţiile pentru care traducerile au fost prea diferite sau neclare trebuiesc exluse din varianta românească. Propoziţiile problematice nu trebuie traduse, pot fi traduceri incorecte. Dacă dorim să creăm o sursă calitativă, trebuie să tratăm atent fiecare propoziţie. În schimb, pentru fiecare unitate lexicală ar trebui de ales un număr de propoziţii româneşti, ce conţin unitatea lexicală respectivă, cu scopul marcării conform metodei FrameNet.

Referinţe bibliografice

Boas, H. C. (2004) Semantic Frames as an Interlanguage for Multilingual Lexical Databases, *First Global FrameNet Workshop*, ISI, Berkeley, California.

Erk, K., Kowalski, A., Pado Sebastian and Pinkal Manfred. (2003) Towards a Resource for Lexical Semantics: A Large German Corpus with Extensive Semantic Annotation. *Proceedings of ACL 2003*, Sapporo.

Fung P., Benfeng C. (2004) "BiFrameNet: bilingual Frame Semantics Resource Construction by Cross-lingual Induction", in *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, August 2004.

Gildea, D., and Daniel Jurafsky. (2002) Automatic Labeling of Semantic Roles. *Computational Linguistics*, 28 (3), 245-288.

Johnson, C., Miriam Petruck, Collin Baker, Michael Ellsworth, Josef Ruppenhofer, and Charles Fillmore, (2003). FrameNet: Theory and Practice. Berkeley, California.

Litkowski, K. C. (2004) Senseval-3 Task: Automatic Labeling of Semantic Roles, Proceedings of Senseval-3: The *Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, ACL 2004*, Barcelona, pp. 9-12.

Lowe, J.B., Baker, C.F. and Fillmore, C.J. (1997): A frame-semantic approach to semantic annotation, In *Proceedings of the SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?,* Washington, USA.

Lupu M., Trandabăţ .D., Husarciuc M. (2005) A Romanian SemCor Aligned to the English and Italian MultiSemCor, *1st International ROMANCE FrameNet Workshop*, în cadrul şcolii de vară EUROLAN 2005, România, pp. 20-27.

Ohara, K. H., Fujii S., Ohori T., Suzuki R., Saito H., Ishizaki S. (2004). "The Japanese FrameNet Project: An introduction." *LREC 2004. Proceedings of the Workshop "Building Lexical Resources from Semantically Annotated Corpora"*, p.9-11.

Subirats, Carlos; Petruck, Miriam. (2003) Surprise: Spanish FrameNet. *International Congress of Linguists. Workshop on Frame Semantics,* Prague, Czech Republic.

Tufiş Dan, Ion Radu, Alexandru Ceauşu, Dan Ştefănescu. (2006): Improved Lexical Alignment by Combining Multiple Reified Alignments. In *Proceedings of the 11th Conference of the EACL*, Trento, Italy, 3-7 April, 2006, pp. 153-160.

DEI MULTIMEDIA: EVOLUȚII, PERSPECTIVE

DUMITRU TODOROI[1], ADRIAN CHIORESCU[2]

[1] *Academia de Studii Economice, Chişinău*

[2] *Universitatea "Al.I.Cuza", Facultatea de Informatică, Iaşi*

todoroi@ase.md , chiorescu@yahoo.com

Rezumat

Lucrarea prezintă evoluțiile proiectului **DEI Multimedia** de la lansarea ideii în anul 2000 până în prezent: crearea bazei de date, scenarii de utilizare, posibilități de conversie, exemplu de interfață (web), concluzii şi perspective de viitor.

1. Introducere

Dicționarul Explicativ Ilustrat Multimedia este rezultatul unui proiect al Academiei de Studii Economice din Moldova denumit *„Romanian Language for the European Community"*.

În cadrul unui lung şir de proiecte de informatizare a limbii române, acest proiect a realizat o *bază de date multimedia* pentru Dicționarul Enciclopedic Ilustrat al Limbii Române (DEI) intitulată **DEI Multimedia**.

Concret, este vorba de o bază de date în SGBD Microsoft Acces 2000. S-a ales acest format datorită posibilităților de stocare a obiectelor (OLE) multimedia cum ar fi imagini, sunete, video.

Structura unui tabel are 7 câmpuri astfel:

`cuvânt, cuvânt_accent, definiție` (câmpuri de tip text)

`cuvânt_audio, definiție_audio, imagine, video` (câmpuri OLE)

Lucrările asupra creării Sistemului **DEI Multimedia** au început după Conferința din 14 - 15 aprilie 2000 de la Chişinău cu inițiativa, planificarea, organizarea şi coordonarea lucrărilor de către Prof. univ. dr. hab. Dumitru TODOROI. Lucrările au fost planificate şi organizate în activități pe grupe, compuse din profesori, doctoranzi, masteranzi şi studenți, în total activând până în prezent 53 de persoane.

Grupul TEXT: Zinaida Todoroi, Claudia Vasilache, Ion Linga şi un colectiv compus din 18 doctoranzi, masteranzi şi studenți de la ASEM. Acest grup a realizat **componenta TEXT** (preluarea din DEI), câmpurile sistemului DEI MULTIMEDIA (în aplicația ACCESS), cu etichetele CUVÂNT, CUVÂNT-ACCENT şi CUVÂNT-DEFINIȚIE şi verificarea acestor câmpuri.

Grupul AUDIO: Nicoleta Todoroi, Silvia Donici, Ştefan Spătaru şi un colectiv compus din 12 doctoranzi, masteranzi şi studenți de la ASEM - a realizat **componenta AUDIO**

(înregistrări audio), câmpurile sistemului DEI MULTIMEDIA (în ACCESS), cu etichetele CUVÂNT-ACCENT - AUDIO şi CUVÂNT-DEFINIŢIE - AUDIO şi verificarea acestor câmpuri. S-au cercetat posibilităţile de iniţiere, introducere, menţinere, utilizare şi distribuire a componentelor AUDIO în sistemul DEI MULTIMEDIA.

Grupul IMAGINI: Diana Micusa, Dumitru Micusa, Nicoleta Todoroi, Igor Coseru şi un colectiv compus din 15 doctoranzi, masteranzi şi studenţi de la ASEM. Acest grup a relizat **componenta IMAGINI**, compusă din câmpul sistemului DEI MULTIMEDIA (în ACCESS) cu eticheta CUVÂNT-IMAGINE şi verificarea acestui câmp. În prima etapă au fost examinate posibilităţile de stocare a imaginilor în ACCESS, optimizări necesare etc. În etapa următoare au fost introduse 2630 de imagini color din DEI şi mai târziu au fost adăugate din diferite surse câteva sute de imagini suplimentare.

Grupul VIDEO: Victor Andronatiev, Zinaida Todoroi, Dumitru Micusa, Nicoleta Todoroi şi un colectiv compus din 8 doctoranzi, masteranzi şi studenţi de la ASEM a cercetat posibilitatea de a compune câmpul VIDEO din sistemul DEI MULTIMEDIA cu suport ACCESS. În continuare s-au înregistrat şi extras secvenţe VIDEO de 10-30 secunde, care ar explica acele cuvintele din DEI, diverse surse video, inclusiv Internet. Astfel de circa 100 de secvenţe s-au acumulat în DEI MULTIMEDIA până în prezent.

La ora actuală, **DEI Multimedia** este doar această bază de date, care, pentru un utilizator neavizat, este cam dificil de folosit. Tocmai această problemă se doreşte a fi principalul subiect tratat de lucrarea de faţă, dl. Adrian Chiorescu realizând chiar o interfaţă web ce este descrisă pe larg în capitolul 4.

2. Generalităţi privind dicţionarele informatizate

Dicţionare se realizează la noi în ţară de foarte multă vreme. Cele mai mari dicţionare, cele lingvistice în special, au necesitat ani şi chiar zeci de ani de muncă, unele dintre ele nefiind nici acum terminate. Alta ar fi fost situaţia poate dacă atunci ar fi existat calculatoarele şi programele de azi. În ultimele decenii s-au făcut progrese enorme în domeniul procesării limbajului natural, lingvisticii computaţionale şi astfel s-au reluat şi finalizat multe astfel de proiecte.

Crearea de baze de date lexicale (LDB) este o preocupare a mai multor ţări balcanice. În acest scop a fost iniţiat proiectul CONCEDE ce îşi propune crearea de metode universale de realizare a acestor baze de date. Proiectul se desfăşoară pentru 6 limbi central europene: bulgară, cehă, estonă, ungară, română şi slovenă.

În ultimii ani numeroase proiecte de informatizare a limbii române s-au desfăşurat şi în Republica Moldova, printre cele mai importante fiind cel de informatizare a Marelui Dicţionar al Limbii Române (MDLR). Cel mai recent este cel al Academiei de Studii Economice din Moldova şi anume realizarea unei *baze de date multimedia* pornind de la Dicţionarul Explicativ Ilustrat. Astfel rezultatul a fost baza de date în MS Acces 2000 intitulată **DEI Multimedia** care face şi obiectul de studiu al acestei lucrări în următoarele capitole.

La baza realizării **DEI Multimedia** au stat ideile şi tehnologiile folosite pentru MDLR, adică reunirea a diferite sub-dicționare pe anumite direcții cum ar fi: TEXT, AUDIO, IMAGINI şi VIDEO.

3. Scenarii de utilizare – posibile interfețe utilizator

Utilizarea de către publicul larg a bazei de date multimedia în SGBD MS Acces 2000 (**DEI Multimedia**) este la ora actuală destul de anevoioasă pentru că trebuie să se folosească facilitățile MS Acces-ului pentru navigare, căutare, interogare etc. Ori nu toată lumea cunoaşte limbajul SQL de exemplu pentru o căutare cât mai precisă.

Din acest motiv ar trebui create aplicații care să interacționeze cu această bază de date şi să dispună de interfețe cu utilizatorul cât mai prietenoase şi uşor de folosit. Interogările cu baza de date trebuie să fie transparente pentru utilizator.

Este imposibil de realizat o „aplicație perfectă" care să placă şi să folosească eficient absolut tuturor utilizatorilor şi de aceea ar trebui create mai multe astfel de aplicații, fiecare având un „grup țintă" de utilizatori. Am identificat 5 astfel de grupuri țintă care să acopere cât mai mult din publicul larg vorbitor de limbă română şi nu numai:

3.1. Publicul preşcolar şi şcolar de clase mici (I, II)

DEI Multimedia ar fi ideal în acest scop dacă peste el s-ar construi o aplicație cu o interfață cât mai atrăgătoare pentru copii, cât mai veselă, viu colorată, animată. Aici modul de prezentare este cel care contează foarte mult şi interfața trebuie astfel organizată încât să acorde un spațiu mai larg afişării imaginilor şi a clipurilor video.

3.2. Elevi de ciclu primar (III, IV) şi ciclu gimnazial

O interfață pentru acest grup de utilizatori ar trebui să fie, la fel ca şi în cazul anterior, veselă, viu colorată, atrăgătoare şi sugestivă, uşor de folosit, pentru ca elevii să nu aibă nevoie de ajutorul profesorului şi să o poată utiliza singuri. În acest caz, organizarea interfeței trebuie să ofere în egală măsură spațiu de afişare atât textelor cât şi imaginilor. De asemeni nu trebuie să lipsească opțiunea de căutare.

3.3. Elevi de liceu, studenți, utilizatori cu studii de nivel mediu sau superior

La acest grup deja interfața ar trebui să fie „serioasă". Aplicația nu mai trebuie să fie orientată *mod de prezentare* ci orientată *funcționalitate*. Bineînțeles că nu trebuie pierdut din vedere aspectul estetic, interfața trebuie să fie atrăgătoare, dar în acelaşi timp simplă şi ergonomică, să asigure rapiditatea funcționării.

3.4. Cei ce învață limba română, turişti străini, studenți străini

Deşi pare un grup restrâns, nu trebuie deloc pierdut din vedere, mai ales că în acest caz este necesară o aplicație cu totul specială. Pe lângă faptul că interfața ar trebui prezentată în mai multe limbi de circulație internațională, aceasta trebuie să scoată în evidență

foarte mult controalele audio pentru a se putea studia cu mare atenție pronunția cuvintelor.

3.5. Mediul academic lingvistic, studenți sau profesori de litere

Deşi aceştia sunt „creatorii" dicționarelor, chiar ei au nevoie de multe ori de un dicționar cu un acces foarte rapid la informație atât pentru uz personal cât şi pentru uz didactic.

O aplicație destinată acestui grup de utilizatori va trebui să ofere tot felul de posibilități de generare şi listare de rapoarte, analize, statistici, de salvare, exportare în diverse formate a datelor pentru ca aceştia să le folosească ulterior mai ales în scop didactic.

4. Aplicația web: DEI Multimedia online

Studiind necesitățile fiecărui grup de utilizatori şi făcând o „medie" a acestora am creat aplicația **DEI Multimedia *online*** ce se doreşte a fi universală, adică să poată fi utilizată cu succes de utilizatori din toate grupurile.

Este o aplicație web (online) şi tocmai de aceea prezintă unele avantaje suplimentare: nu trebuie actualizată de utilizatori (este actualizată permanent de echipa DEI Multimedia), asigură foarte uşor feedback-ul, este accesibilă oricând, de oriunde şi de către oricine are o conexiune Internet (independentă de platformă), în schimb, fiind online, nu excelează cu modul grafic de prezentare fiind orientată mai mult spre funcționalitate. Este o aplicație cu un design extrem de simplu dar plăcut (Figura 1).

Am hotărât ca acest exemplu să fie un site web, pentru că nu există unul asemănător pentru limba română. Deşi există numeroase dicționare online complete ale limbii române (www.dexonline.ro), ele prezintă numai definiții în mod text, nici unul nu este multimedia.

DEI Multimedia *online* este o aplicație web distribuită, construită pe arhitectura Client/Server în triplă legătură. Această structură este cea mai întâlnită la aplicațiile distribuite şi se constituie din 3 părți: *interfața cu utilizatorul*, *funcționalitatea* şi *baza de date*, fiecare dintre părți putând fi stocată pe calculatoare diferite (de aici distribuită). Astfel, utilizatorul, operează cu interfața prin intermediul browserului web de pe calculatorul propriu. Acesta emite cereri către *serverul web* pe care se găseşte propriu-zis codul sursă (*funcționalitatea*) şi primeşte informațiile dorite. Browserului web îi rămâne sarcina de a formata informațiile pentru afişare. Dacă sunt necesare informații din baza de date, atunci serverul web este cel care le cere de la *serverul de date* (interogări SQL) şi le prelucrează, acest lucru fiind transparent pentru calculatorul client (utilizator).

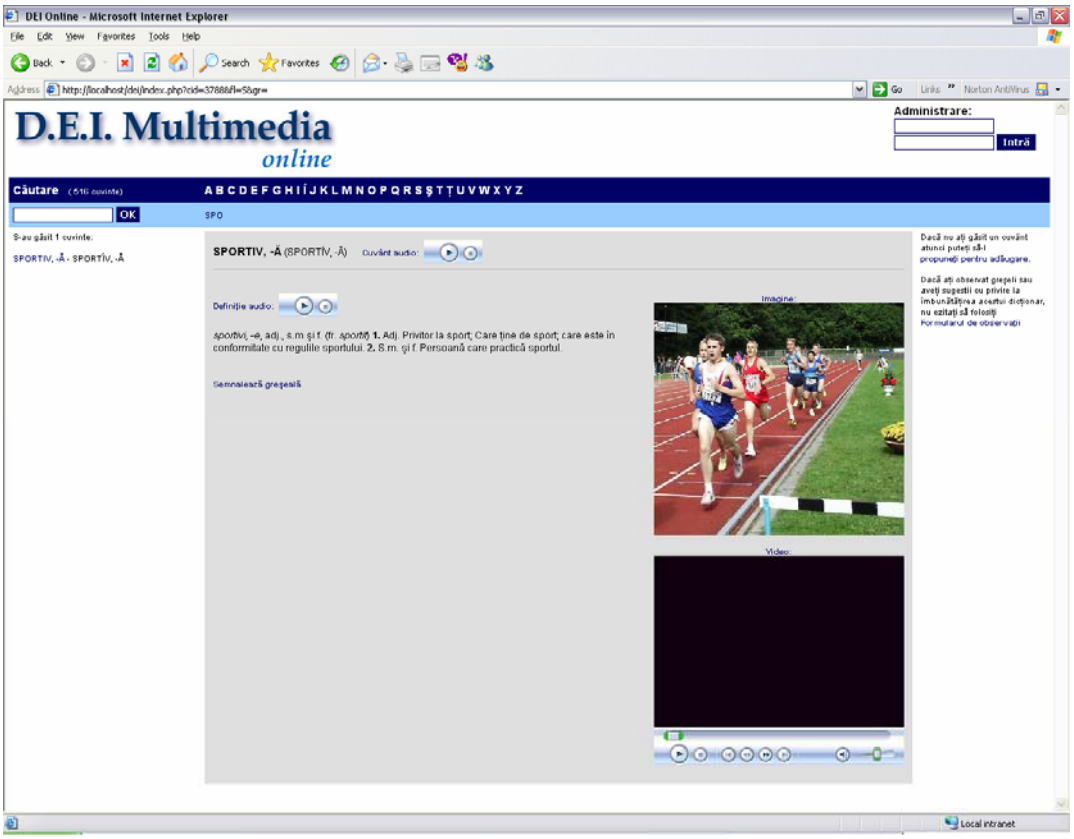


Figura 1 – Interfața cu utilizatorul

Aplicația are de fapt două interfețe: cea prezentată până acum, dedicată utilizatorilor obișnuiți ce caută cuvinte (modulul front-end) și o interfață dedicată administratorilor site-ului, unde aceștia, după ce se autentifică, pot opera modificări, adăugări, ștergeri etc. din baza de date. Acesta se numește *Modul de administrare* (modulul back-end).

Tot în acest modul, administratorii analizează propunerile de cuvinte făcute de utilizatori și le acceptă sau nu în baza de date, de asemenea, tot aici pot vedea și greșelile sau observațiile trimise de utilizatori.

5. *Concluzii și perspective*

Așa cum și-a propus, această lucrare tratează modalitățile prin care baza de date MS Acces 2000, **DEI Multimedia**, poate fi făcută accesibilă publicului larg vorbitor de limbă română și nu numai.

Acest lucru se poate realiza direct construind aplicații (în orice limbaj de programare vizual sub Windows) ce se conectează la baza de date și facilitează interacțiunea utilizatorului cu aceasta folosind interfețe utilizator cât mai accesibile.

DEI Multimedia acum este introdus și în Consorțiul pentru informatizarea limbii române.

Ca perspective de viitor, **DEI Multimedia** doreşte să participe într-un proiect INTAS pentru 2007-2009 sub titulatura *"MULTIMEDIA Multilingual Dictionaries System for Republic of Moldova, Georgia, Armenia, and Azerbaijan in the Process of its Adhesion to the European Community Structures"*.

Referințe bibliografice

Todoroi D., Todoroi, Z., Micusa, D. (2004). Procesarea limbajului natural în baza limbii computerizate române. // *România şi Republica Moldova: Problemele competitivității economiilor naționale. Posibilități de valorificare pe piața internă, europeană şi internațională*, Bucureşti, INCE, 2004, p. 369-375.

Todoroi, N., Todoroi, Z., Todoroi, D. (2004). Complexity Degrees of Illustrated Encyclopaedic Dictionary MULTIMEDIA. *Proceedings of the Int. Symp. "Inovative Applications of Information Technologies in Business and Management"*, October 22-23, 2004, Iaşi, Romania, pp. 23-27.

Todoroi, D., Cristea, D., Tufiş, D., Todoroi, Z. (2003). Limba Română – Limba comunității europene. (LR – LCE – 2000). *Economica*, Nr. 1(44), p. 99-105.

Todoroi D., Micusa, D., Todoroi, Z., I. Linga, I. Covalenco, N. Objelean, S.Spataru, S.Lungu, V. Turcanu, E. Cozlov, N. Ambrozii, V. Slobodeanu, I. Coseru, C. Suruceanu. (2002). Dictionarele multimedia ale limbii române. Secvențe de implementări şi experimentări. *Limba Româna în Societatea Informațională – Societatea Cunoaşterii*, Ed. Expert, Academia Română, Bucureşti, p. 401-421.

Todoroi, D., Todoroi, Z., Micusa, D. (2001). Romanian Computerized Language – One of the European Community Languages. *Proceedings of the 26th Annual Congress of the American Romanian Academy of Arts and Sciences (ARA)*, Montreal, Quebec, Canada, July 25-29, 2001, pp. 133-137. (Rom)

Todoroi, D. (2001). The Computerized Romanian Natural Language Processing Development-Projects-Perspectives. // *INFORMATION SOCIETY. The Proceedings of the 5th International Symposium on Economic Informatics*, May 2001, Ed Economica, Bucharest 10-13 May 2001, pp. 927-935.

Micusa, D., Jucan, T., Todoroi, D. (2002). The E-T-M Formalism for NLP Adaptable Processors' Interactions. *Proc. of the Intern. Conf. "Globalisation And University' Economic Education"*, Vol. II, October 24-27, 2002, Iasi, Romania, pp. 200-218.

MAPAREA CUVINTELOR DINTR-UN LEXICON PE ONTOLOGIE

NATALIA BURCIU, ANTONINA BÎRLĂDEANU

Universitatea Tehnică a Moldovei, Facultatea Calculatoare Informatică şi Microelectronică, Chişinău

natusicb@gmail.com, toni_birlad@yahoo.com

Rezumat

Mapare înseamnă corespondență, adică fiecărui element dintr-o mulțime îi corespunde un alt element din altă mulțime. În termenii acestui proiect – Maparea cuvintelor dintr-un lexicon pe ontologie – maparea este un proces complex care constă în crearea unei ontologii a termenilor juridici, crearea unui lexicon pe baza unui corpus de texte din domeniul juridic, crearea unui adnotator care adnotează fiecare termen juridic din text cu conceptul corespunzător ontologiei.

1. Introducere

Aplicațiile Software pentru Procesarea Limbajului Natural sunt în continuă dezvoltare, în special în domeniile Ontologiilor, Rezumării Informației, Extragerii Informației, Webului Semantic, Traducerii Automate etc. (van Harmelen, Fensel, 1999). Aplicațiile soft pentru Traduceri Automate, Extragerea Informației se dezvoltă, în special, în domeniul guvernamental şi juridic pentru a obține traduceri (documente, texte) mai utile şi mai perfecte decât la nivel de conversație sau texte mai puțin standardizate.

Sunt multe realizări în aceste domenii dar foarte puține pentru limba română. Proiectul „Maparea cuvintelor dintr-un lexicon pe ontologie" are ca obiectiv îmbogățirea realizărilor în domeniu şi a resurselor sistemelor informaționale pentru limba română.

2. Descrierea proiectului

Maparea cuvintelor dintr-un lexicon pe ontologie constă în crearea unei ontologii a termenilor juridici, crearea unui lexicon, crearea unei aplicații – adnotator semiautomat – ce adnotează termenii juridici din texte cu conceptele corespunzătoare din ontologie. Semantici formale sunt de obicei încorporate în ontologii. O „adnotare semantică" prezintă o descriere mult mai precisă a cunoştințelor conținute în texte şi a semanticii acesteia în domeniul juridic. O adnotare semantică trebuie să fie bine definită, uşor de înțeles de către experții din domeniu şi să nu fie ambiguă. Pentru a respecta aceste cerințe, o adnotare semantică trebuie să bazeze pe un model formal al domeniului, de exemplu ontologia. Formalizarea schemei de adnotare utilizînd ierarhia ontologică permite adnotatorului semiautomat să aleagă nivelul corect al detaliului de adnotare, să diminueze ambiguitatea şi să reducă erorile în timpul procesului de adnotare (van Harmelen, Fensel, 1999)

2.1. Ontologia termenilor juridici. Fragment

Ontologia este definită ca specificație a conceptualizării (Horrocks, 2000). Ontologia termenilor juridici constă din 44 clase amplasate ierarhic, cuprinde peste 140 sloturi şi există în format RDFS. Un fragment din ontologie este în figura 1.

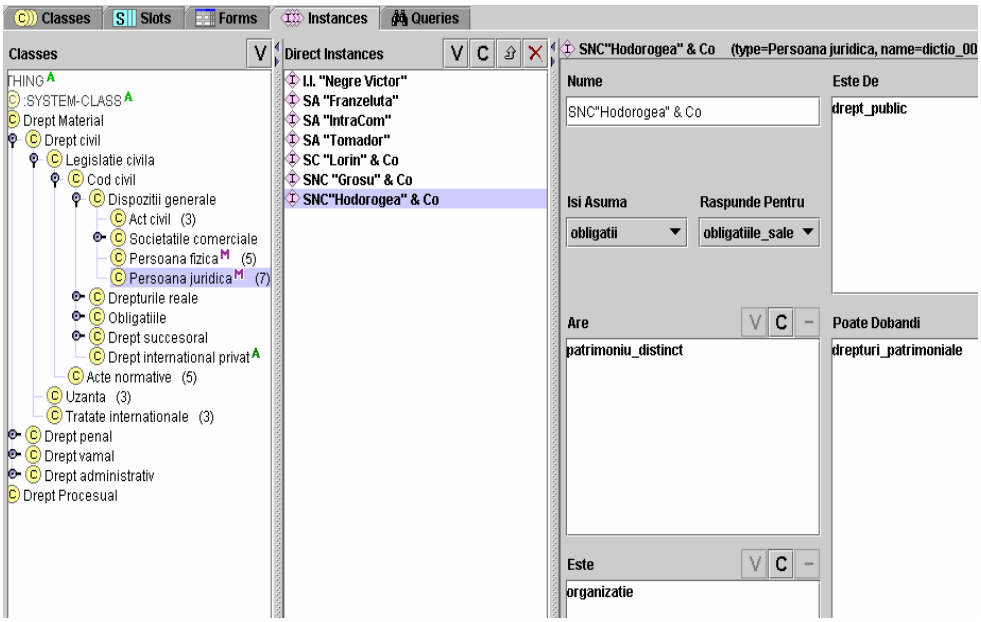


Figura 1: Fragment din ontologia termenilor juridici

2.2. Lexiconul

Lexiconul a fost creat manual dintr-un corpus de texte din domeniul juridic. Fiecare termen juridic este adnotat cu conceptul corespunzător din ontologie via taguri XML ca în tabelul de mai jos:

Decizia	<sem>decizie</sem>
instantei	<sem> instanta judecatoreasca </sem>
recunoasterea	<sem>recunoasterea dreptului</sem>
recunoasterea	<sem>recunoasterea hotararii</sem>
dreptului	<sem> drepturi reale </sem>
spatiu	<sem>patrimoniu </sem>
locativ	
contractului	<sem> contract de locatiune </sem>

Figura 2: Fragment din lexicon

2.3 Adnotatorul semiautomat

Adnotarea Semantică este o tehnologie de bază pentru conținutul inteligent şi este foarte utilă pentru o mulțime de aplicații inteligente orientate pe conținut (Vintar, 2003). Aplicația soft – adnotator semiautomat – a fost creat cu ajutorul limbajului de programare C++. El verifică fiecare cuvînt din text dacă este în lexicon şi respectiv dacă este găsit adnotat copie tagul XML cu conceptul şi îl alipeşte termenului dat în text. Deasemenea, dacă adnotatorul găseşte în lexicon că termenul aparține mai multor concepte, el permite utilizatorului să aleagă conceptul corect, corespunzător contextului în care se află termenul juridic. Aceasta poate fi văzută în figura următoare:

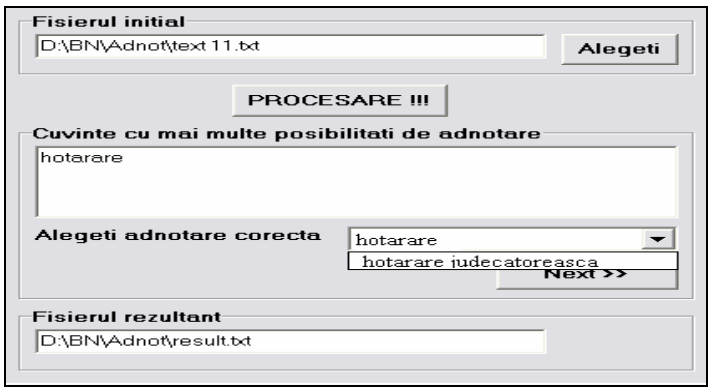


Figura 3: Interfața adnotatorului semiautomat

Fragment de text adnotat:

```
Contractul<sem> contract de vanzare-cumparare</sem> de vinzare-
cumpararea autoturismului<sem> patrimoniu</sem> a fost
recunoscut valabil din motivul ca alin. 2 art<sem> acte
normative</sem>. 49 C.C. nu stabileste un termen concret in
timpul caruia tranzactia executata trebuie sa fie intocmita.
C.D.<sem> persoana fizica</sem> a indicat ca in baza
tranzactiei orale din 12.08.1997 cu P.V.<sem> persoana
fizica</sem> a procurat automobilul<sem> patrimoniu</sem> "Alfa-
Romeo" 164 TD numarul de stat CDP-408 cu pretul de 9.400 lei,
insa nu a dovedit sa intocmeasca contractul<sem> contract</sem>
la notariat<sem> notar</sem> deoarece masina<sem>
patrimoniu</sem> a fost ridicata de la el<sem> persoana
fizica</sem> de catre colaboratorii<sem> persoana fizica</sem>
Considerind ca in aceasta tranzactie el<sem> persoana
fizica</sem> este cumparator<sem> persoana fizica</sem> de
buna credinta, reclamantul<sem> reclamant</sem> solicita sa
fie recunoscuta tranzactia de vinzare-cumparare a
automobilului<sem> patrimoniu</sem> "Alfa-Romeo" dintre
el<sem> persoana fizica</sem> si P.V.<sem> persoana fizica</sem>
   valabila, iar Departamentul sa fie obligat sa-i<sem> persoana
fizica</sem> intoarca automobilul<sem> patrimoniu</sem> care
a fost ridicat ilegal.
Prin hotarirea<sem> hotarare</sem> Judecatoriei<sem> instanta
judecatoreasca</sem> sect.Buiucani mun.Chisinau din 07.09.1998
actiunea<sem> actiunea</sem> a fost admisa.
```

3. Concluzii

Textele cu termenii adnotaţi cu conceptele corespunzătoare din ontologia termenilor juridici, care descrie acest domeniu, va îmbunătăţi procesul de extragere a informaţiei din texte şi documnte cu conţinut juridic pentru limba română. Textele vor contribui deasemenea la obţinerea unor rezultate mult mai calitative în Traduceri Automate prin dezambiguizarea termenilor juridici. Mai mult ca atît, datorită faptului că adnotarea se bazează pe ontologie ne face să utilizăm formalisme standardizate, aşa ca RDF şi OWL care permit reutilizarea acestor adnotări de către alte instrumente de adnotare sau instrumente de căutare.

Referinţe bibliografice

van Harmelen, F., Fensel, D. (1999). Practical Knowledge Representation for the Web. *In Proceedings of the IJCAI Workshop on Intelligent Information Integration*.

Horrocks, I. (2000). The ontology interchange language oil: The grease between ontologies. *Technical report*, Dep. of Computer Science, Univ. of Manchester, UK/ Vrije Universiteit Amsterdam, NL/ AIdministrator, Nederland B.V./ AIFB, Univ. of Karlsruhe, DE.

Vintar, S. (2003). Using parallel corpora for translation-oriented term extraction. Internet [http://www2.arnes.si/~svinta/babel.rtf].

CREAREA RESURSELOR LINGVISTICE CU AJUTORUL UNUI LIMBAJ SPECIALIZAT

ŞTEFAN DIACONESCU

SOFTWIN, Bucureşti
sdiaconescu@softwin.ro

Rezumat

Lucrarea de faţă prezintă o metodă ce permite tratarea relativ unitară a mai multor capitole din lingvistică prin intermediul unui limbaj de reprezentare a cunoştinţelor lingvistice numit GRAALAN (Grammar Abstract Language). Acest limbaj oferă unui lingvist posibilitatea descrierii eficiente a cunoştinţelor lingvistice privind o limbă naturală precum şi corespondenţa între două limbi naturale.

1. Introducere

Există numeroase şi fructuoase încercări de uniformizare a reprezentării cunoştinţelor lingvistice. O asemenea uniformizare ar oferi un avantaj foarte mare în dezvoltarea unor studii, statistici şi, în cele din urmă, aplicaţii lingvistice care să poată trata într-un mod asemănător diverse limbi naturale sau să poată compara (stabili corespondenţe) într-un mod unitar între diverse limbi naturale. Din păcate diversele capitole lingvistice au suferit abordări întrucâtva independente, cum ar fi subcategorizarea (EAGLES, 1996b) adnotarea (EAGLES, 1996b), lexiconul (EAGLES, 1993), etc. astfel încât este uneori dificil de aplicat tratamente unitare.

Comunicarea de faţă prezintă un limbaj de reprezentare a cunoştinţelor lingvistice numit GRAALAN (Grammar Abstract Language). Acest limbaj permite unui lingvist descrierea eficientă a cunoştinţelor lingvistice privind o limbă naturală precum şi corespondenţele între două limbi naturale.

2. Caracteristicile generale ale GRAALAN

Din punct de vedere teoretic, GRAALAN se bazează în special pe următoarele noţiuni: gramatici generative de dependenţe (GDG - Generative Dependency Grammar) (Diaconescu, 2002), arbori de dependenţe (DT - Dependency Tree) (Diaconescu, 2002) şi arbori atribut - valoare (AVT - Attribute Value Tree) (Diaconescu, 2005).

Pornind de la aceste noţiuni, GRAALAN poate descrie diverse capitole lingvistice conforme cu gramaticile convenţionale ale limbilor naturale: alfabetul, despărţirea în silabe, morfologia, sintaxa, regulile de flexiune, formele de flexiune, lexiconul, corespondenţe lexicale între două limbi (inclusiv între expresii multicuvânt MWE - Multiword Expression), corespondenţe morfologice, corespondenţe sintactice.

GRAALAN este în esenţă un limbaj descriptiv care permite însă eventual şi legătura cu anumite subprograme de tip procedural scrise în alte limbaje de programare.

În principiu, descrierile GRAALAN vor putea fi convertite printr-un compilator adecvat în formatul XML care este mai adecvat exploatării ulterioare prin diverse programe.

3. Descrierea alfabetului

În GRAALAN se pot preciza pentru o anumită limbă: alfabetul fonetic utilizat în descrierea limbii (care poate fi un subset al IPA (International Phonetic Alphabet) (IPA, 2005), alfabetul normal şi caracterele speciale.

În afară de acestea se mai pot defini: i) grupe de caractere (diftongi, triftongi, etc.), transcrise cu caractere normale (eventual speciale) dar şi fonetice; ii) clase alfabetice (de exemplu clasa vocalelor, clasa consoanelor, etc.)

Caracterele folosite în GRAALAN se consideră codificate în UNICODE (ISO, 1992).

4. Descrierea despărțirii în silabe

În GRAALAN sunt considerate trei tipuri de despărţire în silabe: i) Despărţirea eufonică a cuvintelor scrise cu alfabetul normal şi respectând modul de pronunţie; ii) Despărţirea fonetică a cuvintelor scrise cu alfabetul fonetic şi respectând de asemenea modul de pronunție; iii) Despărțire morfologică - analogă cu despărţirea eufonică însă respectând şi restricții ce țin cont de structura morfematică a cuvântului.

Primele două tipuri au reguli specifice. Ulitmul tip nu are reguli speciale deoarece ea acționează ca o despărțire eufonică cu restricțiile suplimentare privind morfemele obținute din consultarea lexiconului.

5. Descrierea morfologiei

În GRAALAN, morfologia unei limbi (mai exact ansamblul categoriilor lexicale şi al valorilor lor), se reprezintă sub forma unui arbore atribut valoare (AVT) (Diaconescu, 2005) în care nodurile de tip atribut corespund categoriilor lexicale iar nodurile de tip valoare corespund valorilor categoriilor lexicale. În plus, cele două tipuri de noduri mai au ataşate diverse alte tipuri de informații: numele, abrevieri, (eventual) ataşamente procedurale, etc.

În secțiunea corespunzătoare morfologiei se poate indica de asemenea dacă anumitor situații de flexiune distincte le corespund forme flexionate identice.

6. Descrierea lexiconului

Lexiconul GRAALAN este un ansamblu de intrări de diverse tipuri: i) Morfeme (rădăcini, prefixe, sufixe, prefixoide, sufixoide etc.); ii) Cuvinte care la rândul lor pot fi: intrări principale de tip lemă, intrări suplimentare (care însoțesc o lemă), intrări principale care nu sunt însă leme; iii) MWE-urile cărora li se indică şi structura sub forma unui arbore de dependențe (DT); iv) Structuri morfologice analitice sau analitico-sintetice (forme flexionate formate din mai multe cuvinte) analoge MWE-urilor; v) Structuri sintactice de asemenea analoge MWE-urilor.

În funcție de tipul lor, intrările în lexicon mai pot avea asociate şi alte tipuri de informații: semantice, etimologice, morfologice, etc.

Lexiconul în general nu este scris direct în GRAALAN ci se creează cu ajutorul unui instrument specializat.

7. Descrierea regulilor de flexiune

Intrarea din lexicon care se poate flexiona (lema de exemplu) identifică o regulă compusă de flexiune aflată în secțiunea GRAALAN a regulilor de flexiune. Regula compusă este o listă de reguli de bază. O regulă de bază este de fapt un arbore atribut valoare care indică mai multe situații de flexiune, câte una pentru fiecare frunză a sa. Fiecare situație de flexiune (deci frunză) are asociată una sau mai multe reguli de flexiune elementare. O regulă de flexiune elementară conține: i) O condiție de aplicare a regulii; ii) O secvență de transformări care trebuie făcute asupra lemei (sau asupra altei forme de flexiune) pentru a obține forma de flexiune curentă exprimată în alfabetul normal; iii) Analog cu (ii) pentru alfabetul fonetic; iv) În cazul formelor analitico-sintetice - o caracterizare sub forma unui AVT pentru fiecare cuvânt component şi relațiile care se află între diversele cuvinte componente.

Pe baza regulilor de flexiune aplicate intrărilor din lexicon se pot obține formele din secțiunea GRAALAN a formelor de flexiune.

8. Descrierea formelor de flexiune

Secțiunea GRAALAN corespunzătoare formelor de flexiune conține câte o intrare pentru fiecare formă de flexiune. O intrare conține: i) Forma de flexiune în alfabet normal şi fonetic; ii) Identificarea în lexicon a intrării căreia îi corespunde forma respectivă de flexiune; iii) Caracterizarea formei de flexiune sub forma unui ansamblu de categorii lexicale cu valorile lor (AVT); iv) Despărțirea în silabe.

Formele de flexiune nu sunt scrise în general direct în GRAALAN ci se creează cu ajutorul unui instrument specializat.

9. Descrierea sintaxei

Sintaxa se descrie în GRAALAN sub forma unei liste de reguli sintactice etichetate (care respectă principiile gramaticilor de dependențe generative (Diaconescu, 2002)).

O regulă are un membru stâng care conține un neterminal însoțit de un AVT format din categorii lexicale şi/sau sintactice) şi un membrul drept care conține unul sau mai mulți alternanți. Un alternant este format din trei subsecțiuni:

a) Subsecțiunea sintactică care conține o secvență de NTPA: Neteminali, Terminali, Pseudo terminali, Acțiuni (subprograme procedurale). Neterminalii şi terminalii au accepțiunea obişnuită. Pseudoterminalii sunt neterminali care, dacă ar avea reguli care să îi descrie, acestea ar conține direct terminali din lexicon. Acțiunile sunt subprograme procedurale care ar putea fi utilizate în anumite tratamente specifice dacă este cazul.

Fiecare NTPA conține un nume, un AVT format din categorii lexicale şi/sau sintactice, modul de legare (relaționare) cu alți NTPA.

b) Subsecțiunea de dependențe unde se descriu relațiile de dependență între NTPA-uri ale alternantului. Relațiile de dependență pot fi de tip de regență / subordonare sau de tip coordonare.

c) Subsecțiunea de acord care descrie acordul între NTPA-urile alternantului sub forma unor condiții complexe.

Descrierea sintaxei în GRAALAN este reversibilă adică poate fi folosită şi în procesul de analiză sintactică prin care se generează din textul de suprafață un arbore de dependențe ca formă de adâncime, şi în procesul de generare din arborele de dependențe a textului de suprafață.

10. Descrierea corespondențelor bilingve

Secțiunea GRAALAN privitoare la corespondențele bilingve descrie corespondențe între următoarele tipuri de elemente aparținând la două limbi diferite:

a) Corespondențe între MWE-uri care sunt reprezentate în lexicon sub forma unor arbori de dependențe se exprimă prin echivalarea între expresia sursă şi expresia țintă corespunzătoare dar şi prin regulile de transformare care indică modul în care extensiile expresiei sursă din instanțe reale sunt preluate de expresia țintă.

b) Corespondențe între cuvinte. Este un caz particular al corespondenței între MWE-uri în care expresiile echivalate au câte un singur cuvânt.

c) Corespondențe între structuri sintactice. Este un caz particular al corespondenței între MWE-uri în care cele două expresii pot avea drept caracterizări de noduri nu numai categorii lexicale (cu valorile lor) ci şi categorii sintactice (cu valorile lor).

d) Corespondențe între structuri morfologice. Este un caz particular al corespondenței între MWE-uri în care cel puțin expresia sursă corespunde unei forme flexionate analitico-sintetice.

e) Corespondențe între subarbori morfologici. Este o corespondență între diverse seturi de categorii lexicale (cu valorile lor) organizate sub forma unor AVT-uri.

Informațiile din secțiunea de corespondențe bilingve GRAALAN se pot folosi în aplicații de generare a unor dicționare sau în aplicații de traducere automată.

11. Concluzii

Descrierile de cunoştinţe lingvistice pot fi formulate direct în GRAALAN sau, în anumite cazuri (cum ar fi de exemplu pentru formele flexionate sau pentru lexicon) pot fi create cu ajutorul unor instrumente (programe) speciale care generează text GRAALAN. Textul GRAALAN obținut pe o cale sau pe alta se compilează cu un compilator adecvat care traduce textul GRAALAN în XML, creindu-se astfel o Baza de cunoştinţe lingvistice XML creată prin intermediul GRAALAN va putea fi exploatată într-un mod unitar pentru diverse studii sau pentru elaborarea de aplicații informatice.

Deoarece textul GRAALAN se realizează în mai multe tranşe, o componentă specială GRAALAN Link va determina legăturile între aceste tranşe şi compatibilitatea lor.

Un compilator GRAALAN este în curs de implementare şi unele cunoştinţe lingvistice privind limba română au fost deja scrise în GRAALAN.

Referinţe bibliografice

Diaconescu, S. (2002). Natural Language Understanding Using Generative Dependency Grammar, în Max Bramer, Alun Preece and Frans Coenen (Eds), *Proceedings of ES2002*, Cambridge UK, Springer, pp.439-452.

Diaconescu, S. (2003). Natural Language Agreement Description for Reversible Grammars, în Tamás D. Gedeon, Lance Chun Che Fung (Eds.), *Proceedings of AI 2003*, Perth, Australia, pp. 161-172.

Diaconescu, S. (2004) Multiword Expression Translation Using Generative Dependency Grammar, în *Proceedings of ESTAL 2004 - ESPAÑA for NATURAL LANGUAGE PROCESSING*, Alicante, Spain.

Diaconescu, S. (2005). Some Properties of the Attribute Value Trees Used for Linguistic Knowledge Representation, Proceedings of IICAI-05, INDIA.

EAGLES (1996a). Recommendations for the Morphosyntactic Annotation of Corpora.

EAGLES (1996b). Preliminary Recommendations on Subcategorisation.

EAGLES (1993). Lexicon architecture Draft Report, EAG--LSG/IR--T1.1.

IPA (2005) International Phonetic Association (2005): Handbook of IPA.

ISO/IEC 10646 (1992). Information technology -- Universal Multiple-Octet Coded Character Set (UCS).

RESURSE LINGVISTICE ROMÂNEŞTI ÎN FORMAT ELECTRONIC. *BIBLIA 1688*

BOGDAN-MIHAI ALDEA[1], GABRIELA HAJA[2]

[1]*Facultatea de Informatică, Universitatea "Al.I.Cuza", Iaşi*
[2]*Institutul de Filologie Română „A. Philippide", Academia Română, Iaşi*

{bogdan.aldea, gabihaja}@ gmail.com

Rezumat

Crearea resurselor textuale în format electronic prelucrat reprezintă o prioritate pentru procesul de informatizare a cercetării lingvistice româneşti. Un domeniu puțin cercetat la noi este cel al achiziționării în format electronic şi al prelucrării textelor româneşti vechi. Lucrarea de față descrie rezultatele la care s-a ajuns în prelucrarea a două cărți din *Biblia de la 1688*, ms. 45 şi ms. 4389 (sec. XVII) în vederea generării automate a indicelui de cuvinte.

1. Introducere

Grație colaborării ştiințifice dintre cercetători ai Institutului de Filologie Română „A. Philippide", cercetători ai Institutului de Informatică Teoretică – Academia Română, Filiala Iaşi, cercetători de la Facultatea de Informatică şi de la Facultatea de Litere a Universității „Alexandru Ioan Cuza" din Iaşi, la inițiativa specialiştilor informaticieni, s-a demarat la Institutul „Philippide" din Iaşi un amplu proces de informatizare a cercetării filologice româneşti. Meritul specialiştilor de la acest institut este acela că toate eforturile lor în această direcție s-au concentrat asupra creării unor instrumente şi resurse specifice proiectelor Academiei Române, dintre care le amintim pe cele prioritare: *Dicționarul limbii române (DLR), Dicționarul general al literaturii române (DGLR), Noul Atlas lingvistic român, pe regiuni. Moldova şi Bucovina (NALR – MB).*

Cu referire la cercetările lexicografice din domeniul limbii române, a fost finalizat, în 2005, grantul *Dicționarul limbii române (DLR) în format electronic. Studii privind achiziționarea*, rod al colaborării între lingviştii lexicografi de la Iaşi şi cercetători de la Facultatea de Informatică din Iaşi. Rezultatele acestui proiect au fost publicate (Haja et al., 2005), şi trebuie subliniat că necesitatea definitivării unei forme electronice **integrale** a *Dicționarului limbii române* se impune cu o tot mai mare stringență.

Finalitatea cercetărilor din cadrul proiectului încheiat în 2007 ca şi ale celui în desfăşurare la Institutul „Philippide" este realizarea unui *Dicționar al limbii române informatizat (DLRI),* creat ca instrument şi resursă lexicografică, punct de plecare în constituirea unui dicționar al limbii române care să poată fi permanent actualizat şi îmbogățit prin reeditări consecutive, comparabil cu lucrări fundamentale pentru culturile lumii – tezaure ori dicționare ale limbii electronice – precum cele realizate de lexicografia franceză, italiană, spaniolă, ca să amintim doar trei dintre limbile romanice europene, ori de lexicografia anglosaxonă europeană şi americană.

După ce a fost stabilită soluția achiziționării în format electronic a formei tipărite a *DLR* şi s-a creat instrumentul de achiziție, prelucrare şi consultare a acestuia, DLRex, un alt element care este indispensabil realizării *DLRI*, anume **informatizarea colecției de texte** din care sunt excerptate atestările după care se redactează *DLR*, a intrat în atenția noastră. Pentru început, am optat pentru crearea formatului electronic al ediției unei lucrări care face parte din *Bibliografia DLR*: prima traducere integrală în română a *Vechiului* şi *Noului Testament*, citată de literatura de specialitate sub numele de *Biblia de la Bucureşti (BB)* sau *Biblia de la 1688*, un monument al limbii şi culturii româneşti (Andriescu, 1997; Miron, 1988, 2004).

2. Tradiție şi actualitate

2.1. Monumenta linguae dacoromanorum. Biblia de la 1688

Reeditarea critică a *BB*, într-o ediție cu format enciclopedic, în care sunt cuprinse şi două variante de traducere realizate în acelaşi secol XVII, dar rămase în formă manuscrisă, *Manuscrisul 45* (ms. 45) şi *Manuscrisul 4389* (ms. 4389), a fost inițiată de Paul Miron, profesor la Albert-Ludwigs-Universität din Freiburg, Germania, şi realizată la Iaşi, prin concursul specialiştilor români şi germani, lingvişti, istorici literari, informaticieni şi istorici, oameni de cultură implicați într-un proiect amplu de recuperare şi valorificare a textelor fundamentale ale culturii române.

Din această ediție, proiectată în 20 de părți, au fost tipărite, în seria *Monumenta linguae dacoromanorum*, la Editura Universității „Alexandru Ioan Cuza" din Iaşi, şapte volume (în ordinea cărților din biblice: *Pentateuhul*, 1988–1997, *Iosue, Judicum, Ruth* 2005, *Liber Psalmorum* 2003). În prezent se lucrează, în cadrul grantului *Resurse lingvistice în format electronic. Monumenta linguae dacoromanorum. Biblia 1688. Pars VII. Regum I, Regum II – ediție critică şi corpus adnotat*, la cel de-al optulea volum.

Obiectivele acestui proiect sunt, pe lângă continuarea monumentalei ediții în forma sa tipărită, definirea unui format electronic al acestui volum, adnotarea semiautomată, la nivel de cuvânt, a textului românesc vechi, crearea unui program de indexare a textului. La finele proiectului se vor fi creat premisele constituirii unui nucleu de corpus de limbă română veche, necesar cercetărilor lingvistice în genere, a celor lexicografice în special, şi va fi definitivat instrumentul de realizare a formatuului electronic al întregii ediții.

În proiect sunt implicați cercetători din domeniul lingvisticii şi al informaticii care participă la activități diferite (cercetarea filologică a textului şi crearea instrumentelor informatice de prelucrare a textului) şi de comune (prelucrarea textului), ca în Figura 1.

Lingvişti: - transcrierea interpretativă a textelor; - stabilirea textului, colaționarea, corectura, revizia; - pregătirea volumului pentru tipar.	Informaticieni: - realizarea unui instrument de achiziționare în format electronic a textelor din sec. XVII; - achiziționarea unor eşantioane de text din *Monumenta linguae dacoromanorum. Biblia 1688. Pars VI* şi *Pars VII*; - generarea indicelui de cuvinte.
Lingvişti + informaticieni: - adnotarea la nivel de cuvânt a textelor; - definitivarea formatului electronic al volumului.	

Figura 1: Distribuirea activităților.

Potrivit tradiției create prin editarea volumelor de până acum, pe lângă reproducerea textului tipărit la 1688 însoțit de transcrierea interpretativă a acestuia, alături de cele două manuscrise, se propune o variantă apropiată de limba română contemporană a textului vechi, stabilită de partenerii de proiect din Freiburg. Aparatul critic al ediției constă în realizarea notelor de transcriere, a comentariilor filologice şi, acolo unde este cazul, istorice, a indicelui de cuvinte pentru textul tipărit. O inovație pe care şi-o propune proiectul în realizarea formei electronice a volumului în curs de finalizare o constituie lărgirea indexării asupra manuscriselor şi asupra variantei actualizate a textului. Dintre activitățile descrise mai sus, au fost finalizate următoarele: transcrierea interpretativă a textelor (*BB*, ms. 45, ms. 4389); colaționarea şi corectura textelor; achiziționarea unor eşantioane de text din *Pars VI*, achiziționarea unor eşantioane din *Pars VII*; realizarea unui instrument de achiziționare, prelucrare, indexare a textelor numit convențional AdBB (Ad < adnotare; BB < *BB*).

2.2. AdBB şi funcționalitățile sale

AdBB reprezintă un instrument de achiziționare, prelucrare şi consultare a *BB*, a ms. 45 şi a ms. 4389 în format electronic. Principalele funcționalități ale acestei aplicații create, sunt: permite trecerea textului *Bibliei* (sec. XVII) din format .rtf (Word) în format XML; permite vizualizarea şi corectarea fişierelor XML; generează grupuri de ocurențe ale unei lemme; permite editarea grupurilor; generează şi vizualizează indicele de cuvinte.

2.2.1. Trecerea din format .rtf în XML

Se alege din meniu opțiunea RTF → XML pentru conversia unui text din format Word în format XML:

Figura 2: AdBB – meniu. Captură de ecran.

În urma parsării fişierului .rtf, rezultă un fişier XML cu următorul format:

```
<?xml version="1.0" encoding="UTF-8"?>
<biblia>
      <carte name="NUME_CARTE">
         <capitol id="NR_CAPITOL">
              <verset id="NR_VERSET"> Conținutul primului verset </verset>
              <verset id="NR_VERSET"> Conținutul versetului al doilea </verset>
              <verset id="NR_VERSET"> Conținutul versetului al treilea </verset>
              …….
         </capitol>
         <capitol id="NR_CAPITOL">
         </capitol>
      </carte>
      <carte name="NUME_CARTE">
         <capitol id="NR_CAPITOL">
         …….
         </capitol>
      </carte>
</biblia>
```

2.2.2. Vizualizarea fişierelor XML

Pentru vizualizarea fişierului XML se poate alege cea de-a doua opţiune din meniu, în acest fel putându-se revizui corectitudinea parsării.

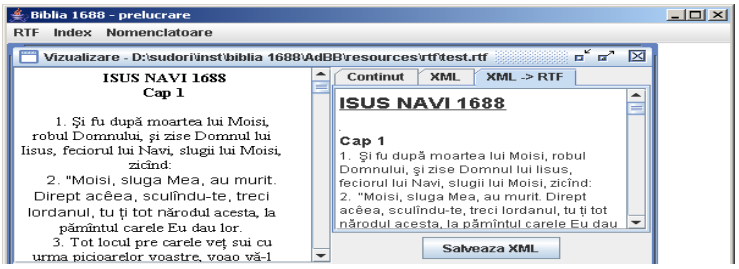


Figura 3: AdBB. Vizualizarea textului .rtf şi XML, înainte de corectură. Captură de ecran.

2.2.3. Generarea grupurilor de ocurenţe

Pentru a genera grupurile de forme flexionare se alege din meniul „Index" opţiunea „Generare Grup":

1) Pentru gruparea cuvintelor dintr-un fişier, se generează vocabularul fişierului selectat şi se elimină o serie de cuvinte cum ar fi: pronumele, articolele ş.a.m.d (vezi nomenclatoare)[1], rezultând un fişier XML, cu grupuri.

2.2.4. Editarea grupurilor forme flexionare

Pentru editarea grupurilor de forme flexionare, se alege cea de-a doua opţiune din meniul „Index": „Editează Gr. Cuvinte", şi se alege pentru deschidere fişierul XML generat la pasul anterior.

Figura 4: AdBB. Editarea grupurilor de ocurenţe generate. Captură de ecran.

Astfel rezultă 3 coloane: Cuvinte, Grupuri şi Cuvinte Grup. Prima coloană cuprinde cuvintele cu o singură apariţie în text sau care nu prezintă forme flexionare, a doua cuprinde lista tuturor grupurilor generate automat, iar ultima cuprinde formele paradigmatice, existente în text, ale fiecărei intrări din coloana a doua, cu posibilitatea modificării acestora de către specialiştii lingvişti.

2.2.5. Generarea şi vizualizarea indexului

[1] Această serie de cuvinte este conţinută într-un fişier intern ce poate fi editat şi anume „romanian.stoplist". StopList este o lista de cuvinte pentru care nu se doreşte generarea indexului şi care poate fi modificată selectând din meniul „Nomenclatoare" opţiunea „Editeaza StopList"

Pentru orice fişier XML rezultat în urma parsării şi pentru orice fişier de grupuri de ocurenţe, se poate genera un index. Se selectează a treia opţiune din meniul „Index”, „Generează index”, rezultând o interfaţă din care se selectează fişierul XML dorit, precum şi fişierul de grupuri aferent acestuia şi apoi se generează indicele de cuvinte. În urma generării, rezultă un fişier XML cu următorul format:

```
<?xml version="1.0" encoding="UTF-8"?>
<indecsi>
<data>
<grupFilePath>FISIERUL_XML_CU_GRUPURI</grupFilePath>
<bibleFilePath>FISIERUL_XML_CU_BIBLIA</bibleFilePath>
</data>
<cuvinte>
      <grup name="NUME_GRUP_#1">
         <cuvant value="CUVANT_#1">
            <index>
                <carte>NUME _CARTE</carte>
                <capitol>NR_CAPITOL</capitol>
                <verset>NR_VERSET</verset>
            </index>
         </cuvant>
      </grup>
</cuvinte>
</indecsi>
```

Pentru vizualizarea indicelui, se alege ultima opţiune din meniul „Index” şi se selectează fişierul XML generat la pasul anterior. Fiecărui cuvânt fiindu-i precizat locul în text, în formatul următor: [NUME_CARTE ; NR_CAPITOL ; NR_VERSET]

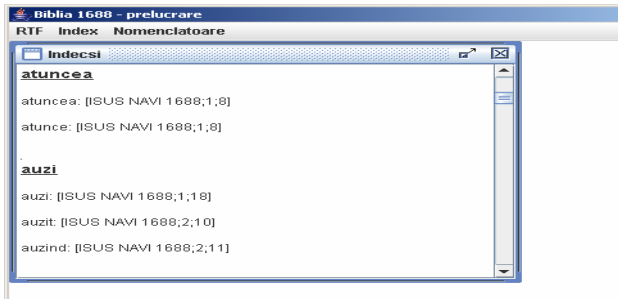


Figura 4: AdBB. Indexul, cu trimiteri, editat. Captură de ecran.

2.3. Necesitatea rafinării AdBB

Pentru o mai bună utilitate a AdBB, este necesară rezolvarea următoarelor chestiuni puse, deocamdată, de specialiştii lingvişti consultaţi cu privire la rezultatele aplicaţiei: **1**) posibilitatea extragerii de contexte (de diferite dimensiuni) în care apar ocurenţele. Ideal al fi ca, din fişierul „Index”, să se poată accesa, prin link-uri, textul în format XML sau .rtf, din care să se selecteze citatele, cu posibilitatea editării acestora în DLRI. Ar fi un câştig, în acest sens, corelarea AdBB cu DLRex, astfel încât, pornind de la o formă paradigmatică ce apare în DLRI, aceasta să poată fi căutată automat în corpusul *BB*; **2**) cuvintele cu o singură apariţie în text trebuie, de asemenea, indexate; există şi în prezent această posibilitate, dar trebuie simplificaţi paşii necesari; **3**) cuvintelor cu frecvenţă foarte ridicată (cuvintele de relaţie, pronumele etc., inventariate în „StopList”) ar trebui să li se precizeze măcar frecvenţa; până în prezent, au fost prelucrate, prin AdBB, eşantioane mici de text; **4**) randamentul aplicaţiei va putea fi verificat după ce vor fi supuse prelucrării fişiere .rtf care să cuprindă peste 100000 de ocurenţe.

3. În loc de concluzii

Acest tip de cercetare interdisciplinară nu poate fi decât constructiv. Ceea ce s-a realizat, într-o perioadă relativ scurtă, deschide o cale nouă cercetării filologice româneşti, în care instrumentele electronice sunt adaptate diverselor tipuri de scriitură, demonstrându-se faptul că vechimea textului nu pune probleme majore mijloacelor actuale de prelucrare a limbii în forma sa scrisă.

Odată create, aceste instrumente facilitează cercetarea filologică, al cărei specific este minuțiozitatea şi acribia, studiul comparativ şi contrastiv al textelor, analiza grafiei şi „arheologia" lingvistică, reducând spectaculos perioada de documentare şi aceea de fişare a textelor, oferindu-i specialistului posibilitatea concentrării asupra fazei de analiză şi interpretare, permițându-i concluzii mai ferme, pe baza unei evidențe cvasiexhaustive a faptelor de limbă. Acesta este doar un exemplu de ameliorare a cercetării, determinată de actualizarea instrumentelor şi de informatizarea resurselor, într-un domeniu atât de vast precum acela descris.

Mulțumiri. Cercetarea descrisă aici s-a desfăşurat în cadrul grantului *Resurse lingvistice în format electronic. Monumenta linguae dacoromanorum. Biblia 1688. Pars VII. Regum I, Regum II – ediție critică şi corpus adnotat* (2006–2007). Autorii mulțumesc Ministerului Educației şi Cercetării, CNCSIS, pentru susținerea financiară a proiectului.

Referințe bibliografice

Andriescu, Al. (1997). *Locul* Bibliei de la Bucureşti *în istoria culturii, literaturii şi limbii române literare* în *Studii de filologie şi istorie literară*, Iaşi, Editura Universității „Alexandru Ioan Cuza", 90-208.

Haja, G., Forăscu, C., Dănilă, E., Aldea, B. M. (2005). *Dicționarul limbii române (DLR) în format electronic. Studii privind achiziționarea*, Iaşi, Editura Universității „Alexandru Ioan Cuza".

Miron, Paul (1988). *O nouă ediție a Bibliei lui Şerban* în *Monumenta linguae dacoromanorum. Biblia 1688. Pars I. Genesis*, Iaşi, 3-6.

Miron, Paul (2004). *Prefață la ediția Freiburg und München a cărții* Ruth, în *Monumenta linguae dacoromanorum. Biblia 1688. Pars VI. Iosue, Iudicum, Ruth*, Iaşi, Editura Universității „Alexandru Ioan Cuza", 5-6.

RESURSE ROMÂNEŞTI ÎN CADRUL PROIECTULUI LT4EL

DIANA TRANDABĂŢ[1,2], ADRIAN IFTENE[1], IONUŢ PISTOL[1],
CORINA FORĂSCU[1,3], DAN CRISTEA[1,2]

[1]*Facultatea de Informatică, Universitatea "Al. I. Cuza", Iaşi*
[2]*Institutul de Informatică Teoretică, Academia Română, Filiala Iaşi*
[3]*Institutul de Cercetări pentru Inteligenţă Artificială, Academia Română, Bucureşti*

{dtrandabat, adiftene, ipistol, corinfor, dcristea}@info.uaic.ro

Rezumat

Proiectul LT4eL are ca scop realizarea unei tehnologii multilingve care să ajute la realizarea şi exploatarea obiectelor de învăţare utilizate în cadrul unui sistem de eLearning. La baza acestora stă un corpus semnificativ de documente, colectat iniţial, apoi adnotat manual şi automat pe diferite niveluri. Tehnologiile dezvoltate în proiect vor facilita operaţiile de creare a obiectelor de învăţare de către profesori cât şi de regăsire a lor de către studenţi, inclusiv prin criterii de natură semantică. Lucrarea prezintă etapele de colectare şi prelucrare a acestui corpus pentru limba română.

1. Introducere

Proiectul LT4eL[1] (Tehnologii Lingvistice pentru eLearning) îşi propune utilizarea unor tehnologii multilingve, unelte lingvistice şi tehnologii ale web-ului semantic pentru a perfecţiona regăsirea şi accesul la materiale de învăţare în sistemele de management al învăţării, prin generarea semi-automată a unor metadate descriptive. Astfel, va fi dezvoltat un extractor de cuvinte cheie şi un detector de definiţii şi termeni definiţi, adaptate tuturor limbilor implicate în proiect (bulgară, cehă, engleză, germană, malteză, olandeză, poloneză, portugheză şi română).

Tehnologia ce va fi dezvoltată în cadrul proiectului va facilita accesul personalizat la cunoştinţele din sistemele de management al învăţării şi va favoriza descentralizarea şi cooperarea în managementul conţinutului didactic (Monachesi et al., 2006).

După o trecere în revistă în secţiunea 2 a cerinţelor specifice proiectului, cu accent pe domeniile din care s-au extras resursele, în secţiunea 3 vom prezenta succint etapele de prelucrare a resurselor, de la forma iniţială în care au fost colectate din diferite surse, până la forma în care vor fi folosite drept corpus de antrenare/test în proiect.

2. Colecţia de obiecte de învăţare

Pentru a îmbunătăţi managementul, distribuţia şi regăsirea materialului de învăţare prin ataşarea semi-automată de metadate este necesară, într-o primă etapă, observarea modului în care aceste metadate sunt marcate manual. Astfel, prima cerinţă a proiectului

[1] http://www.lt4el.eu

a fost colectarea şi normalizarea unor obiecte de învățare, obiectiv realizat prin intermediul unui portal special dezvoltat[2] de partenerii din Universitatea „Al.I.Cuza" Iaşi (Pistol et al, 2006).

Cuvântul de ordine al proiectului LT4eL este multilingvismul. Cu nouă limbi implicate, resursele care vor constitui baza de plecare a extractoarelor automate de cuvinte cheie şi definiții trebuie să fie comparabile pentru toate limbile (din aceleaşi domenii). Necesitatea ca domeniile reprezentate în proiect să fie relativ uniform acoperite în toate cele nouă limbi, să aibă o mare deschidere spre sisteme de eLearning, iar documentele să îndeplinească simultan şi criteriul de accesibilitate cu restricții minime în privința drepturilor de autor au dus la alegerea domeniului informatic, cu precădere a celui dedicat predării de noțiuni informatice către începători, şi a domeniului eLearning. Aceste două domenii mari au fost rafinate în mai multe subdomenii, printre care: *Writing a diploma paper*, *Making an interview*, *Using MS Word*/*Excel*/*PowerPoint*/*Latex*/*XML*, *Creating Web pages*, *Accessing the Internet*, *eLearning*, *eMarketing*, *Impact of use of computers in society*, *Impact of eLearning on education*, *Calimera Documents* etc.

Corpusul colectat pentru limba română conține 56 de documente din aproape toate domeniile şi subdomeniile avute în vedere între partenerii proiectului; la ora actuală acest corpus însumează 683.357 cuvinte. Descrierea fiecărui obiect de învățare se face printr-un nume, un set restrâns de cuvinte cheie[3] şi datele privind drepturile de autor[4]. Din motive statistice, pentru fiecare resursă se calculează numărul de cuvinte.

3. Formatul obiectelor de învățare

Resursele lingvistice au avut, la momentul colectării lor, diferite formate (nivelul 1 din Figura 1.): *.doc*, *.pdf*, *.html*, *.txt* etc. Pentru o prelucrare unitară, s-a hotărât definirea unui format comun la care să fie aduse toate resursele partenerilor. Acest format a fost unul de tip XML şi el urma să păstreze doar puține informații relative la formatarea documentelor (precum fontul subliniat, înclinat sau îngroşat), adică atâtea câte s-ar putea dovedi utile în extragerea automată a cuvintelor cheie sau a definițiilor (de exemplu este foarte mare probabilitatea ca un cuvânt subliniat sau îngroşat să fie un termen cheie).

Deoarece colecția de documente din proiect proveneau, aşa cum s-a menționat, din nouă limbi diferite, fiecare cu convențiile proprii asupra setului de diacritice, s-a convenit asupra utilizării formatului UTF-8, care pare a fi cel mai potrivit păstrării unitare a unor colecții de documente multilingve în vederea unor prelucrări similare. Aducerea la formatul XML UTF-8 (notat Base-XML în Figura 1) a reprezentat, aşadar, primul nivel de prelucrare a documentelor primare. Până la transformarea lor în obiecte de învățare, acestea au fost suferit ulterior prelucrări lingvistice (nivelul 2 din Figura 1.), adnotări

[2] http://consilr.info.uaic.ro/uploads_lt4el/

[3] În etapele preliminare ale proiectului aceste cuvinte cheie au fost incluse în informațiile adăugate pentru fiecare document, pentru a ajuta la selecția domeniilor, deci implicit a documentelor, cu care se lucrează în etapele ulterioare. În prezent aceste cuvinte cheie selectate inițial manual nu mai sunt incluse în informația ataşată fiecărui document.

[4] Documentele vor putea fi făcute publice la sfârşitul proiectului, dar până atunci majoritatea lor au fost oferite de către autori doar pentru cercetare.

manuale şi automate la cuvinte cheie şi definiţii (nivelul 3) şi au fost plasate într-o ierarhie a schemelor de adnotare (Cristea et al., 2006).

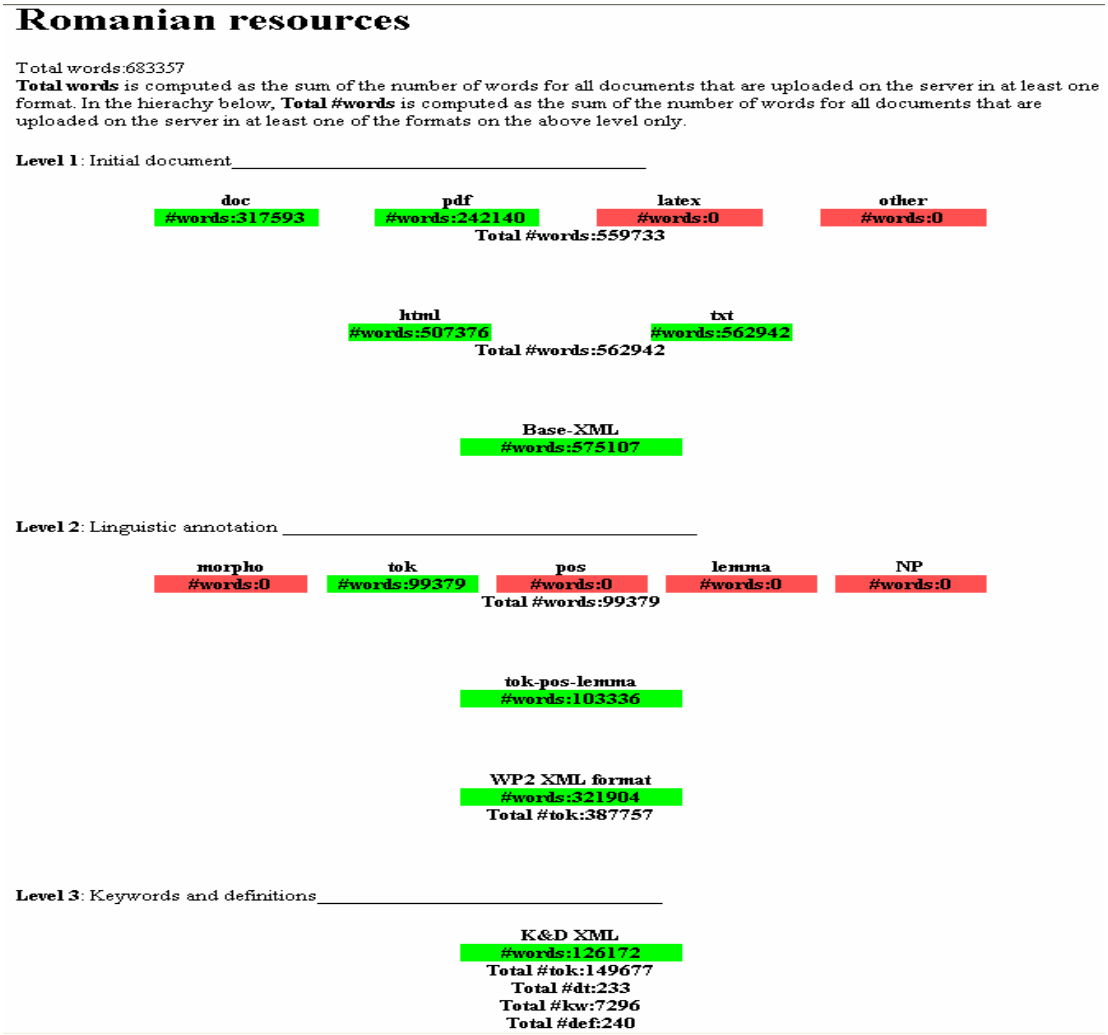


Figura 1: Reprezentarea diferitelor nivele de adnotare a resurselor româneşti.

3.1. Nivelul lingvistic

Procesarea lingvistică a documentelor este importantă pentru a permite extragerea unor informaţii morfologice ce pot ajuta în detectarea automată a cuvintelor cheie şi a definiţiilor. Pe acest nivel, resursele au fost adnotate, folosind serviciul Web pus la dispoziţie de ICIA[5], astfel: împărţirea în unităţi lexicale, marcarea informaţiei morfo-sintactice şi marcarea lemelor formelor flexionate (Ion, 2006).

Ţinând cont de contextul multilingv şi de faptul că fiecare partener a venit cu instrumente de prelucrare diferite, ce manifestă, în general, formate de ieşire diferite, s-a convenit şi pentru această etapă asupra unui format comun la care să se aducă adnotarea

[5] Institutul pentru Cercetări în Inteligenţă Artificială, Academia Română, http://www.racai.ro/

lingvistică. Astfel, există etichete pentru segmentarea în paragrafe, în propoziții şi la nivel de cuvânt. Fiecare cuvânt, de exemplu, trebuie marcat folosind etichete <tok>:

```
<tok rend="" base="Uniunii_Europene" ctag="Ed" id="t961">Uniunii_Europene</tok>
```

Atributele elementului *tok* sunt: *id*, un identificator unic; *rend*, care conține informația de formatare din formatul XML al primului nivelul, dacă ea există; *base*, care conține forma lematizată a intrării lexicale şi *ctag*, unde este trecută informația morfo-sintactică.

3.2. Nivelul post-lingvistic

Adnotările care au urmat nivelului lingvistic s-au făcut în două etape: o *adnotare manuală* şi o *adnotare automată*. Motivul acestei duble adnotări a fost acela de a permite apoi compararea lor în scopuri de evaluare a adnotării automate. În fiecare din aceste două etape s-a avut în vedere adnotarea la cuvinte cheie şi adnotarea la definiții.

3.2.1. Adnotarea cuvintelor cheie

Cuvintele cheie (unități lexicale formate fie dintr-un singur cuvânt, fie din expresii multi-cuvânt) sunt marcate (semi)automat de creatorii obiectelor de învățare; utilizatorii sistemului de învățare folosesc cuvintele cheie pentru a găsi documentele care conțin referi la anumite noțiuni. Din acest motiv cuvintele cheie trebuie să fie reprezentative pentru obiectul de învățare din care provine, să rezume subiectul textului sau să fie un obiectiv central al documentului. În adnotarea cuvintelor cheie s-a avut în vedere şi posibilitatea ca aceleaşi noțiuni să fie uneori referite prin sinonime în acelaşi text.

Un exemplu de adnotare a unui cuvânt cheie este:

```
<markedTerm id="k36" comment="" dt="n" kw="y" status="">
<tok rend="" base="Uniunii_Europene" ctag="Ed" id="t961">Uniunii_Europene</tok>
</markedTerm>
```

Cuvintele cheie sunt marcate cu <markedTerm>. Pentru că aceasta este o etichetă comună cu cea folosită pentru marcarea termenilor definiți dintr-o definiție, diferența dintre cele două adnotări este dată de atributele *dt* şi respectiv *kw*. Acestea pot lua valorile *y* (*yes*) şi *n* (*no*). Pentru exemplul de mai sus, valoarea atributului *dt* este *n*, ceea ce înseamnă ca sintagma nu este un termen definit în acest context, iar valoarea atributului *kw* este *y*, ceea ce înseamnă ca sintagma este aici un cuvânt cheie.

Celelalte atribute ale elementului <markedTerm> sunt un *id*, a cărui valoare trebuie să fie unică în document, *status*, un atribut de confirmare, care poate lua valoarea *?* sau *??* daca adnotatorul nu este sigur, respectiv este foarte nesigur, că ceea ce a marcat este corect, şi *comment,* care poate conține comentarii.

În ceea ce priveşte adnotarea automată, în proiect s-au implementat trei metode pentru extragere a cuvintelor cheie: TF/IDF, Residual IDF (RIDF) şi o versiune ajustată a RIDF (RIDF este înmulțit cu rădăcina pătrată a frecvenței termenilor). Programul generează un model de limbă, folosind fişierele adnotate manual, şi aplică acest model pe restul documentelor (Lemnitzer, Degórski, 2006). Momentan suntem în stadiul de validare a rezultatelor obținute de extractor.

3.2.2. Adnotarea definițiilor

Prin definiție se înțelege o explicație concisă a înțelesului unui cuvânt sau a unei sintagme, o descriere a înțelesului unui concept sau a tipului său. O definiție are două părți: elementul definit şi explicația propriu-zisă. Un exemplu de definiție extrasă din corpus este:

[Cetățenia Uniunii Europene] *DEF PART1* ***, prevăzută în tratatul de la Roma şi mai apoi în cel de la Maastricht [este caracterizată de drepturi, de obligații şi de implicarea în viața politică]*** *DEF_PART2*.

unde elementul definit este *Cetățenia Uniunii Europene*, iar definiția propriu-zisă este marcată între paranteze []. Se observă că atributiva care determină termenul definit nu a fost considerată ca făcând parte din definiție. Notarea definițiilor care au astfel de întreruperi în secvența textuală (formate din mai multe părți) este exemplificată mai jos:

```
<definingText comment="" id="def37" status="" continue="y" def="dt35" part="1">
  <markedTerm id="dt35" comment="" dt="y" kw="n" status="">
    <tok rend=" /b, /p, p" base="cet&#259;&#355;enie" ctag="Ncfsry" id="t960">
Cet&#259;&#355;enia </tok>
    <markedTerm id="k36" comment="" dt="n" kw="y" status="">
      <tok rend="" base="Uniunii_Europene" ctag="Ed" id="t961">Uniunii_Europene</tok>
    </markedTerm>
  </markedTerm>
</definingText>
<tok rend="" base="," ctag="COMMA" id="t962">, </tok>
<tok rend="" base="prevedea" ctag="Vmp--sf" id="t963">prev&#259;zut&#259;</tok>
<tok rend="" base="&#238;n" ctag="Spsa" id="t964">&#238;n</tok>
<tok rend="" base="tratat" ctag="Ncmsry" id="t965">Tratatul</tok>
<tok rend="" base="de_la" ctag="Spca" id="t966">de_la</tok>
<tok rend="" base="Roma" ctag="Np" id="t967">Roma</tok>
<tok rend="" base="(0.67)&#351;" ctag="Vmis1s" id="t968">&#351;i</tok>
<tok rend="" base="mai" ctag="Rp" id="t969">mai</tok>
<tok rend="" base="apoi" ctag="Rgp" id="t970">apoi</tok>
<tok rend="" base="&#238;n" ctag="Spsa" id="t971">&#238;n</tok>
<tok rend="" base="acela" ctag="Pd3msr" id="t972">cel</tok>
<tok rend="" base="de_la" ctag="Spca" id="t973">de_la</tok>
<tok rend="" base="Maastricht" ctag="Np" id="t974">Maastricht</tok>
<definingText comment="" id="def38" status="" continue="y" def="dt35" part="2">
  <tok rend="" base="fi" ctag="Vaip3s" id="t975">este </tok>
  <tok rend="" base="caracteriza" ctag="Vmp--sf" id="t976">caracterizat&#259;</tok>
  <tok rend="" base="de" ctag="Spsa" id="t977">de</tok>
  <tok rend="" base="drept" ctag="Ncfp-n" id="t978">drepturi</tok>
  <tok rend="" base="," ctag="COMMA" id="t979">,</tok>
  <tok rend="" base="de" ctag="Spsa" id="t980">de</tok>
  <tok rend="" base="obliga&#355;ie" ctag="Ncfp-n" id="t981">obliga&#355;ii</tok>
  <tok rend="" base="(0.62)&#351;" ctag="Ncmpry" id="t982">&#351;i</tok>
  <tok rend="" base="de" ctag="Spsa" id="t983">de</tok>
  <tok rend="" base="implicare" ctag="Ncfsrn" id="t984">implicare</tok>
  <tok rend="" base="&#238;n" ctag="Spsa" id="t985">&#238;n</tok>
  <tok rend="" base="via&#355;&#259;" ctag="Ncfsry" id="t986">via&#355;a</tok>
  <tok rend="" base="politic" ctag="Afpfsrn" id="t987">politic&#259;</tok>
</definingText>
```

Pentru elementul definit se foloseşte eticheta `<markedTerm>` cu atributul *dt="y"*, după cum s-a arătat în secțiunea anterioară. Pentru marcarea definiției este folosită eticheta `<definingText>` cu atribute asemănătoare etichetei `<markedTerm>`, dar având în plus atributul *part* pentru a marca definițiile întrerupte (valoarea acestui atribut este 0 dacă definiția este continuă, sau un număr începând cu 1 şi continuând sa crească pentru fiecare parte, dacă definiția are alte elemente intercalate) şi atributul *def* pentru a indica *id*-ul elementului definit, care ajută şi la unificarea definițiilor formate din mai multe părți. În exemplul de mai sus, definiția se referă la termenul definit cu *id*-ul *dt35*.

Adnotarea automată a definiţiilor se face utilizând o gramatică realizată de fiecare partener pentru limba respectivă. În afară de problemele care apar datorită dificultăţii de a surprinde toate modurile de exprimare a unei definiţii (mai ales dacă se evită lexicalizarea în exces), de a trata definiţiile întrerupte, de a stabili unde se termină o definiţie etc., mai apar şi probleme de ordin tehnic, care ţin de modul de redactare a documentelor. Astfel, în exemplul dat anterior, în textul de intrare virgula nu apare la sfârşitul atributivei, aşa cum cer normele gramaticale, lucru care poate crea dificultăţi unei reguli de detecţie a propoziţiilor intercalate.

4. Concluzii

În cadrul proiectului LT4eL s-au colectat 56 de documente care însumează peste 600.000 de cuvinte. Acestea au fost aduse la un format unitar XML şi adnotate la nivel lingvistic (segmentare în cuvinte, adnotare morfo-sintactică şi lematizare). Ulterior, o parte din ele a fost adnotată manual la cuvinte cheie şi definiţii. Momentan se lucrează la îmbunătăţirea rezultatelor obţinute cu extractorul automat de cuvinte cheie şi de definiţii şi la validarea acestora. Una din direcţiile de lucru viitoare implică adnotarea semantică a obiectelor de învăţare conform ontologiei dezvoltate în cadrul proiectului, proces deja început pentru limba engleză.

Referinţe bibliografice

Cristea, D., Forăscu, C., Pistol, I.. (2006). Requirements-Driven Automatic Configuration of Natural Language Applications. In Bernadette Sharp (Ed.): Natural Language Understanding and Cognitive Science, *Proceedings of the 3rd International Workshop on Natural Language Understanding and Cognitive Science - NLUCS 2006*, in conjunction with ICEIS 2006, Cyprus, Paphos, May 2006. INSTICC Press, Portugal. ISBN: 972-8865-50-3.

Ion, R. (2006). *Metode de dezambiguizare semantică automată. Aplicaţii pentru limbile engleză şi română*. Teză de doctorat în curs de susţinere la Academia Română.

Lemnitzer, L., Degórski, L. (2006): Language Technology for eLearning -- Implementing a Keyword Extractor. The fourth EDEN Research Workshop "*Research into online distance education and eLearning. Making the Difference*", 25-28 October, 2006 in Castelldefels, Spain

Monachesi, P., Cristea, D., Evans, D., Killing, A., Lemnitzer, L., Simov, K., Vertan, C. (2006). Integrating Language Technology and Semantic Web techniques in eLearning. *Proceedings of ICL* 2006.

Pistol, I., Trandabăţ, D., Iftene, A., Cristea, D., Forăscu, C. (2006). Prelucrarea resurselor româneşti în cadrul proiectului LT4eL. În acest volum.

TEHNICI DE VALIDARE ŞI CORECŢIE FOCALIZATĂ A ADNOTĂRII MORFO-SINTACTICE ÎN CORPUSURI DE MARI DIMENSIUNI

DAN TUFIŞ, ELENA IRIMIA

Institutul de Cercetări pentru Inteligenţă Artificială, Academia Română, Bucureşti

{tufis, elena}@racai.ro

Rezumat

Articolul descrie procesul de realizare şi corectare a RoCo-News, corpus jurnalistic pentru limba română, de dimensiune medie, abundent în nume proprii, numerale şi entităţi denumite. Textul, iniţial ne-procesat, a fost segmentat cu MtSeg, apoi adnotat morfo-sintactic cu tagger-ul TNT. Ulterior, RoCo-News a fost re-adnotat şi lematizat cu tagger-ul TTL realizat la RACAI, iar în final validat şi corectat. Datorită resurselor umane limitate, a constrângerilor temporale şi a dimensiunilor corpusului, validarea de mână a fiecărei unităţi lexicale a fost exclusă. Etapa de validare a necesitat o metodologie coerentă pentru identificarea automată a cât mai multe erori de adnotare morfo-sintactică şi lematizare. Procesul de validare manuală s-a concentrat apoi doar pe aceste posibile erori descoperite în mod automat.

1. Introducere. Descrierea corpusului RO-CO News

Pentru cercetătorii în domeniul lingvisticii computaţionale, importanţa dezvoltării resurselor, în special a corpusurilor, este evidentă. Reprezentând o "colecţie de fragmente de text într-o anumită limbă, selectate şi ordonate potrivit unor criterii lingvistice, în scopul de a fi utilizate ca mostre ale limbii respective" (Sinclair, 1991), un corpus ajută la formarea unei imagini comprehensive asupra limbii sincronice.

Un corpus jurnalistic este o bună sursă de informare asupra cuvintelor nou intrate în vocabularul unei anumite limbi, despre entităţile denumite, abrevierile comune şi multe alte aspecte ale stilului funcţional. În cele ce urmează, vom descrie corpusul RoCo-News, care, în forma sa actuală, reprezintă rezultatul unei munci de un an de validare şi corectare a procesării şi adnotării automate.

RoCo-News este un corpus jurnalistic pentru limba română de dimensiuni medii. Conţine aproximativ 7 milioane de unităţi lexicale, numărul de unităţi distincte depăşind 231.000. Diferitele articole din corpus, disponibile iniţial în diverse formate (doc, rtf şi pdf) au fost convertite în formatul ASCII, cu diacriticele codificate ca entităţi SGML.

O analiză preliminară a textului a relevat abundenţa numelor proprii şi a expresiilor numerice, date deloc surprinzătoare pentru un registru jurnalistic. Specific acestui tip de text, titlurile şi numele de autori apar ca paragrafe distincte, iar datorită structurii gramaticale parţiale pot conţine mai multe erori de adnotare decât paragrafele reale.

Textul neprelucrat a fost segmentat cu segmentatorul MtSeg, dezvoltat în contextul proiectului MULTEXT (http://aune.lpl.univ-aix.fr/projects/Multext/), bazându-ne pe

resursele de segmentare dezvoltate de proiectul MULTEXT-East (http://nl.ijs.si/ME/) şi adnotat morfo-sintactic folosind tagger-ul TnT (Brants, 1988). Tagger-ul a fost antrenat pe un corpus validat manual care include romanul "O mie nouă sute optzeci şi patru" al lui Orwell (aproximativ 110.000 unităţi lexicale), "Republica" lui Platon (aproximativ 140.000 de unităţi lexicale) şi mai multe articole din câteva jurnale cu acoperire naţională (aproximativ 140.000 de unităţi lexicale). Setul de etichete (tagset) al modelului de limbă este derivat dintr-un tagset mai mare, complet compatibil cu specificaţiile morfo-sintactice MULTEXT-East. Tagset-ul redus utilizat în corpusul RoCo-News este tagset-ul ascuns din metodologia de adnotare stratificată ("tiered tagging", vezi (Tufiş, 1999) şi (Tufiş şi Dragomirescu, 2004), pentru mai multe detalii). Tagset-ul redus conţine 93 de etichete pentru cuvinte şi 10 etichete pentru punctuaţie.

În continuare, corpusul adnotat a fost lematizat. Procesul de lematizare a fost, în esenţă, o procedură de căutare într-un lexicon de dimensiuni mari, conţinând peste 600.000 de intrări de forma: <formă-cuvânt> <lemă> <etichetă>. Pentru unităţile lexicale care nu se regăsesc în această resursă (şi care nu sunt marcate ca nume proprii), lema este furnizată de lematizorul cuprins în modulul statistic de adnotare şi lematizare TTL (Ion, 2006), realizat la RACAI. Aplicaţia foloseşte un set de reguli (specific fiecărei categorii gramaticale flexionare) induse automat din lexicon, care generează leme candidat pentru cuvântul necunoscut, şi apoi modele Markov (antrenate pe leme din lexicon) pentru a ordona candidaţii. Candidatul cu cea mai mare probabilitate câştigă. Procedura funcţionează foarte bine, cele mai multe dintre erori sunt în cazul cuvintelor necunoscute care aparţin paradigmelor flexionare neregulate (Tufiş, 1989) sau atunci când adnotarea morfo-sintactică a formei ocurenţă a fost greşită. În ansamblu, ţinând cont de vasta acoperire a lexiconului şi de rata mică de eroare a lematizorului statistic, probabilitatea unei erori de lematizare este neglijabilă.

2. Trei tehnici utilizate pentru identificarea erorilor posibile în RoCo-News

2.1. Lematizare şi re-segmentare

Lematizarea este un procedeu mai simplu decât adnotarea şi, de aceea, se poate face automat cu mai multă acurateţe. În mare parte dintre cazuri, perechea formată de ocurenţa unui cuvânt şi una dintre etichetele sale legale ar trebui să identifice în mod unic lema acelei unităţi lexicale, dacă ea este înregistrată în lexicon. Cum majoritatea cuvintelor dintr-un text nou sunt cuvinte de uz general, deci teoretic prezente în lexicon, putem presupune că lematizarea unui text folosind această procedură se face cu mare acurateţe. Lexiconul este actualizat în mod constant, pe măsură ce întâlnim cuvinte noi în textele la care lucrăm. Totuşi, lematizorul statistic apelat pentru cuvintele necunoscute poate produce leme greşite, în special dacă unităţile lexicale sunt adnotate în mod eronat. Pentru a nu introduce erori în lexicon, tripletele < **formă-cuvânt lemă etichetă** > sunt subiectul validării unui expert, înainte de a fi incluşi în lexicon.

Procedura de lematizare, re-segmentare şi corectare a erorilor identificate în decursul acestui proces este descrisă pe scurt în cele ce urmează:

a) Dacă unitatea lexicală curentă nu este marcată printr-un asterisc şi a fost adnotată cu o etichetă de semn de punctuaţie sau de categorie gramaticală fără flexiune, lema este identică formei ocurenţă.

b) Dacă unitatea lexicală curentă este marcată de tagger ca necunoscută, verificăm dacă eticheta morfo-sintactică este NP (nume propriu), caz în care lema este din nou considerată ca identică formei ocurenţă a lemei. Raţionamentul este că în limba română, numele proprii (cele străine şi cele masculine) sunt rareori flexionate. Pe de altă parte, numele proprii feminine pot avea flexiune, dar cele mai frecvente dintre ele se află deja în lexicon. Unităţile lexicale adnotate în mod consecutiv prin eticheta NP sunt concatenate şi considerate drept o singură unitate lexicală, iar lema sa este concatenarea lemelor unităţilor lexicale care o compun. Tripletele necunoscute <formă-cuvânt NP lemă> obţinute prin concatenare au fost adăugate în fişierul NumeProprii. Toate numele proprii din acest fişier au fost validate, iar erorile corectate. Corecţiile au fost operate şi în corpus. Câteva dintre erorile tipice au fost reprezentate de unităţile lexicale ale căror caractere erau în întregime scrise cu majuscule (făceau parte din titluri de articole) sau erau nume proprii independente, a căror concatenare nu era necesară.

c) Dacă o unitate lexicală nerecunoscută de TnT nu este etichetată ca şi NP, atunci este căutată în lexicon (mult mai mare decât lexiconul tagger-ului) împreună cu eticheta sa. În cazul în care este găsită, lema este copiată din respectiva intrare în lexicon. Altfel, lematizorul probabilistic este apelat pentru unitatea lexicală curentă iar tripletul <formă-cuvânt etichetă lemă> este salvat în fişierul denumit NuSuntÎnLexicon, pentru inspecţie şi validare ulterioară. Conţinutul acestui fişier a fost clasificat şi analizat în ordine descrescătoare a frecvenţei tripletelor sale.

Au fost identificate mai mult de 20.000 de erori, majoritatea datorate conversiei eronate a anumitor diacritice în entităţi SGML. O astfel de eroare sistematică, o dată observată, este relativ uşor de corectat. Un caz special de unităţi lexicale nerecunoscute de tagger este reprezentat de numere. Ele sunt adnotate în mod sistematic ca şi numerale, dar există numeroase cazuri în care segmentarea acestora a fost eronată, datorită utilizării ca separator între grupurile de trei cifre a caracterului spaţiu, în loc de virgulă sau punct. Astfel, segmentatorul a considerat că are de a face cu unităţi lexicale diferite. Pentru astfel de cazuri am procedat la concatenarea grupurilor de cifre.

Printre unităţile lexicale necunoscute am găsit de asemenea şi adrese web sau de e-mail. Aceste grupuri de cuvinte speciale au fost adnotate sistematic de către TnT ca şi NN. Importanţa de necontestat pe care aceste tipuri de unităţi textuale şi-au câştigat-o motivează decizia de a introduce în tagset două etichete noi: NNWEB şi NNMAIL. Toate ocurenţele adreselor de web şi e-mail au fost re-adnotate corespunzător acestui tagset extins.

2.2. Utilizarea analizei clasei închise pentru identificarea erorilor

Divizarea categoriilor lexicale în două tipuri diferite de clase este tradiţională în lingvistică: clasele închise sunt acelea enumerabile (ex: clase precum determinatori, prepoziţii, verbe modale sau auxiliare), în timp ce clasele deschise sunt categoriile mari şi productive precum verbele, substantivele şi adjectivele.

(Dickinson şi Meurers, 2002) au exploatat ideea că, pentru detectarea erorilor, se poate utiliza în mod practic conceptul de clasă închisă. Poate fi uşor de observat, susţin ei, că aproximativ jumătate din etichetele oricărui tagset corespund claselor lexicale închise. O categorie lexicală clasă închisă conţine un număr redus de cuvinte, enumerabil. În mod frecvent, aceste cuvinte pot face parte din mai multe categorii clase închise (ex.: în limba română există cuvinte care pot fi în acelaşi timp: prepoziţii sau conjuncţii; prepoziţii şi auxiliare etc.). În funcţie de granularitatea tagset-ului, o categorie clasă închisă poate acoperi mai multe etichete (ex.: tipuri de conjuncţii, prepoziţii, pronume). Ţinând cont de frecvenţa mare a cuvintelor clase închisă, pentru un corpus mare putem presupune că aceste tipuri de cuvinte apar în cele mai multe (dacă nu în toate) dintre contextele posibile şi este, deci, de aşteptat ca toate etichetele acestora sa fie regăsite în corpus. Bazându-ne pe aceste consideraţii, am decis să facem o serie de teste cu privire la cuvintele clasă închisă din RoCo-News. Astfel, am extras din lexicon o listă L1, de etichete clasă închisă, fiecare dintre ele indexând mulţimea cuvintelor care ar putea să primească acea etichetă. Din această listă, am calculat o alta, L2, conţinând cuvinte din L1 ce indexează două sau mai multe etichete de clasă închisă. Apoi am extras din RoCo-News toate perechile <*cuvânt, etichetă*> a.î. eticheta să fie de clasă închisă. Dacă *cuvânt* nu a fost în mulţimea din L1 indexată de *etichetă,* am verificat ocurenţa respectivului cuvânt în context. În marea majoritate a cazurilor, am găsit o eroare de adnotare, dar ocazional am descoperit erori şi în lexicon (o posibilă etichetă de clasă închisă nu a fost înregistrată pentru anumite cuvinte). Bazându-ne pe L2, am extras toate cuvintele care au fost văzute în corpus doar cu un subset al etichetelor de clasă închise posibile. Au fost descoperite din nou anumite erori în lexicon (cuvinte care erau în mod eronat asociate cu anumite etichete de clasă închisă).

2.3. Utilizarea evaluării auto-referenţiale pentru o mai bună identificare a erorilor

Cea de-a treia tehnică utilizată în corectarea corpusului RoCo-News se bazează pe ipoteza evaluării auto-referenţială (Tufiş, 1999), care spune că un corpus consistent şi corect adnotat, re-adnotat cu modelul de limbaj învăţat din el însuşi (evaluare auto-referenţială), ar trebui să aibă majoritatea unităţilor lexicale adnotate în mod identic. Procentul de etichete identice depinde de dimensiunea corpusului, dar de obicei este mai mare de 97.5%-98%.

După ce am efectuat corecturile descrise în secţiunile precedente, am luat această versiune ca referinţă pentru procedura de evaluare auto-referenţială descrisă în cele ce urmează. Am antrenat tagger-ul TnT pe corpusul referinţă, construind un nou model de limbă. Am re-adnotat RoCo-News cu acest nou model de limbă şi am comparat noua adnotare cu cea de referinţă. Am descoperit 96.8% unităţi lexicale adnotate identic şi am extras diferenţele. Sortând diferenţele în ordine inversă a frecvenţei lor în corpus, am examinat pe rând, în context, primele 100 tipuri de diferenţe (reprezentând aproximativ 8-10.000 de ocurenţe diferite) iar expertul însărcinat cu validarea lor a stabilit care dintre cele două etichete era cea corectă (dacă vreuna dintre ele a fost corectă). Unele dintre diferenţe au fost explicate prin inconsistenţa sau incompletitudinea corecturilor din etapele precedente. Alte diferenţe au apărut deoarece corecturile au modificat contextele pentru unităţile lexicale vecine şi astfel, conform modelului de limbă auto-referenţial, multe dintre unităţi au apărut în contexte diferite şi au primit etichete diferite. Corectarea

tuturor erorilor descoperite în analiza primelor 100 tipuri de diferențe încheie procedura. Având în vedere dimensiunea corpusului, realizarea ei necesită foarte mult timp. Procedura a fost repetată de mai multe ori, cu o scădere continuă a numărului de diferențe; la final, numărul de diferențe s-a stabilizat la 0.4% din dimensiunea corpusului (rămânând 25.500 diferențe dintre care 6353 distincte).

3. Concluzie

Am descris o procedură semi-automată prin ale cărei mijloace am construit un corpus jurnalistic pentru limba română cu înalt nivel de acurateţe a adnotării şi lematizării. Deşi rezultate analizelor şi natura erorilor sunt dependente de limbă, de tagger şi de tagset, scheletul acestei abordări poate fi adaptat şi aplicat cu uşurinţă într-un alt context. Tipul de analiză pe care am descris-o poate oferi indicaţii importante despre cuvintele/tipurile de cuvinte care ar putea fi adnotate nesatisfăcător într-un alt corpus din acelaşi registru.

Metoda nu asigură eliminarea tuturor erorilor existente, dar câştigul în acurateţe este substanţial iar faptul că nu este nevoie de examinarea cuvânt cu cuvânt constituie, din punct de vedere al economiei de timp şi de resurse umane, un mare avantaj, expertul care validează putându-se concentra pe acele erori pe care procedurile le evidenţiază ca fiind frecvente şi/sau importante.

Referinţe bibliografice

Brants, T. (1998). TnT - A Statistical Part-of-Speech Tagger. Instalation and User Guide, University of Saarland, *Computational Linguistics*, March 1998.

Dickinson, M., Meurers, W. Detmar (2003). Detecting Errors in Part of Speech Annotation. In *Proceedings of the 11th conference of the EACL-03*, Budapest, Hungary.

Ion, R. (2006). *Metode de dezambiguizare semantică automată. Aplicaţii pentru limbile engleză şi română*, Teză de doctorat în curs de susţinere la Academia Română, Bucureşti, România, 145 p.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*, Oxford University Press.

Tufiş, D. (1989). It Would Be Much Easier If *WENT* Were *GOED*., Harry Somers, Mary McGee Wood (eds.), *Proceedings of the 4th European Conference of the Association for Computational Linguistics*, Manchester.

Tufiş, D. (1999). Tiered Tagging and Combined Classifiers. F. Jelinek, E. Nöth (eds) Text, *Speech and Dialogue, Lecture Notes in Artificial Intelligence* 1692, Springer, 1999, pp. 28-34.

Tufiş, D., Dragomirescu, L. (2004). Tiered Tagging Revisited. *Proceedings of the 4th LREC Conference.* (pp. 39—42). Lisbon, Portugal.

ROGER – UN CORPUS PARALEL ALINIAT

MONICA GAVRILĂ, NATALIA ELIȚA

Departamentul de Informatică, Facultatea de Matematică, Informatică şi Ştiinţele Naturii, Universitatea din Hamburg

{gavrila, elita}@informatik.uni-hamburg.de

Rezumat

Colecţiile de texte în format electronic (corpusurile) sunt folosite foarte des în comunitatea procesării limbajului natural. Deoarece este o resursă primară, alegerea unui corpus potrivit influenţează direct rezultatul pe care utilizatorul doreşte să îl obţină şi, implicit, prin folosirea acestui rezultat, cercetările şi rezultatele ulterioare.

Această lucrare prezintă un corpus multilingv (**ro**mân-**g**erman-**e**nglez-**r**us), aliniat şi paralel, de aproximativ 2300 propoziţii.

Lucrarea este organizată în patru secţiuni. Prima secţiune descrie importanţa unui corpus în prelucrarea limbajului natural (PLN), cu accente asupra traducerii automate bazate pe exemple. Cea de-a doua secţiune prezintă resurse existente şi problemele întâmpinate în găsirea unui corpus adecvat, iar cea de-a treia informaţii asupra corpusului creat: RoGER. Concluziile şi lista de referinţe bibliografice încheie lucrarea de faţă.

1. Importanţa corpusului în PLN şi în traducerea automată bazată pe exemple

Colecţiile de texte în format electronic (corpusurile) sunt foarte des folosite în comunitatea procesării limbajului natural (PLN). Deoarece este o resursă primară, alegerea unui corpus potrivit influenţează direct rezultatul pe care utilizatorul doreşte să îl obţină şi, implicit, prin folosirea acestui rezultat, cercetările şi rezultatele ulterioare. De aceea, câteva aspecte sunt foarte importante:

- dimensiunea corpusului;

- tipul textului;

- relevanţa textului;

- tipul corpusului (monolingv / multilingv, aliniat / nealiniat, etc.)

Corpusul este folosit atât pentru antrenarea unui sistem, cât şi pentru testarea lui.

Corpusurile sunt utilizate pentru rezolvarea diverselor probleme din PLN: traducerea automată, analiza textelor cu metode statistice, dezambiguizarea sensurilor cuvintelor, construirea modelelor limbilor, etc.

În ceea ce priveşte traducerea automată bazată pe exemple, aceasta are la bază un corpus ce trebuie să îndeplinească anumite condiţii legate de dimensiune, tipul textului, etc.

Pentru a elabora un astfel de sistem, în primul rând este nevoie de un corpus paralel aliniat[1]. Din acest corpus sunt extrase exemple de traducere, care vor fi folosite ulterior la traducerea unor texte noi. Aceste exemple pot fi salvate în formate diferite (de exemplu: arbori, şabloane – *engl:* „templates", etc), iar în funcție de modul de salvare câteodată este nevoie de adnotarea corpusului.

Pentru extragerea automată de exemple este necesar a avea corpusul, şi informația aferentă lui, într-un format uşor de accesat şi procesat de calculator.

În traducerea automată bazată pe exemple, o mare importanță o are balanța corectă între lungimea şi similaritatea exemplelor din baza de exemple: cu cât exemplele sunt mai lungi, cu atât este mai greu să găseşti un exemplu potrivit pentru textul ce urmează a fi tradus, şi cu cât exemplele sunt mai scurte, cu atât creşte posibilitatea de ambiguitate.

2. Resurse existente

Aplicația pe care dorim să o realizăm, şi pentru care avem nevoie de un corpus, este un sistem de traducere automată bazată pe exemple, în care este specificată informația semantică, prin adnotarea semantică a corpusului. Adnotarea semantică se va face având la bază o ontologie.

În cercetarea noastra, inițial, am încercat să găsim un corpus care să îndeplinească cerințele traducerii automate bazate pe exemple: paralel şi aliniat. În plus, doream ca acesta să fie în patru limbi: român-englez-german-rus, şi domeniul descris în text să fie restricționat. Am dorit să avem cu corpus al cărui domeniu este restricționat pentru a uşura munca necesară realizării ontologiei.

Din resursele existente, analizate de noi, fac parte şi cele din tabelul de mai jos:

Tabel 1. O parte a corpusurilor analizate

	Denumire	**Conținut**	**Observații**
1	Corpusul paralel Român-Englez (Rada Mihalcea www.cs.unt.edu/~rada/downloads.html)	Arhive de ziare	Traduceri incomplete, nu sunt toate cele patru limbi
2	Corpus Român-Englez-Rus (www.azi.md)	Colecție de ştiri	Traduceri incomplete, nu sunt toate cele patru limbi
3	Corpus paralel German-Englez (www.iccs.informatics.ed.ac.uk/~pkoehn/publications/de-news)	Colecție de ştiri	Traduceri incomplete, nu sunt toate cele patru limbi
4	JRC-Acuis (wt.jrc.it/lt/Acquis), detalii în (Steinberger et al., 2006)	Colecție de texte din legislatia UE (1950-2005)	Lipseşte rusa

[1] Un corpus paralel este un corpus în două sau mai multe limbi, în care textele dintr-o limbă sunt traduse în celelalte limbi. Un corpus paralel aliniat este un corpus paralel în care este realizată corespondența dintre traduceri.

	Denumire	**Conținut**	**Observații**
5	OPUS (logos.uio.no/opus/kdedoc.html)	Documentații, manuale	Traduceri parțiale

Din analiza acestor resurse, unele dintre problemele întâlnite au fost:

- informația eronată asupra conținutului textului;

- traduceri incomplete;

- traduceri incorecte;

- domeniul corpusului este prea extins,

- corpusul nu conține toate cele patru limbi dorite, etc.

3. RoGER

Motivația noastră de a crea RoGER constă în faptul că resursele existente şi descrise în secțiunea anterioară nu corespund în totalitate cerințelor noastre asupra corpusului, din diverse motive: limbile considerate, traduceri inexacte, domeniul corpusului, etc.

RoGER este un corpus:

- paralel,

- aliniat la nivel de propoziție,

- specializat (domeniu tehnic) - textele sunt preluate dintr-un manual de utilizare a unui aparat electronic,

- multiligv: român – german - englez – rus,

- realizat în proporție de peste 80 % manual,

- neadnotat (la nivel semantic, sintactic sau morfologic),

- în care diacriticele sunt neglijate.

Textul inițial a fost procesat, în sensul că unele noțiuni au fost înlocuite cu "meta-noțiuni", astfel că: numerele au fost înlocuite cu *NUM*, denumirile de pagini web cu *WWWSITE*, imaginile cu *PICT*. De asemenea, pentru a uşura procesul de traducere automată bazată pe exemple (exemple salvate ca şabloane), unele abrevieri au fost extinse.

Pentru alinierea la nivel de propoziție şi corectarea traducerilor a fost efectuată o verificare manuală a corpusului. Verificarea şi corectarea traducerilor s-au realizat în momentul creării alinierii corpusului.

În tabelul de mai jos se pot găsi câteva date statistice referitoare la corpus:

Tabel 2. RoGER - statistici

	Engleză	**Germană**	**Română**	**Rusă**
Dimensiune corpus (propoziții)	2333	2333	2333	2333
Dimensiune corpus (cuvinte)[2]	26096	25850	27142	22383
Dimensiune vocabular	2012	3104	3031	3883
Vocabular (cu nr. de apariții mai mare ca doi)	1231	1575	1698	1904
Lungimea medie a propoziției	11	11	11	9

Referitor la dimensiunea şi scopul pentru care a fost creat (traducerea automată bazată pe exemple), RoGER poate fi caracterizat, conform datelor menționate în (Somers, 1999), ca un corpus de dimensiune medie.

El se situează la jumătatea listei (compusă din 30 exemple) menționate în (Somers, 1999), deasupra corpusului folosit în sistemul Gaijin (Veale and Way, 1997) - 1836 exemple.

Corpusul este salvat în format XML. Mai jos se găseşte un exemplu:

```
<?xml version="1.0" encoding="UTF-8"?>
<sentences>
.................
  <sentence id="1010">
   <en>Press Options and some of the following options may be
available .</en>
  <de>Druecken Sie Optionen . und einige der folgenden Optionen
sind ggf. verfuegbar .</de>
   <ro>Apasati Optiuni dupa care unele din urmatoarele optiuni
pot fi disponibile .</ro>
   <ru>Нажмите Вар-нты и выберите одну из перечисленных ниже
функций .</ru>
  </sentence>
.....................
</sentences>
```

În tabelul 3 sunt incluse câteva exemple de cuvinte foarte frecvente în corpus, şi analizându-le, ne putem da seama că în mare parte ele sunt aceleaşi în toate limbile considerate. Multe dintre ele sunt prepoziții, conjuncții (şi (ro), und (ge), and (en), и (ru)), dar sunt şi cuvinte purtătoare de sens - *engl.*:"content words" - (selectați (ro), waehlen (ge), select (en), выберите (ru)).

[2] Cuvintele se numără la nivel de şir de caractere („selectat” şi „selectate” se numără ca şi două cuvinte diferite).

Tabel 3. Cuvintele cele mai frecvente în RoGER

Română	Germană	Engleză	Rusă
de, şi, pentru, în, la, selectați, apăsați, este, un, dacă, să, pe, nu, o, care, Dvs, pagina, meniu, din, setari	Sie, und, die, der, waehlen, druecken, das, oder, um, zu, den, wenn, auf, fuer, auf, von, in	the, to, and, select, a, in, press, or, you, for, of, is, on, your, service, settings, menu	и, в, для, нажмите, на, выберите, или, Меню, Если

Cuvintele cele mai frecvente sunt:

- română: "de" (1459 ori)

- engleză: "the"(2075 ori)

- germană:"Sie" (1677 ori)

- rusă: "и" (799 ori)

Pentru a fi folosit ulterior în sistemul de traducere automată bazată pe exemple, asupra căruia lucrăm, intenționăm să extindem corpusul până la (minim) 2600 propoziții, introducând artificial ambiguități[3], şi să îl adnotăm semantic.

4. Concluzii

RoGER, corpusul realizat de noi - multilingv, paralel, aliniat - reprezintă o resursă utilă nu numai în antrenarea şi testarea unui sistem de traducere automată bazată pe exemple, ci şi pentru alte aplicații de PLN.

Referințe bibliografice

Steinberger R., Pouliquen B., Widiger A., Ignat C., Erjavec T., Tufiş D., Varga D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. *LREC'2006*, pag. 2142-2147. Genoa, Italia, 24-26 mai.

Herold S. (1999). Review Article: Example-based Machine Translation. *Machine Translation* 14: 113 – 157.

Tony V., Way W. (1997). Gaijin: A Template-Driven Bootstrapping Approach to Example-Based Machine Translation. *NeMNLP'97. New Methods in Natural Language Processing,* Sofia, Bulgaria, septembrie.

[3] Pentru a putea demonstra utilitatea semanticii in traducerea automată bazată pe exemple.

TIMEBANK 1.2:
O VERSIUNE ADNOTATĂ ÎN LIMBA ROMÂNĂ

CORINA FORĂSCU[1, 2], RADU ION [2]

[1]*Facultatea de Informatică, Universitatea "Al.I.Cuza", Iaşi*
[2] *Institutul de Cercetări pentru Inteligență Artificială, Academia Română, Bucureşti*

corinfor@info.uaic.ro, radu@racai.ro

Rezumat

Informația temporală s-a dovedit foarte relevantă mai ales în Prelucrarea Limbajului Natural. De proveniență lingvistică, teoriile temporale au fost studiate şi formalizate cu predilecție pentru limba engleză. Lucrarea prezintă cercetările efectuate pentru obținerea corpusului paralel TimeBank, englez-român, care să fie folosit, printr-un import automat, la validarea acestor teorii pentru limba română. Corpusul va constitui şi baza de lucru pentru dezvoltarea unor instrumente de prelucrare temporală a limbajului natural.

1. Introducere

Teoria logicii temporale s-a dovedit extrem de relevantă încă din anii '70, mai ales în Inteligența Artificială. Posibilitatea de a identifica şi analiza informația temporală este de mare importanță pentru multe dintre aplicațiile Prelucrării Limbajului Natural precum rezumarea multi-document, sistemele de întrebare răspuns, structura temporală a discursului, regăsirea şi extragerea informației, traducere automată, etc.

Dacă pe plan internațional în domeniul marcării informației temporale, s-au dezvoltat atât standarde de adnotare precum Timex2 (Ferro et al., 2005a) şi TimeML (Sauri et al., 2006), cât şi corpusuri adnotate conform cu acestea, precum ACE-TERN (Ferro et al., 2005b) sau TimeBank (Pustejovsky et al., 2006), predominant pentru limba engleză, pentru limba română cercetări anterioare (Forăscu, Solomon, 2004) au arătat că o adnotare manuală a unui corpus în limba română este foarte costisitoare, atât ca timp cât şi ca resurse umane implicate şi, mai mult, adnotările sunt deseori incomplete.

Lucrarea prezintă activitățile de creare ale unui corpus paralel englez-român, folosind ca sursă corpusul englez de ştiri TimeBank, pentru ca adnotarea temporală din acesta să fie apoi transferată în varianta română a corpusului, obținându-se astfel atât un corpus paralel cât şi o sursă exemplificată de fundamentare a teoriilor temporale pentru română.

Secțiunea a doua a lucrării motivează necesitatea unui corpus paralel adnotat temporal, incluzând şi fundamente ale informației temporale şi ale principalului standard de adnotare temporală. În continuare sunt prezentate corpusul englezesc, modul de realizare a variantei româneşti a acestuia (secțiunea 4), prelucrările realizate asupra corpusului paralel (secțiunea 5). În încheiere se dezvăluie câteva obiective de viitor ale cercetării.

2. Necesitatea unui corpus paralel, adnotat temporal

Cînd un corpus este privit ca o colecție de documente selectate şi ordonate conform unor criterii lingvistice stabilite, el permite punerea în evidență, informarea şi fundamentarea unor teorii lingvistice specializate. Cum la ora actuală engleza este limba cu cea mai densă realizare de documente adnotate, ea este deseori utilizată ca sursă din care să se transfere adnotări specifice asupra altor limbi. Pentru limba română o serie de corpusuri paralele au fost deja create (Cristea, Forăscu, 2006), însă niciunul care să permită evidențierea informației temporale în limbajul natural.

2.1. Informație şi adnotare temporală

Cf. (Mani et al., 2005), informația temporală este reprezentată în limbajul natural prin:

- expresii temporale exprimate prin grupuri nominale, prepoziționale sau adverbiale – ore (timp al zilei), date, durate: *acum trei ore, mai 1984, anii 90, 5 februarie 2007*, etc.; aceste expresii temporale referă timpul ca:
    - punct (moment): Am deschis uşa la *ora doisprezece*.,
    - interval: *Am fost plecat ieri.*
- expresii ce denotă evenimente exprimate în principal, pe lângă adjective, clauze predicative sau grupuri frazale prepoziționale, prin:
    - propoziții, mai exact prin centrul (eng. *head*) sintactic, anume verbul principal: *Ion a plecat la munte.*
    - grupuri nominale: *Greva va continua şi în zilele următoare.*

Expresiile ce denotă timpul pot avea:

- referințe explicite (specificate), care referă la o intrare într-un sistem calendaristic / orar: *amiază*, *11.10.2006 (midday, 11.10.2006)* ;
- referințe implicite (sub-specificate) - pot fi evaluate doar prin intermediul unui timp indexat: *anul viitor, săptămâna trecută, acum două ore*;
- referințe vagi (nespecificate, neancorate), care nu pot fi corelate cu un punct sau interval exact de timp: *după-amiază, în câteva săptămâni, acum câteva zile*.

Evenimentele exprimate prin verbe pot fi temporal ancorate:

- indirect, prin categoria morfologică a timpului şi
- direct, prin modificatori adverbiali (adverbe de timp şi frecvență, grupuri nominale şi prepoziționale şi clauze subordonate).

Pentru a codifica toate tipurile de expresii temporale, evenimente şi relații între acestea, a fost creat standardul TimeML (Pustejovsky et al., 2006), ale cărui fundamente s-au pus încă din 2002. Standardul reuneşte multe dintre eforturile anterioare de adnotare temporală, diferind de acestea prin separarea reprezentării evenimentelor şi a expresiilor temporale de legăturile de ancorare, ordonare sau dependență ce apar în texte.

Standardul TimeML defineşte 7 etichete: `EVENT`, `MAKEINSTANCE` (pentru evenimente şi instanțe ale acestora – doar instanțele vor participa în legături temporale), `TIMEX3` (pentru expresii temporale de tip `DATE, TIME, DURATION, SET`, complet specificate, non- şi sub-specificate), `SIGNAL` (pentru elemente lexicale de legătură) şi `TLINK`, `ALINK`, `SLINK` (pentru legături temporale, aspectuale şi respectiv de subordonare între expresii şi evenimente).

3. Corpusul TimeBank – versiunea engleză

Realizarea corpusului TimeBank a început în 2002 în cadrul proiectului TERQAS[1]. Corpusul conține în versiunea actuală 183 de fișiere de rapoarte de știri în limba engleză, adnotate conform cu TimeML v.1.2. (Pustejovsky et al., 2006). Documentele provin din evaluarea rezumatelor DUC 2001 şi din corpusurile ACE incluse în cataloagele LDC2003T11 şi LDC99T42. Documentele conțin şi alte marcaje XML: formatul documentelor, informație structurală, nume de entități (`ENAMEX`, `NUMEX` din MUC7), marcaje de propoziție.

Adnotarea temporală inițială a corpusului este considerată „preliminară" întrucât s-a arătat (Boguraev, Ando, 2006) că apar greşeli sistematice datorate dimensiunii relativ reduse a corpusului şi datorate inconsistențelor în adnotare: legături temporale sau de subordonare inconsistente sau incomplete, clasificarea evenimentelor – în perfectare, adnotare incompletă a timpului şi aspectului unor evenimente.

TimeBank 1.2. este versiunea actuală – din 2006[2] - a corpusului, conformă cu specificațiile TimeML 1.2.1. Structura şi adnotarea corpusului sunt, în esență, aceleaşi cu cele din prima versiune a corpusului. TimeBank 1.2 este distribuit prin LDC (Pustejovsky et al., 2006). Statisticile pe TimeBank 1.2 sunt ilustrate în Tabelul 1.

Adnotarea documentelor a început cu o fază de preprocesare, când unele articole lexicale de tip evenimente (`EVENT`) şi semnale (`SIGNAL`) au fost marcate cu unele clase, timpuri sau aspecte ale acestora. După această etapă 5 adnotatori umani au verificat preprocesările şi corectitudinea adnotărilor conforme cu specificația TimeML 1.2.1.

Tabel 1: Statistici asupra corpusului TimeBank 1.2

TimeML tags	#
events	7935
instances	7940
timexes	1414
signals	688
alinks	265
slinks	2932
tlinks	6418
TOTAL	**27592**

General	#
propoziții	4715
unități lexicale	61042
unități lexicale unice	10586

Corpusul TimeBank este în revizie continuă, pentru următoarele distribuții avându-se în vedere: evenimentele compuse, legăturile dintre argumente, evenimentele generice, relațiile temporale între data creării documentului şi evenimentele de tip `REPORTING`, o distincție mai clară între data creării şi data publicării unui articol de ştiri.

4. Crearea corpusului pentru limba română

Textul englezesc a fost repartizat inițial în vederea traducerii la două masterande în Lingvistică Computațională, Facultatea de Informatică Iaşi, cu un set minimal de recomandări, pentru a obține traduceri unitare şi alinieri satisfăcătoare cu originalele.

Ori de câte ori a fost posibil, traducerile au fost unu la unu. Alinierea la fraze/propoziții s-a obținut astfel direct prin notările care au marcat traducerile. S-a recomandat folosirea

[1] *Temporal and Event Recognition for Question Answering Systems*, http://www.timeml.org/site/terqas/index.html
[2] http://www.timeml.org/site/timebank/timebank.html

echivalenţilor de traducere cu aceeaşi parte de vorbire, cuvintele româneşti trebuind să fie cât mai "apropiate" de corespondentele lor englezeşti: atunci când cuvântului englezesc îi poate fi asociat în româneşte un cognate, acesta va fi preferat unei expresii (*sporadic -> sporadic* şi nu *mai rar*). S-au tradus toate cuvintele şi nu s-au introdus în traducere, din motive stilistice, cuvinte sau expresii fără corespondent în engleză. S-a folosit scrierea cu diacritice, conformă cu normele lingvistice în vigoare. Timpurile verbelor s-au păstrat pe cât posibil, modificările fiind acceptate doar pe temeiuri lingvistice, nu stilistice. S-a păstrat formatul din engleză pentru date, momente ale zilei şi numere.

Varianta actuală pentru limba română a fost verificată manual, urmărindu-se evitarea unor inconsistenţe şi lipsuri în traducere, care nu ar fi permis o aliniere a unor elemente temporale esenţiale. În cele 4.715 propoziţii sunt 65.375 unităţi lexicale (inclusiv semne de punctuaţie), din care 12.640 sunt unice.

5. Prelucrări ale corpusului paralel

5.1. Adnotări ale corpusurilor TimeBank englez şi român

În vederea alinierii lexicale a celor două jumătăţi ale corpusului, s-a utilizat o preadnotare unitară a textelor care să poată fi folosită de aliniatorul lexical YAWA (Tufiş et al., 2006). Această procesare preliminară se referă la segmentarea la nivel de cuvânt, adnotarea cu etichete morfosintactice şi lematizarea textelor în engleză şi română. Modulul TTL (Ion, 2006) oferă aceste adnotări şi în plus, asigură o reprezentare uniformă a textelor adnotate în termenii codificării corpusului paralel într-un format XML similar cu formatul XCES (Ide et al., 2000).

Segmentarea la nivel de cuvânt trebuie să ia în calcul faptul că spaţiul nu este singurul delimitator de cuvinte şi nici nu este întotdeauna delimitator de cuvinte. Atât în engleză cât şi în română există expresii idiomatice care trebuie considerate ca unităţi lexicale în procesul de aliniere (*bun_simţ*, *take_a_look*). Adnotarea morfosintactică se face cu ajutorul unui adnotator probabilistic care implementează adnotatorul TnT (Brants, 2000) bazat pe Modele Markov Ascunse. Setul de etichete morfosintactice este compatibil cu specificaţiile MULTEXT-East[3] fapt care permite reprezentarea uniformă a informaţiei morfosintactice în engleză şi în română.

Lematizorul implementat în TTL este de asemenea unul probabilistic. O lemă candidată se generează pe baza unei mulţimi de reguli extrase automat dintr-un lexicon care conţine pentru fiecare formă ocurenţă a unui cuvânt, lema şi eticheta morfosintactică a acesteia. Lema unei noi forme ocurenţe a unui cuvânt de o etichetă morfosintactică dată este lema cea mai probabilă dintre toate lemele candidate după Modelul Markov al tuturor lemelor de aceeaşi etichetă din lexicon (Ion, 2006).

Tot ca o cerinţă a alinierii lexicale, s-au recunoscut, folosind expresii regulate peste secvenţe de etichete morfosintactice, grupuri nominale şi prepoziţionale nerecursive, compuşi verbali (*s-a dus*), adjectivali (*cea mai frumoasă*) şi adverbiali (*tare de tot*).

[3] http://nl.ijs.si/ME

5.2. Alinierea lexicală a corpusurilor

Alinierea lexicală a corpusului paralel a fost realizată cu YAWA (Tufiş et al., 2006) pe ieşirea modulului TTL. Corpusul paralel TimeBank 1.2 a fost aliniat la nivel de unitate lexicală din română în engleză urmându-se patru faze specifice acestui aliniator:

1. alinierea cuvintelor conținut (substantive, verbe, adjective şi adverbe) folosind un dicționar de echivalenți de traducere extras automat (Tufiş, 2002);
2. pe scheletul de aliniere de la pasul anterior se aliniază cuvintele aflate în acelaşi grup sintactic cu cuvintele aliniate utilizându-se reguli de aliniere. De exemplu, dacă avem un substantiv românesc aliniat la unul englezesc care este precedat de un determinator, aliniază determinatorul englezesc la substantivul românesc;
3. pe scheletul de aliniere de la pasul 2, aliniază toate blocurile de indecşi consecutivi care au rămas nealiniați (Tufiş et al., 2006);
4. corectează alinierea de la 3.

Fazele 2 şi 4 sunt evident dependente de perechea de limbi aliniate dar regulile de aliniere şi cele de corecție nu sunt integrate în corpul aliniatorului astfel încât să poată fi schimbate atunci când se doreşte alinierea altei perechi de limbi.

6. Obiective viitoare

În vederea obținerii unui transfer optim al adnotărilor temporale din limba engleză, corpusul paralel aliniat este în prezent validat manual în proporție de 60%. O aliniere perfectă va fi folosită în continuare atât pentru îmbunătățirea performanțelor aliniatorului, cât şi la importul adnotărilor TimeML în varianta română a corpusului. După validarea manuală a acestui import, liste de activatori[4] (*eng. trigger*) lexicali vor fi extrase pentru a fi folosite, eventual în combinație cu metode specifice de învățare automată, pentru crearea şi antrenarea unui adnotator temporal pentru limba română. Pentru evaluarea adnotatorului se vor avea în vedere şi alte domenii pe lângă cel de ştiri, precum beletristică, legislație etc. Ca planuri de lungă durată se pot menționa folosirea adnotărilor temporale combinate cu cele de discurs pentru determinarea structurii temporale a discursului, rezumarea multi-document şi folosirea ontologiilor temporale pentru a obține inferențe despre evenimente în timp.

Mulțumiri. Autorii sunt recunoscători Ministerului Educației şi Cercetării, de a cărui finanțare au beneficiat în cadrul proiectului CEEX 29 ROTEL şi CEEX 132 InterOb. Pentru sfaturile şi sprijinul primit, autorii mulțumesc coordonatorului comun de doctorat, prof. dr. Dan Tufiş, precum şi prof. dr. Dan Cristea.

Referințe bibliografice

Armstrong, A. (1996). *Multext: Multilingual Text Tools and Corpora. Lexikon und Text*, pp. 107–119.

[4] Cuvinte care semnalează un anumit fenomen lingvistic; în acest caz, de exemplu: expresii temporale (*azi, septembrie*), semnale (*să, şi, că, când*).

Boguraev, B., Ando, R. (2006). Analysis of TimeBank as a Resource for TimeML Parsing. In *Proceedings of LREC 2006*, Genoa, Italy, pp. 71-76.

Brants, T. (2000). TnT – a statistical part-of-speech tagger. In *Proceedings of the 6th Applied NLP Conference*, ANLP-2000, Seattle, WA, pp. 224–231.

Cristea, D., Forăscu, C. (2006). Linguistic Resources and Technologies for Romanian Language. In *Journal of Computer Science of Moldova*, Academy of Science of Moldova, vol. 14, nr. 1(40), pp. 34-73, ISSN 1561-4042.

Ferro, L., Gerber, L., Mani, I., Sundheim, B., Wilson, G. (2005a). *TIDES 2005 Standard for the Annotation of Temporal Expressions,* April 2005.

Ferro, L., Gerber, L., Hitzeman, J., Lima, E., Sundheim, B. (2005b). *ACE Time Normalization (TERN) 2004 English Training Data v 1.0*, Linguistic Data Consortium, Philadelphia, ISBN 1-58563-331-3.

Forăscu, C., Solomon, D. (2004). Towards a Time Tagger for Romanian. In *Proceedings of the ESSLLI Student Session*, August 2004, Nancy, France.

Ion, R. (2006). *Metode de dezambiguizare semantică automată. Aplicaţii pentru limbile engleză şi română*. Teză de doctorat în curs de susținere la Academia Română.

Ide, N., Bonhomme, P., Romary, L. (2000). XCES: An XML-based Encoding Standard for Linguistic Corpora. In *Proceedings of the Second International Language Resources and Evaluation Conference*., pp. 825-830.

Mani, I., Pustejovsky, J., Gaizauskas, R. (eds.) (2005). *The Language of Time: A Reader*. Oxford University Press, ISBN-13: 978-0-19-926853-5, May 2005.

Pustejovsky, J., Verhagen, M., Sauri, R., Littman, J., Gaizauskas, R., Katz, G., Mani, I., Knippen, B., Setzer, A. (2006). *TimeBank 1.2*. Linguistic Data Consortium, Philadelphia, ISBN*:* 1-58563-386-0.

Sauri, R., Littman, J., Knippen, B., Gaizauskas, R., Setzer, A., Pustejovsky, J. (2006). *TimeML Annotation Guidelines, Version 1.2.1*, January 2006.

Tufiş, D., Ion, R., Ceauşu, A., Ştefănescu, D. (2006). Improved Lexical Alignment by Combining Multiple Reified Alignments. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics* (EACL2006), Trento, Italy, 3-7 April, 2006, pp. 153-160.

Tufiş, D., Barbu, A.M. (2002). Revealing translators knowledge: statistical methods in constructing practical translation lexicons for language and speech processing. In *International Journal of Speech Technology*. Kluwer Academic Publishers, no.5, pp.199-209, 2002, ISSN 1381-2416.

RESURSE LINGVISTICE REUTILIZABILE

CONSTANTIN CIUBOTARU, SVETLANA COJOCARU, ELENA BOIAN, ALEXANDRU COLESNICOV, LUDMILA MALAHOVA, VALENTINA DEMIDOV, OLEG BURLACA

Institutul de Matematică şi Informatică, Academia de Ştiinţe a Republicii Moldova

{chebotar, sveta, lena, kae, mal, demidova, burlaca}@math.md

Rezumat

Lucrarea este executată în cadrul proiectului RoLTech[1] şi are drept scop achiziţionarea resurselor electronice pentru limba română. Este prezentată structura bazei lexicale (componenta de bază a resurselor), sunt descrise particularităţile gestionării, verificării integrităţii, corectitudinii şi completitudinii ei. Se expun unele tehnici de verificare formală a bazei.

1. Introducere

Informatizarea continuă a societăţii se manifestă prin aplicarea activă a tehnologiilor informaţionale. În acest context procesarea limbajului natural (PLN) devine o problemă actuală. Putem menţiona trei direcţii importante în PLN:

- elaborarea instrumentarului pentru PLN;
- crearea resurselor lingvistice reutilizabile;
- elaborarea aplicaţiilor în baza acestor resurse.

Un efort important în implementarea produselor program pentru limba română îl prezintă pachetul de programe "Produse program pentru aplicaţii lingvistice" (Boian et al., 2000). Acest pachet a fost utilizat cu succes la implementarea Corectorului de texte pentru limba română RomSP (Boian et al., 2000; Cojocaru, 1997; Colesnicov, 1995). Dezvoltarea lui a condus la implementarea Resurselor Reutilizabile pentru Tehnologia Limbajului Natural (RRTLN), care conţin o bază de date cu informaţie lingvistică la nivel de cuvânt şi un set de programe de gestionare (Boian et al., 2005; Boian et al., 2005; Boian et al., 2003).

O trecere în revistă a produselor program create pentru procesarea limbajului natural este prezentată în (Cristea, Forăscu, 2006). RRTLN pot fi adăugate la clasificarea efectuată constituind o colecţie de intrări lexicale completată cu informaţie morfologică, forme flexionate, traduceri în limba engleză şi rusă, sinonime.

Corectorul RomSP poate fi considerat o aplicaţie lingvistică bazată pe RRTLN. Utilizând RRTLN, apare posibilitatea de elaborare a unui set de noi aplicaţii. De exemplu, dicţionare electronice; sisteme educaţionale pentru studierea morfologiei limbii române, scanere Web şi motoare de căutare capabile să utilizeze formele

[1] Romanian Language Technology, proiect INTAS Ref. Nr. 05-104-7633

flexionate ale cuvintelor limbii române. RRTLN pot fi utilizate şi la elaborarea aplicaţiilor "clasice" pentru PLN (de exemplu, parsere, tokenizatoare, lematizatoare, etc.).

Extinderea, modernizarea şi menţinerea resurselor lingvistice existente, precum şi a produselor program sunt efectuate în paralel cu elaborarea noilor aplicaţii. Începând cu anul 2006 se depun eforturi majore în direcţia dezvoltării platformei RoLTech pentru tehnologia limbii române. Platforma reprezintă o colaborare dintre cercetători din Republica Moldova, România şi Marea Britanie.

2. Proiectul RoLTech

Proiectul RoLTech propune următoarele obiective tehnice de bază:

1. Construirea portalului Web cu resurse lingvistice reutilizabile care vor fi folosite în tehnologia limbajului, produse program pentru tehnologia limbii române (atât surse deschise cât şi cu cod autorizat) şi referinţe la informaţii utile despre limba română.

2. Elaborarea aplicaţiilor bazate pe RRTLN:

   - un sistem de instruire adaptabil pentru morfologia limbii române cu elemente multimedia dedicat nevorbitorilor de limba română;
   - o aplicaţie dedicată vorbitorilor de limba română care are ca scop îmbogăţirea vocabularului în rezultatul căutării în colecţiile de documente în limba română cu ajutorul unui motor de căutare avansat;
   - o aplicaţie Web ce oferă un serviciu interactiv de corectare a textelor în limba română;
   - servicii pentru utilizatori experţi în limba română (de exemplu, un sistem suport pentru elaborarea dicţionarelor specializate).

Resursele şi aplicaţiile create în cadrul proiectului vor fi plasate pe Web-portal. Iniţial se vor crea versiuni prototipice, care mai apoi vor fi extinse, finalizate şi menţinute în continuare.

3. Flexionarea cuvintelor în limba română

Programele de flexionare (Cojocaru, 1997) au contribuit substanţial la acumularea resurselor lingvistice. Flexionarea cuvintelor este efectuată prin două metode: statică şi dinamică.

Metoda statică de flexionare se bazează pe clasificarea descrisă în lucrarea (Lombard, 1981). Algoritmul utilizează o gramatică de flexionare care formalizează procesul de realizare a alternanţelor şi concatenare a seturilor de terminaţii. Pentru limba română această gramatică include 866 reguli şi 320 seturi de terminaţii. Această metodă a fost aplicată pentru flexionarea a circa 30000 de cuvinte-leme.

Metoda dinamică nu utilizează liste de cuvinte, dar încearcă să calculeze paradigma de flexionare utilizând clasificările asemănătoare cu cele descrise în (Lombard 1981).

Algoritmul a fost verificat pe câteva mii de cuvinte în limba română, care nu au fost incluse în acele tabele. De asemenea au fost depistate unele iregularităţi (3% din mulţimea de cuvinte flexionate).

4. Structura bazei de date

Resursele Reutilizabile pentru Limba Română (RRLR) conţin o bază de date (BD) cu informaţie lingvistică pentru limba română la nivel de cuvânt şi un set de programe de gestionare a acestei baze de date. Ca volum RRLR conţin circa un milion de elemente.

În continuare vom descrie unele tehnici formale de verificare a integrităţii şi corectitudinii RRLR ce conţin cuvinte în limba română, derivate morfologice, sinonime, traduceri în limba engleză şi rusă. Descrierea mai detaliată a BD şi a algoritmilor este expusă în (Cojocaru, 2006).

BD a RRLR are şase tabele de bază şi 16 tabele auxiliare. Tabelele de bază sunt: *words, words_engl, words_rus, word flexies, word_synonyms, word_translations*. În primele trei tabele sunt cuvinte în limbile română, engleză şi rusă, cărora li se pun în corespondenţă nişte coduri numerice. Aceste coduri numerice sunt utilizate în celelalte trei tabele. De exemplu, tabelul *word_synonyms* conţine perechea de sinonime, care constă din două numere, ce corespund cuvintelor în limba română situate în tabelul *words_table*.

Tabelele auxiliare conţin diferite coduri utilizate în tabelele de bază: caracteristici morfologice, codurile limbilor, părţilor de vorbire, etc.

5. Popularea bazei de date

Pentru completarea BD cu informaţie morfologică s-a utilizat setul de fişiere produse în cadrul proiectelor precedente. Informaţia pentru traduceri şi sinonime a fost luată din diferite surse lexicografice (Boian et al., 2003).

Fişierele existente au fost transformate într-un format unic elaborat special pentru intrările BD.

Programul de populare a BD produce adiţional un fişier, care avertizează dacă cuvântul a fost deja inserat în BD, arată codul cuvântului şi rezultatul fiecărei operaţii. Erorile sunt marcate şi pot fi uşor depistate. Un alt mijloc de populare a BD cu informaţie morfologică este un program semi-automat care generează toate formele flexionate în baza cuvântului-lemă indicat.

Menţionăm trei surse de erori care apar la popularea BD: erori preluate din surse lexicografice, erori în programele utilizate la procesarea informaţiei şi erori produse de operator (factorul uman). O parte din aceste erori pot fi depistate doar cu implicarea experţilor filologi. O serie de alte erori pot fi depistate în mod automat cu ajutorul unor programe special elaborate.

6. Verificarea BD

Pentru început au fost utilizate metode formale de verificare a validității structurii BD. Aceste metode au fost formulate folosind semantica şi interdependențele câmpurilor BD şi a tabelelor. De exemplu, câmpul *part_code* din tabelul *words* conține numere – coduri ale părților de vorbire din tabelul *parts_of_speech* – şi de aceea ele pot avea numai valori întregi de la 1 (codul pentru verb) până la 10 (codul pentru conjuncție). Următoarea metodă formală a fost aplicată la verificarea cuvintelor. Pentru cuvintele în limba română a fost utilizat corectorul RomSP, care operează cu o listă de cuvinte deja testată de elaboratorii şi utilizatorii acestui produs. Cuvintele limbilor română, rusă şi engleză au fost testate utilizând corectoarele de texte MS Office pentru limbile corespunzătoare.

O metodă efectivă de verificare a fost utilizarea n-gramelor (părți de cuvinte ce conțin exact n litere). Cuvintele care conțin n-grame mai puțin frecvente se consideră a fi cele mai suspicioase.

Tabelele atributelor se pot verifica vizual deoarece ele sunt scurte. O altă metodă de verificare constă în căutarea codurilor atributelor care se folosesc rar sau nici nu se folosesc în tabelele de bază.

Tabelele de bază în BD conțin referințe mutuale. În caz ideal, oricărui cuvânt în limba română ar trebui să i se ataşeze forme flexionate, sinonime şi traduceri. Utilizând codificările din BD se pot căuta, de exemplu, cuvintele care nu au forme flexionate, traduceri, sinonime. Putem obține o listă de cuvinte pentru care lipseşte informația corespunzătoare, care ulterior ar putea fi adăugată în BD, sau o listă de cuvinte cu erori, care pot fi corectate.

Dublarea datelor s-a evitat la etapa de completare a BD. Apariția datelor dublate indica prezența unor erori în programele de completare a BD.

Verificarea statistică a procesului de flexionare ne-a permis să depistăm un şir de erori pentru cazurile când numărul formelor flexionate depăşea numărul admisibil pentru o anumită parte de vorbire, de exemplu, 35–40 pentru verb. Devierea acestor numere ne indică posibile erori.

A fost efectuată verificarea cuvintelor utilizând dicționare paralele. Vom menționa, că utilizarea resurselor paralele s-a dovedit a fi o metodă utilă în PLN (Tufiş, Barbu, 2002). În cazul nostru acestea au fost traducerile în limba rusă. Limba rusă, ca şi limba română, are un grad înalt de flexivitate. Au fost analizate verbe, adjective şi adverbe. Aceste părți de vorbire în limba rusă au terminații tipice. Cazurile, când partea de vorbire nu corespundea celei aşteptate, au fost clasificate drept suspecte şi examinate suplimentar.

7. Concluzii

Proiectul RoLTech, prin natura sa interdisciplinară (combinând informatica cu lingvistica) şi crearea portalului Web dedicat resurselor lingvistice, instrumentarului de procesare şi referințelor la cele mai importante evenimente şi descoperiri, relativ la limba vorbită în România şi Republica Moldova, va ajuta la atingerea mult râvnitei coordonări a activităților ambelor categorii de cercetători: a informaticienilor şi lingviştilor.

Referințe bibliografice

Boian, E., Cojocaru, S., Malahova, L. (2000). Instruments pour applications linguistiques. *La terminologie en Roumanie et en Republique de Moldova*, Hors serie, No. 4,.

Boian, E., Ciubotaru, C., Cojocaru, S., Colesnicov, A., Demidova, V., Malahova, L. (2005). Lexical resources for Romanian. *Scientific Memoirs of the Romanian Academy*, ser.IV, vol. XXVI, Bucureşti, România, pp. 267–278.

Boian, E., Cojocaru, S., Ciubotaru, C., Colesnicov, A., Demidova, V., Malahova, L. (2005). Technologization of Romanian: linguistic resources, applications, tools. *Proceedings of the 4rd International Conference on Microelectronics and Computer Science*. Vol.II, pp.519–522.

Boian, E., Ciubotaru, C., Cojocaru, S., Colesnicov, A., Demidova, V., Malahova, L. (2003). Lexical Resources for Romanian – a project overview. In: *Proceedings of Symposium on Intelligent Systems and Application*, September 19-20, Iasi, Romania, 12 pp. ISBN 973–97737–29.

Cojocaru, S. (1997). Romanian Lexicon: Tools, Implementation, Usage. In: Dan Tufiş, Poul Andersen (eds.). *Recent Advances in Romanian Language Technology*. ISBN 973–27–0626–0, Editura Academiei, I, pp. 107–114.

Cojocaru, S., Colesnicov, A., Malahova, L. (2006). Integrity and correctness checking of a lexical database. *Computer Science Journal of Moldova*, v. 14, Nr. 1(40), pp. 138–151.

Colesnicov, A. (1995) The Romanian spelling checker ROMSP: the project overview. Computer Science Journal of Moldova, v. 3, Nr. 1(7), pp. 40–54.

Cristea, D., Forăscu, C. (2006). Linguistic Resources and Technologies for Romanian Language. *Computer Science Journal of Moldova*, v. 14, Nr. 1(40), pp. 34–73.

Lombard, A., Gadei, C. (1981). *Dictionnaire morphologique de la langue roumaine,* Bucureşti (în franceză).

Tufiş, D., Barbu, A.M. (2002). Revealing Translator's Knowledge: Statistical Methods in Constructing Practical Translation Lexicons for Language and Speech Processing. *International Journal of Speech Technology* 5, pp. 199–209.

Capitolul 3

Aplicații ale tehnologiilor lingvistice

SISTEME DE ÎNTREBARE RĂSPUNS PENTRU LIMBA ROMÂNĂ

ADRIAN IFTENE[1], IONUȚ PISTOL[1], DIANA TRANDABĂȚ[1, 2], GEORGIANA PUŞCAŞU [3], CORINA FORĂSCU[1, 4], DAN CRISTEA[1, 2]

[1]*Facultatea de Informatică, Universitatea "Al.I.Cuza", Iaşi*
[2]*Institutul de Informatică Teoretică, Academia Română, Filiala Iaşi*
[3] *Universitatea Wolverhampton*
[4]*Institutul de Cercetări pentru Inteligență Artificială, Academia Română, Bucureşti*

{adiftene, ipistol, dtrandabat, corinfor, dcristea}@info.uaic.ro, georgie@wlv.ac.uk

Rezumat

În acest articol vom prezenta pe scurt modul în care au fost abordate problemele apărute în dezvoltarea sistemului de întrebare-răspuns pentru competiția CLEF 2006[1], secțiunea română-engleză. Pe lângă etapele clasice de adnotare sintactică şi semantică a corpusurilor, au fost probleme specifice datorate necesității unei traduceri sigure din română în engleză şi necesității evaluării rezultatelor folosind cât mai mult posibil metodologia CLEF. Acest articol descrie gradat paşii necesari implementării acestui sistem, pentru a putea evalua mai bine rezultatele rulărilor noastre în formatul CLEF.

1. Introducere

Primul sistem de Întrebare-Răspuns[2] românesc a fost dezvoltat în anii '80 (Tufiş şi Cristea, 1985) şi era reprezentat de o interfață ce facilita comunicarea cu o rețea semantică (care codifica cunoaşterea). Astăzi sistemele de ÎR folosesc documente text ca bază de cunoaştere şi integrează tehnici de prelucrare a limbajului natural (PLN) pentru a găsi (într-o colecție dată de documente sau prin căutare pe web) răspunsul la o întrebare pusă în limbaj natural.

România a participat pentru prima dată la o competiție CLEF în 2006, în cadrul secțiunii QA@CLEF[3]. Organizatorii au decis că limba sursă (a întrebărilor) să fie româna în timp ce limba țintă (a colecției de documente în care este căutat răspunsul) să fie engleza, datorită inexistenței unui corpus ziaristic din perioada anilor 1994-1995, care ar fi permis şi folosirea românei ca limbă țintă. Astfel, la întrebările puse în română s-au căutat răspunsuri sub forma unor fragmente de text în colecția de documente în engleză.

Ca şi în celelalte interacțiuni multilingve din competiție, sistemul nostru a fost evaluat pe un set de 200 de întrebări în limba română. Aflați pentru prima dată într-o astfel de competiție, intenția a fost obținerea în primul rând a unui sistem funcțional, calitatea rezultatelor fiind lăsată, în acest an, pe locul al doilea.

---

[1] Cross-Language Evaluation Forum: http://www.clef-campaign.org/2006.html
[2] Question Answering (QA) – rom.: Întrebare-Răspuns (ÎR)
[3] Multilingual Question Answering at CLEF: http://clef-qa.itc.it/

2. Descrierea Sistemului

2.1. Prezentare Generală

De regulă, sistemele de ÎR folosesc o arhitectură generală de tip pipe-line, în care prelucrarea parcurge trei etape principale: analiza întrebării, căutarea documentară şi extragerea răspunsului (Harabagiu, Moldovan, 2003). Sistemul creat este o variantă a arhitecturii generale, cu particularizări specifice legate de reprezentare şi procesare pentru fiecare din componentele amintite mai sus. Un modul aparte inclus în sistemul de ÎR este modulul care traduce cuvintele din română în engleză, pentru a face transferul interlingv. Arhitectura şi funcţionalitatea sistemului sunt ilustrate în Figura 1.

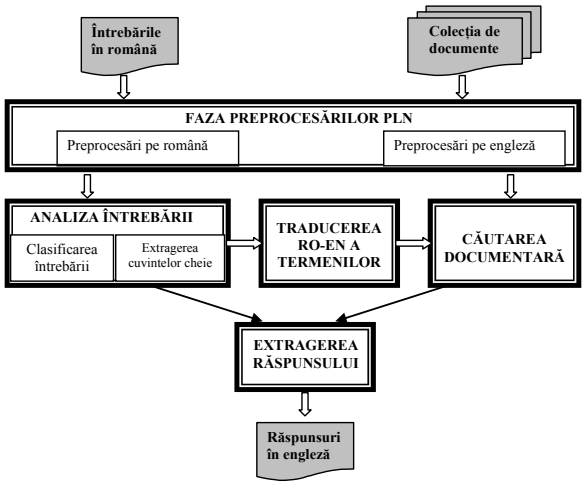


Figura 1: Arhitectura şi funcţionalitatea sistemului

2.2. Preprocesări asupra corpusului

Iniţial, întrebările sunt procesate morfo-sintactic folosind POS tagger-ul românesc dezvoltat de ICIA[4] (Tufiş, 1999; Ion, 2006). Ulterior, se realizează clasificarea numelor proprii, printr-o procedură de recunoaştere bazată pe şabloane, în următoarele clase: PERSOANĂ, LOCAŢIE, MĂSURĂ şi altele. Aceleaşi operaţii de preprocesare sunt efectuate şi pe colecţia de documente englezeşti, folosind o segmentare la nivel de propoziţie. Pentru aceasta s-a folosit acelaşi parser de la ICIA, dar cu un alt model de limbă. Corpusul CLEF englezesc constând din articole din ziarele Los Angeles Times anul 1994 şi Glasgow Herald anul 1995. Acest corpus a fost segmentat la nivel de propoziţie şi cuvânt, iar apoi a fost etichetat la partea de vorbire.

2.3. Analiza întrebării

Această etapă are în vedere, în primul rând, identificarea tipului semantic al entităţii ce ar trebui să sugereze tipul răspunsului aşteptat. În plus se identifică focusul întrebării, tipul întrebării şi mulţimea cuvintelor cheie relevante pentru întrebare. Pentru a putea atinge aceste scopuri, analiza efectuează următorii paşi:

[4] Institutul de Cercetări în Inteligenţă Artificială al Academiei Române: http://www.racai.ro/

a) ***Depistarea grupurilor nominale (GN), extragerea numelor de entități (NE), identificarea expresiilor temporale (ET)***

Identificatorul de nume de entități construit de ICIA determină numele de entități din întrebarea în româneşte. Expresiile temporale sunt de asemenea identificate folosind un identificator şi un normalizator de ET pentru limba română adaptat după varianta pe limba engleză descrisă în (Puşcaşu, 2004).

b) ***Identificarea focusului întrebării***

Focusul întrebării este cuvântul sau secvența de cuvinte care arată ce anume se caută. Se consideră ca focus al întrebării substantivul ulterior pronumelui interogativ din întrebare (ca în *Ce țară*) sau primul grup nominal (GN) al întrebării dacă el apare înainte de verbul principal al întrebării sau de cel ce urmează verbului *a fi*.

c) ***Găsirea tipului răspunsului aşteptat***

Sistemul de analiză a întrebării poate face distincție între următoarele clase de răspuns: PERSOANĂ, LOCAȚIE, ORGANIZAȚIE, TEMPORAL, NUMERIC, DEFINIȚIE şi GENERIC. Atribuirea unei clase unei întrebări analizate este realizată folosind focusul întrebării şi tipul acestuia. De exemplu, în cazul întrebării *În ce oraş a fost omorât Vladislav Listyev?*, focusul întrebării este *oraş*, substantiv ce apare în lista LOCAȚIILOR, şi astfel se determină tipul răspunsului ca fiind LOCAȚIE. Tipul focusului întrebării este determinat folosind WordNet (Fellbaum, 1998).

d) ***Deducerea tipului întrebării***

În acest an, competiția QA@CLEF a făcut distincție între patru tipuri de întrebare: *factoid*, *definiție*, *listă* şi întrebări cu *restricții temporale.* Deoarece restricțiile temporale se pot ataşa oricărui tip de întrebare, alegerea unuia dintre tipurile *factoid*, *definiție* sau *listă* se face înainte de testarea existenței restricțiilor temporale.

e) ***Generarea mulțimii cuvintelor cheie***

Mulțimea cuvintelor cheie este generată automat din lista termenilor importanți ai întrebării în ordinea inversă a relevanței acestora. Prin urmare, mulțimea cuvintelor cheie cuprinde: focusul întrebării, NE şi ET identificate, substantivele rămase, şi toate verbele diferite de cele auxiliare prezente în întrebare. Această mulțime este apoi trimisă modulului de traducere a termenilor, cu scopul de a obține cuvintele cheie englezeşti necesare căutării documentare.

2.4. Traducerea termenilor

Pentru a realiza traducerea termenilor s-a folosit ca resursă WordNet-ul, disponibil atât în limba română (Tufiş et al., 2006) cât şi în limba engleză. Mulțimea cuvintelor cheie extrase în faza de analiză a întrebării este folosită ca intrare pentru faza de traducere a termenilor, şi prin urmare s-au tradus atât substantivele cât şi verbele. Cuvintele componente ale GN sunt traduse unul câte unul: synseturile româneşti care conțin cuvântul în cauză sunt puse în corespondență, prin indexul inter-lingual ILI, cu cele echivalente din engleză, obținându-se astfel mulțimea tuturor traducerilor posibile. Dacă cuvântul ce trebuie tradus nu apare în WordNet, caz destul de frecvent, acesta se caută în alte dicționare disponibile şi, dacă este găsit, se păstrează primele trei traduceri. În cazul

verbelor, se extrage pentru fiecare verb traducerea echivalentă din WordNet la fel ca la substantive. Pentru cazurile de verb-substantival am folosit situaţiile din (Pekar et al, 2004) în selectarea traducerii.

2.5. Crearea indexului şi căutarea documentară

Corpusul englezesc a fost preprocesat iniţial folosind instrumente pentru împărţirea în cuvinte, instrumente pentru găsirea lemei şi a părţii de vorbire, şi unelte de recunoaştere a NE. În rulările noastre indexarea şi căutarea s-a realizat cu motorul Lucene[5].

Colecţia de documente a fost indexată atât la nivel de document cât şi la nivel de paragraf folosind lema cuvintelor conţinute şi a clasele NE (MĂSURĂ, PERSOANĂ, LOCAŢIE, etc). Când nu este găsit nici un paragraf pentru o anumită interogare, se folosesc două strategii: fie se măresc segmentele de la paragrafe la documente, fie se reformulează interogarea folosind pentru anumite cuvinte alte variante de traducere.

2.6. Extragerea răspunsului

Două module de extragere a răspunsului au fost dezvoltate, unul de către UAIC[6] şi altul de către ICIA. Ambele module au ca intrare tipul răspunsului aşteptat, focusul întrebării, mulţimea de cuvinte cheie, părţile de text obţinute în urma căutării pe partea de vorbire, lemă şi informaţii de tip NE şi indicatorul de relevanţă al paragrafelor determinat de Lucene. Procesul de extragere depinde de tipul aşteptat al răspunsului: când răspunsul are ca tip un NE, modulul de extragere a răspunsului identifică în fiecare propoziţie întoarsă de Lucene entităţile de tip NE care au tipul dorit de răspuns. Când tipul răspunsului nu este un NE, procesul de extragere se bazează în principal pe recunoaşterea focusului, în acest caz şabloanele sintactice de găsire a răspunsului bazate pe focus fiind cruciale.

3. Descrierea rulărilor înscrise în competiţie

Au fost înscrise în competiţie trei rulări diferite, cu următoarele detalii:

- **UAIC -** Această rulare a fost obţinută prin parsarea şi analizarea întrebărilor, traducerea cuvintelor cheie, căutarea pasajelor relevante şi căutarea răspunsurilor finale folosind extractorul de răspunsuri realizat de UAIC.
- **RACAI**[7] **-** Această rulare a fost obţinută de asemenea prin parsarea şi analizarea întrebărilor, traducerea cuvintelor cheie, căutarea pasajelor relevante, dar pentru căutarea răspunsurilor finale s-a folosit extractorul de răspunsuri ICIA.
- **DIOGENE -** Cea de a treia rulare a fost obţinută prin conversia rezultatelor modulelor de analiză a întrebării şi traducere a termenilor în formatul cerut de sistemul de ÎR DIOGENE (Kouylekov et al, 2003), şi apoi trimiterea lor ca intrare la modulele DIOGENE de căutare documentară şi de extragere de răspunsuri.

[5] http://lucene.apache.org/
[6] Universitatea "Al.I.Cuza": http://www.uaic.ro
[7] Sigla englezească pentru ICIA.

Datorită numărului mare de rulări înscrise având ca ţintă limba engleză doar rulările UAIC şi RACAI au fost evaluate. În continuare, sistemul RACAI va fi referit ca Sistemul 1, şi sistemul UAIC va fi referit ca Sistemul 2.

4. Analiza Rezultatelor

Rezultatele evaluării oficiale pentru Sistemele 1 şi 2 sunt prezentate în Tabelul 1. Fiecare răspuns a fost evaluat ca fiind NECUNOSCUT (răspunsurile neevaluate), CORECT (răspunsurile corecte), NEJUSTIFICAT (răspunsuri care nu puteau fi găsite în bucăţile de text justificatoare), INCORECT (răspunsurile greşite) sau INEXACT (răspunsuri incomplete).

Tabel 1: Evaluarea rezultatelor pentru cele două sisteme

Evaluarea rezultatelor pentru Sistemul 1		
Z	NECUNOSCUT	400
R	CORECT	35
U	NEJUSTIFICAT	13
W	INCORECT	184
X	INEXACT	7
	TOTAL	**639**

Evaluarea rezultatelor pentru Sistemul 2		
Z	NECUNOSCUT	543
R	CORECT	22
U	NEJUSTIFICAT	4
W	INCORECT	191
X	INEXACT	1
	TOTAL	**761**

Numărul mare de răspunsuri evaluate ca fiind NECUNOSCUT se datorează faptului că s-au determinat 10 răspunsuri pentru aproape toate cele 200 de întrebări, unde 10 a fost numărul maxim de răspunsuri posibile. Cum evaluarea finală a ţinut cont doar de primul răspuns pentru majoritatea întrebărilor (doar în cazul întrebărilor de tip *listă* au fost evaluate primele trei răspunsuri), răspunsurile de pe poziţiile de la 2 la 10 au fost etichetate ca NECUNOSCUT, indicând faptul că nu s-a făcut nici o încercare pentru a le verifica corectitudinea. Folosind un evaluator dezvoltat special pentru aceasta răspunsul corect a fost găsit în primele zece răspunsuri generate de sistemele noastre pentru 35-40% din întrebări. Acesta este un rezultat promiţător, dovedind că extractorul de răspunsuri funcţionează, dar enecesară îmbunătăţirea ordonării răspunsurilor obţinute.

5. Concluzii

Respectând arhitectura clasică a sistemelor de tip ÎR, sistemul dezvoltat implementează cele trei niveluri esenţiale ale unui astfel de sistem, ca şi modulul specific sistemelor interîlinguale care traduce termenii relevanţi ai întrebării din română în engleză.

Rezultatele şi evaluările, deşi nu tocmai satisfăcătoare, vor fi folosite pentru îmbunătăţirea sistemului la ediţia viitoare de QA@CLEF sau la alte competiţii similare precum TREC[8].

O analiză detaliată a relevat un număr important de direcţii pentru îmbunătăţirea substanţială a sistemului. Modulul de traducere a termenilor, cheia performanţelor pentru orice sistem interlingv, este principalul obiectiv. Modulul de extragere a

[8] Text REtrieval Conference: http://trec.nist.gov/

răspunsului va fi modificat astfel încât precizia acestuia să crească. O metodă mai bună de ordonare a răspunsurilor candidate este a treia direcţie prioritară pentru viitor.

Autorii sistemului

- ◊ **Membrii echipei UAIC**: Dan Cristea, Iustin Dornescu, Corina Forăscu, Maria Husarciuc, Adrian Iftene, Ana Masalagiu, Alex Moruz, Ionuţ Pistol, Diana Trandabăţ;
- ◊ **Membrii echipei ICIA**: Alin Ceauşu, Radu Ion, Dan Ştefănescu, Dan Tufiş;
- ◊ **Universitatea Wolverhampton**: Georgiana Puşcaşu, Constantin Orăsan.

Mulţumim lui Milen Kouylekov şi Bernardo Magnini pentru disponibilitatea lor de a procesa ieşirea procesorului nostru de întrebări şi a modulului de traducere cu sistemul DIOGENE, dezvoltat de IRST Trento.

Parţial, acest proiect a fost finanţat de Ministerul Educaţiei şi Cercetării în cadrul proiectului CEEX 29 ROTEL şi de INTAS în cadrul proiectului RolTech (INTAS ref. 05-104-7633).

Referinţe bibliografice

Fellbaum, C. (1998) (ed.) WordNet: An Eletronic Lexical Database. The MIT Press.

Harabagiu, S., Moldovan, D. (2003). Question answering. In R. Mitkov (ed.), *Oxford Handbook of Computational Linguistics*, pp. 560 - 582. Oxford University Press.

Kouylekov, M., Magnini, B., Negri, M., Tanev, H. (2003). ITC-irst at TREC-2003: the DIOGENE QA system. *Proceedings of the TREC-12 Conference*.

Ion, R. (2006). Metode de dezambiguizare semantică automată. Aplicaţii pentru limbile engleză şi română. Teză de doctorat în curs de susţinere la Academia Română.

Pekar, V., Krkoska, M., Staab, S. (2004). Feature weighting for cooccurrence-based classification of words. *Proceedings of the 20th International Conference on Computational Linguistics (COLING-04)*.

Puşcaşu, G. (2004). A Framework for Temporal Resolution. *Proceedings of the 4th Conference on Language Resources and Evaluation (LREC2004)*.

Tufiş, D. (1999). Tagging with Combined Language Models and Large Tagsets. *Proceedings of the TELRI International Seminar on "Text Corpora and Multilingual Lexicography"*.

Tufiş, D., Barbu Mititelu, V., Ceauşu, A., Bozianu, L., Mihăilă, C., Manu Magda, M.,. (2006). Noi dezvoltări ale wordnet-ului românesc. În acest volum.

Tufiş, D., Cristea, D., Stamou, S. (2004). BalkaNet: Aims, Methods, Results and Perspectives. A General Overview. In D. Tufiş (ed.), *Romanian Journal on Information Science and Technology*. Special Issue on BalkaNet. Romanian Academy.

IDENTIFICAREA ŞI EXTRAGEREA AUTOMATĂ A COLOCAȚIILOR DIN TEXTE

DAN ŞTEFĂNESCU, DAN TUFIŞ, ELENA IRIMIA

Institutul de Cercetări pentru Inteligență Artificială, Academia Română, Bucureşti

{danstef, tufis, elena}@racai.ro

Rezumat

Identificarea şi extragerea automată a colocațiilor din texte este necesară în rezolvarea multor probleme dificile de prelucrare a limbajului natural, cum ar fi generarea limbajului natural, rezumarea sau traducerea automată. Colocațiile sunt expresii care de obicei nu pot fi traduse cuvânt cu cuvânt (folosind doar un simplu dicționar şi un model de limbă). Acest lucru se întâmplă deoarece sunt caracterizate de compoziționalitate limitată – înțelesul expresiei nu se obține întotdeauna însumând înțelesurile cuvintelor ce o compun.

1. Introducere

Diverse definiții au fost propuse pentru noțiunea de *colocație*, mai mult sau mai puțin stricte. Iată câteva dintre cele care sunt folosite în lingvistica computațională:

- expresie formată din două sau mai multe cuvinte ce corespunde unui mod convențional de a afirma, de a exprima, anumite lucruri;
- „Două sau mai multe cuvinte ce apar împreună semnificativ de des în interiorul unei ferestre pre-definite într-un corpus dat" (Quasthoff & Wolff, 2002);
- "O secvență de două sau mai multe cuvinte consecutive, ce are caracteristicile unei unități sintactice şi semantice, şi a cărei înțeles exact şi neambiguu nu poate fi obținut direct din înțelesurile sau conotațiile cuvintelor ce o compun" (Choueka, 1988).

Colocațiile pot fi grupuri nominale (*televizor alb-negru*, *arme de distrugere în masă, vin roşu, drept de suită*), locuțiuni verbale (*a aduce atingere, a intra în vigoare, a face obiectul, a lua în considerare*) şi nu numai (*sărac dar cinstit*, *tânăr şi neliniştit, de jur împrejur*).

Colocațiile se pot caracteriza prin (Manning & Schütze, 1999):

- **Non-compoziționalitate** – atunci când înțelesul întregului este diferit de suma înțelesurilor părților;
- **Non-substituționalitate** – atunci când componentele colocației nu pot fi substituite cu sinonime;
- **Non-modifiabilitate** – atunci când colocațiile nu pot fi modificate prin adăugarea de material lexical adițional sau prin transformări gramaticale.

O colocaţie pentru care avem îndeplinite toate cele trei condiţii de mai sus se apropie foarte mult de noţiunea de idiom. Colocaţiile pot fi clasificate după mai multe criterii, ele pot fi de natură lexicală, sintactică, sau de natura semantică, pot fi generale sau specifice unui anumit domeniu, pot avea structură fixă sau structură variabilă.

Literatura de specialitate propune diferite metode pentru găsirea colocaţiilor. Justeson şi Katz (1995) au folosit doar frecvenţa de ocurenţă a cuvintelor în perechi şi un filtru pe părţile de vorbire; Smadja (1990) a folosit o metodă bazată pe media şi dispersia distanţelor dintre (două) cuvinte în corpus, în timp ce alţii (Church et al., 1991) au utilizat Testul *t*, *chi* pătrat, *log-likelihood* sau informaţia mutuală pentru a găsi cuvinte ce apar împreună, în text, mai des decât ne-am putea aştepta să apară întâmplător.

2. Modelarea colocaţiilor

In modelarea noastră, colocaţiile sunt succesiuni de cuvinte (nu neapărat adiacente) care respectă două criterii statistice:

- distanţa dintre cuvinte este relativ constantă;
- apar în aceleaşi contexte de un număr de ori semnificativ din punct de vedere statistic.

Primul criteriu este evaluat folosind abordarea lui Smadja (1990) iar cel de-al doilea se bazează pe calculul raportului Log-Likelihood (LL).

Pentru identificarea colocaţiilor Smadja propune utilizarea mediei şi dispersiei distanţelor dintre cuvintele din corpus (2 câte 2). Dispersia măsoară deviaţia distanţelor de la medie:

$$\sigma^2 = \frac{\sum_{i=1}^{n} (d_i - \mu)^2}{n-1}$$

Dacă două cuvinte se găsesc în corpus mereu la aceeaşi distanţă, dispersia este 0. Dacă distanţele au o distribuţie aleatoare (cum este cazul atunci când cuvintele apar împreună întâmplător), dispersia are valori ridicate. Putem afirma astfel că media şi deviaţia standard (rădăcina pătrată a dispersiei) sunt mărimi care caracterizează distribuţia distanţelor dintre două cuvinte într-un corpus. Smadja demonstrează că se pot descoperi colocaţii căutând perechi de cuvinte pentru care avem deviaţii standard mici.

Scorul Log-Likelihood calculează raportul probabilităţilor a două ipoteze statistice care pot fi emise în descrierea datelor observate într-un text. Ipotezele pe care le luăm în considerare sunt (i)H_0: cele două cuvinte nu au nici o legătură între ele şi apar întâmplător împreună şi (ii)H_1: cele două cuvinte sunt cumva corelate şi apariţia lor împreună nu este întâmplătoare.

- $H_0 : P(w_2|w_1) = p = P(w_2|\neg w_1)$

  (presupunere de independenţă)

- $H_1 : P(w_2|w_1) = p_1 \neq p_2 = P(w_2|\neg w_1)$

  (w_1 şi w_2 nu sunt independente)

Pentru calculele efective se foloseşte un tabel de contingenţă ca mai jos, în care fiecare celulă conține numărul de apariții ale diferitelor combinații de cuvinte pentru care se evaluează scorul LL. Astfel, n_{11} reprezintă numărul de apariții împreună ale cuvintelor w_1 şi w_2, n_{12} reprezintă numărul de apariții ale cuvântului w_1 în contextele în care cuvântul w_2 lipseşte, etc.

	w_2	$\neg w_2$
w_1	N_{11}	n_{12}
$\neg w_1$	N_{21}	n_{22}

Notând cu: $n_{1*} = n_{12} + n_{11}$, $n_{*1} = n_{21} + n_{11}$ şi cu $n_{**} = n_{22} + n_{11}$, formula de calcul este:

$$LL = 2 * \sum_{j=1}^{2} \sum_{i=1}^{2} n_{ij} * \log \frac{n_{ij} * n_{**}}{n_{i*} * n_{*j}}$$

Dacă scorul obținut este mai mare decât un anumit prag ipoteza de nul (H_0) este respinsă cu un anumit grad de certitudine.

Rezultatele obținute folosind o combinație a celor două metode descrise mai sus indică un lucru interesant: utilizarea scorului LL calculat pentru perechi de cuvinte care îndeplinesc anumite criterii ce țin de partea de vorbire, cât şi de media distanței dintre cuvinte, constituie o abordare eficientă. În ceea ce ne priveşte, suntem interesați de extragerea de colocații ce sunt de tip substantiv-substantiv (S-S), substantiv-adjectiv (S-A) / adjectiv-substantiv (A-S) şi substantiv-verb (S-V)/ verb-substantiv (V-S). Trebuie să remarcăm că, în timp ce colocațiile din prima şi a doua categorie se încadrează în general în categoria compuşilor terminologici (termenii cheie se găsesc aici în marea lor majoritate), cele din a treia categorie caracterizează structurile de sub-categorizare verbale.

Următoarele rânduri descriu metoda folosită de noi pentru extragerea colocațiilor de tip substantiv-verb (verb-substantiv). Inițial, textul este lematizat şi adnotat la părți de vorbire. Apoi, o fereastră de 11 cuvinte (acesta este contextul în care se consideră co-ocurențele) parcurge fiecare propoziție din text în aşa fel încât fiecare cuvânt devine la un moment dat centrul ferestrei[1]. Cuvintele ce se introduc în fereastră sunt substantive sau verbe; celelalte părți de vorbire sunt ignorate. Lungimea a fost aleasă astfel încât fereastra să poată cuprinde orice pereche de cuvinte interesantă care ar exista. Am considerat că o distanță de 5 (stânga/dreapta) pentru o astfel de fereastră, în care se găsesc doar cuvinte ce sunt verbe sau substantive, (pentru alte tipuri de colocații considerăm doar substantivele şi adjectivele, sau doar substantivele) este suficientă pentru a găsi perechile interesante. Deşi ar putea exista exemple în care distanța dintre cuvinte este mai mare de 5 (numărând doar cuvintele din categoriile gramaticale care ne interesează pe noi), aceste cazuri sunt rare şi se datorează probabil intercalării unor expresii între cuvintele ce formează perechea interesantă[2]. Toate perechile de cuvinte (sub formă de lemă) ce se formează între centrul ferestrei şi celelalte cuvinte din fereastră, împreună cu distanța[3] dintre cuvintele ce formează aceste perechi, sunt

[1] În acest pas aplicăm metoda lui Smadja. Aceasta ne permite să identificăm perechi interesante de cuvinte ce nu sunt neapărat adiacente.

[2] Este evident că această tehnică funcționează pentru limbi ca româna, engleza, franceza şi multe altele; pentru limba germană, unde verbul stă uneori la sfârşitul propoziției, căutarea perechilor interesante verb-substantiv în funcție de distanța dintre cuvinte trebuie regândită, iar contextul de căutare trebuie extins la nivelul întregii propoziții.

[3] Distanța este negativă dacă perechea e formată de cuvântul-centru împreună cu un cuvânt ce stă în fața sa.

introduse într-o bază de date. După ce a fost parcurs întreg textul, pentru fiecare pereche din baza de date, calculăm media şi dispersia luând în calcul ocurenţele la diferite distanţe. Dispersia reprezintă variaţia distanţei dintre cele două cuvinte în jurul mediei. În cazul în care avem o dispersie mică, media ne indică distanţa uzuală la care se află cele două cuvinte în text. Dispersia este pătratul deviaţiei standard. În cazul nostru, am păstrat ca perechi interesante pe cele pentru care deviaţia standard este sub un prag de 1,5. Valoarea de 1,5 este îndeajuns de mare încât să prindem toate perechile interesante[4]. Pentru perechile interesante extrase, calculăm scorul LL. În acest calcul intră doar perechile de cuvinte de tip S-V / V-S care se află la o distanţă egală cu media perechii pentru care se calculează acest scor. Dacă scorul LL depăşeşte pragul de 9, spunem că perechea este o colocaţie. Pentru un scor de 9, probabilitatea de eroare este mai mică de 0,004[5]. Dacă dorim înăsprirea condiţiilor, putem micşora pragul pentru dispersie şi / sau ridica pragul pentru scorul LL.

Trebuie să remarcăm că, dacă din lista colocaţiilor obţinute, luăm primele *x* în ordinea tăriei cu care sunt legaţi constituenţii din colocaţii[6], tărie care este dată de scorul LL, obţinem o listă de termeni cheie relevanţi pentru documentul din care i-am extras.

3. Evaluarea metodei

Am testat această metodă în contextul ambelor proiecte de care am amintit chiar la începutul acestui articol. Am folosit un corpus românesc, lematizat şi adnotat la părţi de vorbire, cu o mărime de aproximativ 350Mb ce conţine articole ce fac parte din Acquis-ul Comunitar.

În cazul extragerii de termeni cheie, trebuie să facem precizarea că ne aşteptăm ca ei să fie o submulţime a colocaţiilor de tip S-S şi S-A. Pentru o evaluare adecvată am fi avut nevoie de documente în care termenii cheie să fi fost etichetaţi. Cum nu am avut posibilitatea să folosim astfel de documente, ne-am orientat spre utilizarea tezaurului Eurovoc, un tezaur poli-tematic, multilingv, folosit pentru indexarea Acquis-ului Comunitar (AC) (legislaţia şi tratatele interne ale comunităţii europene). Tezaurul conţine 6645 de termeni organizaţi în structuri arborescente. Dintre aceştia, 519 termeni sunt foarte generali, constituind rădăcinile arborilor. Din punctul nostru de vedere, termenii Eurovoc-ului sunt doar o submulţime a întregii mulţimi a termenilor cheie ce pot caracteriza documentele AC. Datorită faptului că ierarhia tezaurului nu este una adâncă, există un anumit grad sau nivel de generalizare la care se opresc termenii din Eurovoc. Cu alte cuvinte, se pot găsi termeni cheie pentru anumite documente din AC, foarte specifici, care însă nu se regăsesc în Eurovoc. Un exemplu în acest sens este *tratat de instituire* care este un anumit tip de tratat, dar care nu face parte din tezaur.

tratat instituire 2(distanţa) 71286.61852(scorul LL) 6175(ocurenţe în text)

[4] A se vedea exemplele din Manning & Schütze – într-o pereche cu o deviaţie standard de peste 2, cuvintele nu au nici o legătură; ajung împreună din întâmplare.

[5] Pragul de 0,004 înseamnă că aproximativ odată la 250 de cazuri avem o situaţie în care deşi două cuvinte apar împreună întâmplător, scorul LL este 9. Pentru un prag de 0,001scorul LL trebuie ales 10,83 (a se consulta tabelele cu pragurile date pentru distribuţiile *chi* pătrat cu un grad de libertate).

[6] În funcţie de cât de caracteristici vrem să fie termenii cheie, *x* poate fi mai mare (chiar câteva sute) sau mai mic (până la câteva zeci).

Perechea **tratat – instituire** aşa cum apare în text:

tratatul/tratat/nsry de/de/s instituire/instituire/nsrn 6052
tratatului/tratat/nsoy de/de/s instituire/instituire/nsrn 70
tratatele/tratat/npry de/de/s instituire/instituire/nsrn 29
tratatelor/tratat/npoy de/de/s instituire/instituire/nsrn 13
tratatul/tratat/nsry prevede/prevedea/v3 instituirea/instituire/nsry 3
tratat/tratat/nsn prevede/prevedea/v3 instituirea/instituire/nsry 2
tratate/tratat/npn de/de/s instituire/instituire/nsrn 2
tratatului/tratat/nsoy este/fi/v3 instituirea/instituire/nsry 1
tratat/tratat/nsn figurează/figura/v3 instituirea/instituire/nsry 1
tratat/tratat/nsn menționează/menționa/v3 instituirea/instituire/nsry 1
tratatul/tratat/nsry privind/privi/vg instituire/instituire/nsrn 1

Se observă că *tratat de instituire* apare foarte des în documentele AC şi, în plus, scorul LL este foarte ridicat. Aceste lucruri susțin ideea că *tratat de instituire* poate fi considerat cuvânt cheie chiar dacă el nu apare în tezaurul oficial.

Iată alte colocații, de tip S-S, considerate de noi termeni cheie pentru documentele AC. Aceşti termeni se regăsesc şi în Eurovoc:

stat membru	1	100653.529	23854	program lucru	2	2893.112575	439
directivă consiliu	1	42393.08842	6556	tratat aderare	2	690.3592391	153

Iată exemple de colocații de tip S-V extrase folosind aceeaşi metodă:

aduce atingere	1	51567.34864	4959	ține seamă	1	22825.70709	2357
înlocui text	3	43992.3067	4114	adopta bruxelles	2	21792.22915	2610
intra vigoare	2	42527.03736	4473	adopta regulament	2	20847.73793	2951
avea tratat	3	32050.11219	5816	lua măsuri	1	19207.12849	2491
face obiect	1	30729.47663	3898	pune aplicare	2	13186.20796	1564
modifica regulament	4	29141.39454	3098	face referire	1	11854.14299	1486
modifica dată	2	27658.4116	3213	informa privire	4	10586.88849	1175
lua considerare	2	27062.0349	2621	îndeplini condiție	1	8382.436218	1334
ține cont	1	26635.12649	2868	intra incidență	2	8119.841768	985
adresa membru	2	25844.0428	2362	îndeplini cerință	1	7473.851703	1223

O parte din acestea, cum ar fi *a intra în vigoare*, *a lua în considerare, a fi adoptat la bruxelles* pot fi considerați termeni specifici limbajului juridic reflectat în corpusul AC.

4. Concluzii şi continuarea cercetărilor

Au fost prezentate o parte din rezultatele obținute, până în prezent, în cadrul unui proiect internațional la care participă: ICIA (RO), Universitatea March Bloch (FR) şi IMS Stuttgart (GE), care are ca obiectiv construirea unui dicționar de colocații pentru cele 3 limbi. Totodată, această tehnică de extragere a colocațiilor a fost folosită cu succes şi în cadrul proiectului CEEX-ROTEL, pentru extragerea automată a termenilor cheie multi-cuvânt din documente arbitrare.

Testele efectuate până acum arată că metoda noastră este una eficientă. In plus, ea nu depinde de limba naturală în care sunt redactate documentele prelucrate. În continuare, sunt prevăzute o serie de experimente şi evaluari noi:

- vor fi extrase şi colocaţii de tip S-A evaluarea extragerii termenilor cheie se va face în raport cu tezaurul Eurovoc;
- se va testa o metodă de extragere a colocaţiilor similară cu cea prezentată, dar în care cele două faze principale sunt interschimbate;
- se vor extrage colocatiile din textele de limbă franceză şi germană;
- se vor extrage diverse statistici din colocaţiile deja extrase;
- se vor compara cross-lingual colocaţiile extrase din documentele AC, folosind tehnologia alinierii la nivel de cuvânt (Tufiş et al., 2006) a corpusului paralel.

Referinţe bibliografice

Choueka, Y. (1988), Looking for needles in a haystack, *Proceedings of RIAO '88*, 609 – 623.

Church, K., Gale, W., Hanks, P., Hindle, D. (1991). Parsing, word associations and typical predicate-argument relations, *Current Issues in Parsing Technology*. Kluwer Academic, Dordrecht, Olanda.

Justeson, J. S., Katz, S. M. (1995). Technical terminology: Some linguistic properties and an algorithm for identification in text, *Natural Language Engineering*, 1:9-27.

Manning, C., Schütze, H. (1999), *Foundations of Statistical Natural Language Processing,* MIT Press, Cambridge.

Quasthoff, U., Wolff, C, (2002). The Poisson collocation measure and its application, *Workshop on Computational Approaches to Collocations*, Viena, Austria

Salton, G., McGill, M. J. (1983). *Introduction to modern information retrieval,* McGraw-Hill.

Salton, G., Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management 24(5)*, 513–523.

Smadja, F. A., McKeown, K. R. (1990). Automatically extracting and representing collocations for language generation. *Proceedings of the 28th annual meeting on Association for Computational Linguistics*, 252-259, Pittsburgh, Pennsylvania.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C. Erjavec, T., Tufiş, D., Varga, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. *Proceedings of the 5th LREC Conference*, pp.2142-2147.

Tufiş, D., Ion, R., Ceauşu, A., Ştefănescu, D. (2006). Improved Lexical Alignment by Combining Multiple Reified Alignments. *Proceedings of the 11th EACL Conference*, pp. 153-160.

SPRE O EXTRAGERE AUTOMATĂ A COLOCAȚIILOR: CAZUL VERBULUI 'A FACE'

AMALIA TODIRAŞCU

[1]*LILPA, Université Marc Bloch, Strasbourg*

todiras@umb.u-strasbg.fr

Rezumat

Articolul prezintă rezultatele unui studiu al proprietăților morfologice, sintactice şi semantice ale locuțiunilor verbale, concentrîndu-se asupra celor care folosesc verbul 'a face'. Studiul este realizat pe baza unor corpusuri specializate (texte juridice) dar şi al unor corpusuri generale (ziare, texte literare, manuale de utilizare).

1. Introducere

Colocațiile sunt expresii des utilizate, care au un sens diferit de cuvintele care o compun, reprezentînd un element dificil în învățarea şi folosirea unei limbi străine, cît şi în cadrul unor sisteme de procesare automată a limbajului natural. Articolul de față îşi propune să prezinte rezultatul studiului proprietăților morfo-sintactice pentru o clasă particulară de colocații, cele care implică folosirea verbelor speciale, care intră în componența multor locuțiuni verbale (a face, a lua etc.), în vederea creării unor resurse lingvistice electronice complete pentru procesare automată. Acest studiu este realizat în cadrul unui proiect de cercetare internațional, implicînd ca parteneri Institutul de Cercetări în Inteligența Artificială al Academiei Române (România), IMS Stuttgart (Germania), INSA Strasbourg şi Universitatea Marc Bloch Strasbourg (Franța). Proiectul "Collocations en contexte: étude et analyse contrastive" (finanțat de către Agenția Universitară pentru Francofonie) are ca obiectiv realizarea unui dicționar de colocații multilingv (română, germană şi franceză), precum şi al unui sistem de extragere automată a colocațiilor, pe baza informațiilor contextuale (morfologice, sintactice). Proiectul îşi propune să identifice principalele proprietăți morfo-sintactice ale celor mai frecvente colocații, precum şi ai constituenților sintactici care se combină cu acestea, în vederea creării unui dicționar electronic multilingv. Metodologia aleasă este deja aplicată pentru limba germană (Heid & Ritz, 2005), fiind bazată pe interpretarea informațiilor contextuale asociate colocațiilor şi va fi aplicată pentru limba română şi franceză.

În cadrul acestui articol, ne concentrăm asupra studiului unei clase specifice de locuțiuni verbale, care implică verbul a face (*a face obiectul, a face apel la, a face referire la* etc.).

2. Metodologia propusă

O astfel de resursă lingvistică care descrie comportamentul colocațiilor este absolut necesară pentru un sistem de traducere automată sau pentru o platformă e-learning pentru învățarea limbilor străine. Folosirea corectă a colocațiilor pune probleme persoanelor care învață şi utilizează o limbă străină, datorită faptului că acestea au un sens care nu poate fi dedus întotdeauna pe baza sensului părților componente. De asemenea, acest tip de expresii pune

probleme deosebite unui sistem de traducere automată, o traducere cuvînt cu cuvînt nu este întotdeauna adaptată. Din aceste motive, mai multe studii s-au orientat spre o identificare automată a acestora. Astfel, metodele statistice fac ipoteza de lucru următoare, colocaţiile sunt expresii care se repetă în mod frecvent, deci criteriul frecvenţei este important pentru a detecta în mod automat colocaţiile într-un text (Smadja, 1993). O altă categorie de metode pleacă de la principiul ca o colocaţie este caracterizată de proprietăţi morfo-sintactice specifice (Hausmann, 2004). Astfel, între elementul de bază al colocaţiei şi consituentul asociat se stabilesc relaţii de dependenţă sintactică (substantiv modificat de un adjectiv, verb şi complementul său, etc.), chiar dacă expresiile sunt discontinue. O serie de metode de extragere a colocaţiilor se concentrează asupra relaţiilor sintactice care au loc între bază şi constituentul asociat (Seretan et al., 2004), dar nici una din metodele acestea nu reuşeşte identificarea corectă a colocaţiilor.

Ipoteza de lucru pe care am adoptat-o în acest proiect este aceea că vom combina metodele statistice cu cele bazate pe informaţie sintactică pentru a obţine o precizie mai mare în cazul extragerii automate a colocaţiilor. În afară de relaţiile sintactice care există între diversele elemente ale unei colocaţii, facem ipoteza că putem stabili o serie de proprietăţi morfologice şi sintactice care permit identificarea colocaţiei cu precizie. Astfel, anumite colocaţii preferă folosirea articolului definit (*face obiectul* dar nu **face un obiect*) sau al pluralului, se pot combina doar cu anumite clase de prepoziţii, sau acceptă unele adverbe între verb şi complementul său direct. Pentru a identifica proprietăţile cele mai interesante, am studiat un corpus paralel multilingv, pentru identificarea proprietăţilor pertinente care permit identificarea unor clase de colocaţii specifice. Proprietăţile identificate pe baza acestui studiu vor fi verificate pe baza unor corpusuri disponibile în fiecare din limbile studiate. Aceste proprietăţi vor fi selecţionate pentru a fi reprezentate în dicţionarul de colocaţii care va fi construit. Dicţionarul va fi integrat în cadrul unui sistem de extragere automată a colocaţillor. În continuare ne concentrăm asupra tipului de analiză lingvistică pe care dorim să îl efectuăm pentru identificarea proprietăţilor pertinente.

3. Locuţiuni verbale, predicate complexe

3.1. Locuţiuni verbale - o analiză generativă

Între colocaţiile care prezintă un interes deosebit din punctul de vedere al analizei pe care o aplicăm, ne-am oprit la o categorie specială de colocaţii, locuţiunile verbale, deoarece acestea reprezintă o clasă foarte numeroasă de expresii în toate limbile europene:

avoir besoin (a avea nevoie), porter bonheur (a purta noroc), to make a decision (a lua o decizie), a-şi aduce aminte, a face obiectul

După Gledhill (2006), aceste expresii sunt formate dintr-un verb care descrie un proces şi un complement care precizează sensul expresiei.

În cadrul unei analize lingvistice generative, verbele care intră în componenţa locuţiunilor verbale sunt considerate ca fiind golite de sens, verbele capătă un rol de verb auxiliar, sensul locuţiunii fiind propus de către complementul acesteia. Intîlnim astfel noţiunea de verb suport (Storrer, 2006), « *light verb* » (Kearns, 1989) sau „constructions converses » (Gross, 1989), care consideră că verbele din această categorie trebuie analizate doar din punct de vedere sintactic, obiectul direct fiind cel care dă sensul locuţiunii. Aceste analize ignoră constituenţii

sintactici care se combină cu acestea şi care pot oferi de asemenea informaţii importante despre gradul de libertate pe care îl avem în folosirea expresiei respective. Dacă putem folosi o locuţiune combinată cu o anumită clasă de prepoziţii, atunci probabil avem de a face cu o adevărată locuţiune verbală. Astfel, am constatat că analiza clasică de tip generativ nu este suficient de completă, şi vom considera în continuare că locuţiunile verbale pot fi tratate ca predicate complexe, cu proprietăţi sintactice şi semantice de sine stătătoare.

3.2. Predicate complexe

(Gledhill, 2006) consideră aceste construcţii verb-substantiv (notate VS) având o serie de proprietăţi sintactice şi semantice care sunt o rezultantă a proprietăţilor verbului şi al substantivului considerate separat. Astfel, construcţiile de tip VS au proprietăţi similare unui verb simplu (morfologie, diateza, complemente). În unele situaţii putem deriva un verb plecînd de la substantiv (*a face apel – a apela*, *a lua o decizie – a decide*), dar acest lucru nu este întotdeauna posibil (*a purta ghinion – *a ghinioni*), sau (*a face obiectul – a obiecta?*). În ceea ce priveşte folosirea diatezei active sau pasive, formele pasive nu sunt întotdeauna posibile (*a lua o decizie - decizia a fost luată, a face obiectul – *obiectul a fost făcut*).

În acelaşi timp, construcţiile VS au proprietăţi specifice unui substantiv : poate fi definit sau nu, poate fi modificat de către o propoziţie relativă sau poate fi transformat în substantiv. Astfel, putem observa folosirea sistematică a articolului definit sau nedefinit (*a face obiectul - - a face un obiect?*; *a face apel – a face un apel?*). Substantivul nu poate fi întotdeauna modificat de către o clauză relativă (*a luat decizia care se impunea*, dar **a facut referirea care trebuia*)

Substantivul, care este complementul verbului, joacă un rol semantic important, precizînd sensul verbului. Cum sensul este acela de proces (stare sau eveniment), complementul nu este doar obiectul verbului, ci reprezintă tipul de proces care are loc (mental sau material).

Terenul a făcut obiectul unui litigiu, care s-a rezolvat la tribunal.

Comisia de disciplină a luat o decizie rapidă privind suspendarea jucătorului.

Astfel, procesele exprimate de verbele *a face* şi *a lua* sunt procese abstracte, exprimate de data aceasta de către complementele directe (obiectul, o decizie).

Verbele *a face, a lua* sunt foarte productive, făcînd parte din componenţa multor locuţiuni verbale. De aceea am studiat proprietăţile morfologice şi sintactice ale celor mai des utilizate locuţiuni în corpusul de lucru.

4. Cîteva rezultate

4.1. Un corpus specializat

Deoarece proiectul este orientat sper un studiu comparativ al colocaţiilor în franceză, română, germană, avem nevoie de corpusuri paralele în cele 3 limbi, alineate la nivel de cuvînt şi de propoziţie. Un corpus care îndeplineşte condiţiile este corpusul AcquisCommunautaire (ACC) (http://langtech.jrc.it). Corpusul conţine 17 milioane cuvinte în limba română, 16 milioane de cuvinte în limba franceză, 15 milioane cuvinte în germană. Documentele conţin articole de lege şi directive legate de legislaţia europeană. Limbajul este specific textelor juridice, stilul

este impersonal, iar expresiile fixe sunt foarte numeroase (*se face trimitere, se face apel la…*).

Avem la dispoziție o versiune a corpusului neetichetată precum şi o versiune etichetată, care ne-a fost pusă la dispoziție de către Institutul de Cercetări pentru Inteligență Artificială al Academiei Române. Pentru a putea realiza o analiză statistică corectă, am eliminat din corpus unele elemente de structură a documentelor (grafice, tabele etc.), deoarece modifică rezultatele analizei efectuate. Versiunea etichetată a fost realizată aplicând TreeTagger (Schmid, 1994) pentru limbile franceză şi germană, iar pentru limba română TTL şi MeTT (Tufiş & Dragomirescu, 2004). Pentru a putea compara proprietățile locuțiunilor interesante identificate în corpusul AcquisCommunautaire, a trebuit să creem un corpus general (CG) care să permită verificarea datelor extrase din corpusul specializat (alcătuit din ziare, romane (1984), manuale de utilizare (Php), care însumează 2 milioane de cuvinte.

4.2. Cîteva observații asupra verbului 'a face'

Am realizat studiul cu ajutorul programului WordSmith care permite identificarea concordanțelor (contextelor unui cuvînt), sortarea acestora în funcție de contextele stîng sau drept. Contextele unui cuvînt sunt reprezentate de o fereastra de n cuvinte (am ales n=5). Am realizat o căutare folosind formele *face/fac/făcut/făceam/făceau/face*, şi am analizat rezultatele care se găsesc în dreapta verbului, imediat după verb sau la un cuvînt distanță sortate în ordinea descrescătoare a frecvenței. Printre 20 cele mai frecvente cuvinte care apar imediat după verb, regăsim multe articole (*unui, unora*), prepoziții (*dintre, din, pentru*), conjuncții (*sau*). Cum pe noi ne interesează în special construcțiile VS, am selecționat doar substantivele care apar imediat după verb şi care au sens în limba română:

Tabel 1: Cele mai frecvente construcții VS pentru verbul '*a face*', extrase din corpusul ACC

Expresie	**Număr de apariții în ACC imediat după verb**	**Număr de apariții în CG**
A face obiectul	2869	6
A face referire	1336	6
A face parte	1038	20
A face trimitere	476	7
A face dovada	209	2
A face față	86	9
A face notificarea	71	0

Pentru corpusul general, care este mult mai limitat decît ACC, frecvențele obținute sunt diferite de cele obținute pentru corpusul ACC. Astfel, marea majoritate ale expresiilor celor mai frecvente în corpusul CG sunt cele de forma *V+Prepoziție*. Doar expresia *a face parte* o regăsim între primele 20 de expresii frecvente.

În ambele corpusuri, am urmărit identificarea unor proprietăți specifice fiecărei construcții în cele două corpusuri pentru limba română. Astfel, am urmărit următoarele aspecte :

1) dacă substantivul este articulat sau nu, dacă acceptă articol definit, sau nedefinit ;

2) proprietățile complementului indirect;
3) folosirea unor prepoziții speciale;
4) folosirea adverbelor între verb şi substantiv ;

Expresia *a face obiectul* este folosită întotdeauna sub această formă (obiectul este articulat, iar articolul este definit). Ea este urmată de un substantiv în cazul genitiv/dativ în majoritatea cazurilor. Modificatorul substantivului este un substantiv reprezentînd un termen juridic sau un proces abstract (*modificării, deciziei, litigiului* etc.). Între verb şi substantiv pot apare diverse adverbe caracteristice: *deja, de asemenea, imediat* etc. Un comportament asemănător a fost constatat şi în limba franceză. Pentru expresia *a face parte*, substantivul nu este articulat, iar expresia este folosită în mod sistematic împreună cu prepoziția *din*. Pentru expresiile *a face referire*, de asemenea putem constata că substantivul este folosit mereu fără articol (definit sau nu), iar prepozițiile care urmează imediat după această expresie sunt *la* şi *în*, urmate de un substantiv indicînd locul în document (articol, paragraf, alineat etc.). De asemenea, pentru expresiile *a face față* sau *a face apel*, substantivele nu sunt articulate şi se folosesc exclusiv în această formă. Expresia *a face față* este urmată de un substantiv în cazul genitiv/dativ, modificatorul substantivului poate fi considerat ca fiind complement indirect al predicatului complex « *a face față* ». Aceste preferințe pentru una din proprietățile morfologice sau sintactice sunt identice în cele două corpusuri (chiar dacă cel general trebuie încă îmbogățit) arată că putem încerca o caracterizare a claselor de colocații cu ajutorul unui ansamblu de proprietăți identificate pe baza unei analize lingvistice.

5. Perspective

Articolul de față se concentrează doar asupra unei clase specifice de locuțiuni verbale, cele generate de verbul '*a face*'. Am identificat unele proprietăți interesante ale locuțiunilor, cum ar fi preferința pentru un anume tip de articol (definit, nedefinit) sau pentru o anumită prepoziție. Aceste proprietăți vor fi specificate în dicționarul de colocații care este în curs de realizare.

Mulțumiri. Autoarea este recunoscătoare organizației AUF (Agence Universitaire pour la Francophonie), care finanțează acest proiect în cadrul rețelei « Lexicologie, Terminologie, Traduction » pe durata mai 2006-martie 2007. De asemenea, autoarea mulțumeşte doamnei Rada Mihalcea pentru corpusul românesc pus la dispoziție de către aceasta.

Referințe bibliografice

Gross, G. (1989). *Les constructions converses du français*, Genève-Paris, Droz.

Gledhill, C. (2006). Vers une analyse systémique des locutions verbales, constructions verbo-nominales et autres prédicats complexes, *La Linguistique systémique fonctionnelle et la langue française* (D.Banks ed.), ERLA, Brest, Université de Bretagne Occidentale

Hausmann, F.J. (2004). Was sind eigentlich Kollokationnen?, *Wortverbindungen – mehr oder weniger fest* (K.Steyer ed.), pp. 309-334

Heid, U., Ritz, J. (2005). Extracting collocations and their contexts from corpora, *Proceedings of COMPLEX-2005, Conference on Computational Lexicography and Text Research*,

Budapest, juin 2005.

Kearns, K.(1989). Predicate Nominals in Complex Predicates, *MIT Working Papers in Linguistics*, 10, 123-134.

Smadja, F. (1993). Retrieving collocations from text: Xtract.*Computational Linguistics*, 19(1): 143-177.

Seretan, V., Nerima, L., Wehrli, E.(2004). A tool for multi-word collocation extraction and visualization in multilingual corpora. *Actes du congrès EURALEX'2004*, Lorient, France, Vol. 2, pp.755-766

Storrer, A. (2006). Corpus-based investigations on German support verb constructions. *Collocations and Idioms: Linguistic, lexicographic, and computational aspects*, London: Continuum Press (Fellbaum, Christiane ed.).

Tufiş, D., Dragomirescu, L. (2004). Tiered Tagging Revisited. In *Proceedings of the 4th LREC Conference*, Lisabona, pp. 39-42.

REZOLUȚIA ANAFOREI PENTRU LIMBA ROMÂNĂ

GABRIELA PAVEL[1], OANA POSTOLACHE[2], IONUȚ PISTOL[1], DAN CRISTEA[1, 3]

[1]Facultatea de Informatică, Universitatea "Al.I.Cuza", Iaşi
[2] Institute of Information Sciences, University of Southern California
[3]Institutul de Informatică Teoretică, Academia Română, Filiala Iaşi

{pavelg, ipistol, dcristea}@info.uaic.ro, oana@isi.edu

Rezumat

În această lucrare se prezintă un model de rezoluție a anaforei pentru limba română, implementat în motorul general de rezoluție RARE, şi paşii făcuți până în momentul de față în direcția rezolvării referințelor anaforice pentru limba română. Au fost adaptate aplicații existente pentru marcarea automată pe un text de intrare a informațiilor necesare rezoluției. Printre acestea, marcatorul de grupuri nominale a fost dezvoltat plecând de la reguli învățate dintr-un corpus adnotat manual. Se prezintă primele rezultate şi dezvoltări preconizate.

1. Introducere

Una dintre problemele domeniului lingvisticii computationale, care rezistă încă asalturilor cercetătorilor şi după aproape 30 de ani de eforturi continue, o reprezintă rezoluția referințelor anaforice. O referință anaforică este o secvență lexicală (numită şi *anafor*), în general realizată printr-un grup nominal, care poate avea diferite interpretări în funcție de contextul în care apare. Secvența lexicală care determină interpretarea anaforului, în general precedându-l pe acesta în text, se numeşte *antecedent*. Relația dintre anafor şi antecedent poartă numele de *relație anaforică*. În general, pentru găsirea acestei relații trebuie luate în considerare o gamă diversă de proprietăți morfologice, sintactice şi semantice ale cuvintelor, în contextul lor de utilizare. Rezoluția anaforei are însemnate aplicații în regăsirea documentară inteligentă, în sistemele de întrebare-răspuns, în sistemele de inferențe textuale etc. În grupul de Tehnologii ale Limbajului Natural de la UAIC[1] s-a încercat soluționarea acestei probleme prin crearea unui motor de rezoluție simbolic, numit RARE (Cristea, Postolache, 2002a), conceput a fi suficient de general pentru a oferi soluții de rezoluție practic în orice context. Într-un scenariu de utilizare al motorului, acesta ar urma să primească în intrare texte româneşti şi să scoată în ieşire lanțuri coreferențiale într-o adnotare XML. Cu el au fost efectuate experimente şi s-au implementat deja modele de rezoluție pentru limba engleză, care sunt la nivelul altor realizări cunoscute în lume (o rată de succes de 61%, un recall de 73%).

În lucrarea de față vom arăta cum poate fi folosit motorul RARE pentru soluționarea problemelor de rezoluție anaforică în limba română. În continuare se prezintă motorul RARE, maniera lui de funcționare pentru limba engleza, şi propunerile pentru dezvoltarea motorului pentru limba română. În final se dau rezultatele preliminare ale procesului de evaluare şi se amintesc planurile de dezvoltare ulterioară.

[1] http://consilr.info.uaic.ro/research/

2. *Motorul de rezoluție RARE*

RARE (*Robust Anaphora Resolution Engine*) este un program care funcționează ca un cadru general de rezoluție a referințelor anaforice. El poate fi considerat un motor de rezoluție pentru că, la fel ca un motor de sistem expert, pentru a funcționa asupra unui text de intrare, trebuie să fie alimentat cu un "program" ce descrie comportarea lui în operațiunile de identificare a antecedenților anaforilor. RARE are la bază o concepție asupra relației anaforice (Cristea et al., 2002a) conform căreia *nivelul textului*, populat cu expresii referențiale (notate RE în Figura 1), trebuie separat de nivelul semantic, sau *al cogniției*, unde rezidă reprezentări ale entităților de discurs (notate DE în Figura 1). Intermediar acestor două niveluri se află plasat un *nivel al proiecțiilor* informațiilor din text. Elementele acestui nivel sunt structuri de atribute (PS în Figura 1). Programul care pune în mişcare motorul RARE, numit **model**, are în componență patru elemente: un *set de atribute* care caracterizează descrierile obiectuale ale elementelor celor trei niveluri; un *set de surse de cunoaştere*, ca proceduri elementare capabile să găsească valorile corespunzătoare setului de atribute ale nivelului proiecțiilor; un *set de reguli sau euristici* capabil să răspundă la întrebarea dacă expresia referențială curentă este menționată pentru prima dată sau a mai fost menționată în textul precedent, caz în care să decidă cărui obiect de tip DE de pe nivelul cogniției îi corespunde obiectul PS curent de pe nivelul proiecției; un *domeniu de accesibilitate referențială*, care descrie un set de reguli de limitare a căutării unui antecedent (distanță şi căutare liniară versus ierarhică).

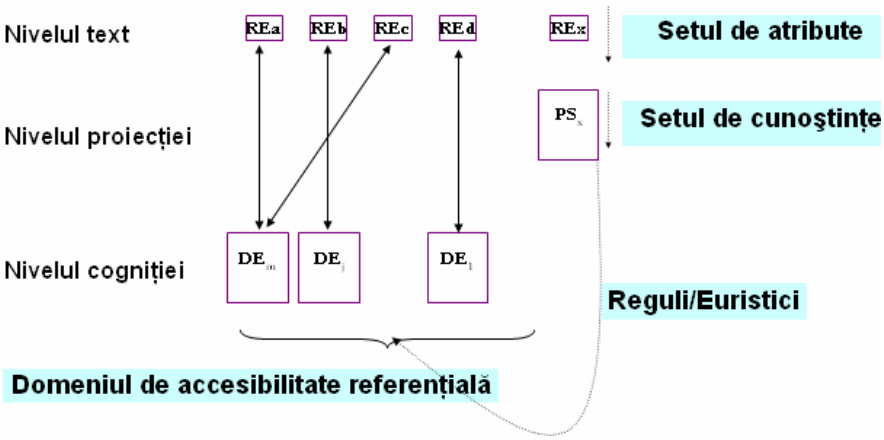


Figura 1: Motorul RARE

3. *Implementări RARE ale rezoluției pentru limba engleză*

Implementările de sisteme de rezoluție anaforică efectuate până în prezent utilizând RARE (Cristea, Postolache, 2005) au urmărit recunoaşterea antecedenților nepredicaționali pentru anafori pronominali cât şi pentru anafori generali de tip grup nominal (nume comune sau proprii) în limba engleză. Problemele de rezoluție au inclus cazuri elementare de rezoluție bazată pe acorduri în gen şi număr, dar şi cazuri speciale, cum sunt cele în care apar dezacorduri în gen şi număr între anafor şi antecedent, diferență în lemă, recuperarea antecedenților distribuiți, ori rezoluții amânate. O parte din aceste cazuri necesită implicarea unor surse de cunoaştere deosebit de sofisticate, mergând până la capacitatea de a recunoaşte şi opera cu restricții semantice şi pragmatice în context ori de a manipula cunoaştere generală

despre lume. În principiu, dacă astfel de surse pot fi găsite, atunci ele ar putea fi incluse în modelul motorului şi rezoluții de acest gen s-ar putea realiza. Din pacate însă, resursele posibil a fi angrenate actualmente într-un proces de rezoluție nu sunt capabile de mai mult decât de simulări de mică anvergură, insuficiente pentru a descrie complexitatea situațiilor reale cu care se pot confrunta anumite procese de rezoluție.

Majoritatea sistemelor de rezoluție actuale accesează în intrare un text pe care s-au plasat deja o seamă de notații (Mitkov, 2001), în principal legate de identificarea expresiilor referențiale şi de proprietățile morfo-sintactice ale acestora. Astfel de preprocesări influențează semnificativ performanțele sistemelor şi vor fi utilizate şi în implementarea descrisă în această lucrare, care este o adaptare la limba română a precedentelor create în colectivul nostru, pentru limba engleză.

Detaliem în cele ce urmează câteva elemente ale unui model RARE construit pentru limba engleză (Postolache, Cristea, 2004)[2]. Setul de atribute conține: lema cuvântului, numărul lexical, partea de vorbire, rolul sintactic al RE în propoziție sau o legătură de dependență funcțională, o indicație asupra întinderii de text acoperită de RE, setul de RE incluse (unde e cazul), indicația dacă nu nume propriu este nume de familie şi dacă numele proprii mici reprezintă nume masculine ori feminine etc. Toate aceste atribute pot primi valori prin accesarea unor proceduri (surse de cunoaştere) ce fac apel la adnotări anterior plasate în formatul de intrare al textului de către preprocesoare, sau le determină ad-hoc în faze incipiente ale procesului de rezoluție. Componenta a treia (setul de reguli ori euristici de rezoluție) implementează trei tipuri de reguli: demolatoare – responsabile de invalidarea unui anumit candidat (de exemplu, RE incluse nu pot niciodată fi coreferențiale); certificatoare – care, dimpotrivă, stabilesc cu precizie un anumit candidat ca antecedent (de exemplu, pe baza identității numelor proprii); reguli cu scor – prin a căror aplicare se măreşte sau se micşorează un scor global asociat unei perechi formate din PS-ul curent şi un DE candidat a fi considerat antecedent (o astfel de regulă, de exemplu, calculează probabilitatea ca RE-urile corespunzătoare DE-ului candidat să poată fi referite printr-unul din pronumele *he*, *she*, *it* sau *they*; o alta măreşte scorul dacă anaforul şi antecedentul se acordă în număr; legături de sinonimie şi hipernimie, determinate prin accesul la WordNet se constituie în alte reguli cu scor, ş.a.m.d.). În sfârşit, domeniul de accesibilitate implementează un tip de căutare liniară, înapoi dinspre anafor spre începutul textului, precum şi o limitare la un număr de propoziții (parametru).

4. Implementarea modelului românesc

Intrarea a fost analizată în prealabil cu serviciile web ale ICIA (Institutul de Cercetare în Inteligență Artificială al Academiei Române), care se ştie că identifică corect caracteristicile morfo-sintactice în mai bine de 98% din cazuri. Notațiile ANA obținute la ieşire conțin informații morfologice referitoare la parte de vorbire, gen, număr, caz, articol, persoană, mod şi timp verbal. Pentru uşurința de tratare ulterioară, etichetele ANA complexe, care înglobează condensat toate aceste informații morfo-sintactice, au fost decodificate la liste de perechi atribut-valoare într-o notație XML alternativă.

Exemplul de mai jos ilustrează secvența de text *o zi frumoasă* cu notațiile ICIA, ca rezultat al preprocesării utilizate de RARE:

[2] Pentru amănunte suplimentare a se vedea (Postolache, 2004).

```
<root>
     <W id="W0" LEMA="un" ANA="Tdfsr">o</W>
     <W id="W1" LEMA="zi" ANA="Ncfsrn">zi</W>
   <W id="W2" LEMA="frumos" ANA="Afpfsry">frumoasa</W>
</root>
```

Recodarea explicatorie a perechii atribut-valoare ANA corespunzătoare lexemului *zi*, produce:

```
<W ID="W1" LEMA="zi" POS="N" NUM="SG" NGEN="F" />
```

Identificarea grupurilor nominale s-a realizat cu un extractor antrenat pe corpusul *1984* (George Orwell) adnotat inițial la parte de vorbire cu *pos-tagger*-ul ICIA. Grupurile, marcate poziţional manual în corpus, au fost selectate la unică apariţie şi sortate, după care asupra lor s-au aplicat o seamă de reguli de generalizare. Rezultatul a fost o listă de şabloane, care notează poziţional marcaje ANA. De exemplu, şablonul de mai jos:

```
{0={ana=ts}, 1={ana=nsrn}, 2={ana=a}, flagPos=[0, 1, 2]}
```

este capabil să recunoască o secvenţă lexicală formată dintr-un articol nehotărât, un substantiv şi un adjectiv, ca în secvenţa *o zi frumoasă*. Aplicatorul de şabloane va încadra apoi grupul între etichete <NP></NP>. Astfel, secvenţei menţionate i se asociază următoarea adnotare ca ieşire a detectorului de grupuri nominale:

```
<NP HEADID="W1" ID="NP1">
     <W ID="W0" LEMA="un" POS="DET" NUM="SG" NGEN="F" />
     <W ID="W1" LEMA="zi" POS="N" NUM="SG" NGEN="F" />
     <W iD="W2" LEMA="frumos" POS="A" NUM="SG" NGEN="F" />
</NP>
```

Doar o parte dintre atributele modelului englezesc au fost reţinute în modelul RARE de rezoluţie anaforică pentru limba română: partea de vorbire, numărul, genul şi lema. Datorită inexistenţei la momentul actual al unui parser sintactic pentru limba română[3] atributul care indică rolul sintactic al RE-ului a fost eliminat.

În modelul românesc au fost menţinute toate regulile demolatoare şi certificatoare, ele aplicându-se identic în română ca şi în engleză. O parte a regulilor englezeşti cu scor au fost menţinute în implementarea românească, ca de exemplu cele de testare a numărului şi a lemei. În acest set au trebui însă operate şi modificări care să reflecte diferenţele care există între cele două limbi. De exemplu, în română a putut fi adăugată o regulă care măreşte scorul în cazul unei potriviri în gen între anafor şi antecedent, atribut inexistent pentru substantivele limbii engleze. Testarea potrivirii în gen poate fi luată în considerare în cazul în care pronume referă substantive (grupuri nominale) de acelaşi gen (LuperFoy, Rich, 1988). De exemplu, în secvenţa:

Maria scrie poezii… Versurile ei sunt frumoase.

aplicarea acestei reguli poate duce la concluzia corectă că pronumele personal în dativ *ei* referă *Maria*, dacă există disponibilă o sursă de cunoaştere specializată capabilă să recunoască drept feminin genul substantivului propriu *Maria*. Această sursă ar trebui să genereze următoarea notaţie pentru aceste două secvenţe lexicale:

```
<NP ID="N1">
  <W ID="W7" NUM="SG" NGEN="F" POS="N" LEMA="Maria">Maria</W>
</NP>
```

[3] În curs de elaborare, v. (Moruz et al., 2006).

```
<NP ID="N9">
  <W ID="W14" NUM="SG" NGEN="F" ROLE="" POS="N" LEMA="ea">ei</W>
</NP>
```

Aplicarea regulii de coreferențialitate pe criterii de acord în gen produce o ieşire RARE de genul:

```
<DE ID="2" reList="N1,N9" />
```

care indică că grupurile nominale N1 şi N9 sunt coreferențiale.

De asemenea, regula care verifică posibilitatea ca un antecedent nume comun să fie referit printr-unul din pronumele *he*, *she*, *it*, *they* a fost modificată pentru a putea lucra cu pronumele româneşti, *el*, *ea*, *ei*, *ele*, cât şi cu toate variantele lungi ori prescurtate ale acestora.

S-au menținut regulile de verificare a sinonimiei şi hipernimiei prin accesul la WordNet-ul românesc[4]. De exemplu, pentru aceeaşi secvență ca mai sus, în WordNet se găseşte că lemele *poezie* (la singular) si *versuri* (la plural) sunt sinonime (fac parte dintr-un acelasi synset):

```
<W LEMA="poezie" SYN_ID="ENG20-05981555-n" />
<W LEMA="versuri" SYN_ID="ENG20-05981555-n" />
```

Secvențele corespunzătoare şirurilor *poezii* şi *versurile* au următoarea reprezentare în intrarea motorului:

```
<NP ID="N2">
  <W ID="W9" NUM="PL" NGEN="F" ROLE="" POS="N"
     LEMA="poezie">poezii</W>
</NP>

<NP ID="N8">
  <W ID="W13" NUM="PL" NGEN="F" ROLE="" POS="N" LEMA
     ="versuri">Versurile</W></NP>
```

5. Evaluare şi dezvoltări ulterioare

Textul folosit în acest exercițiu a fost construit de autori în ideea de a fi scurt dar foarte bogat în expresii referențiale. El conține doar 33 de grupuri nominale, relațiile anaforice fiind adnotate manual ca lanțuri coreferențiale (Mitkov, 2001). În această etapă am fost interesați numai de realizarea unui prototip, îmbunătățirea lui urmând a se realiza în continuare.

Ca întotdeauna, probarea performanțelor unui sistem de rezoluție trebuie făcută prin compararea ieşirii motorului, care notează lanțurile coreferențiale determinate automat, cu cele ce s-au notat manual pe corpusul considerat standard. Aşa cum s-a exemplificat mai sus, în urma rulării motorului se obține o listă de entități de discurs (DE-uri), fiecare dintre acestea având asociată o listă de expresii referențiale (RE-uri) găsite a fi în relația de coreferință. În felul acesta fiecare DE are semnificația unui lanț de coreferențialitate (trivial, acestea putând a avea şi lungimea 1, reprezentând entități cu unică menționare). Corpusul de probă a fost marcat în maniera identică ieşirii motorului RARE: o listă de DE-uri, fiecare listând RE-urile lanțului.

Pentru evaluare s-a folosit evaluatorul asociat motorului. Rezultatele evaluării au următoarele valori:

```
SUCCESS_RATE: 0.696969696969697
MUC_PRECISION = 0.25
```

[4] http://multiwordnet.itc.it/online/multiwordnet

```
MUC_RECALL = 0.6
MUC_F-Measure = 0.35294117647058826
```

Pe viitor se urmăreşte în primul rând adăugarea de noi reguli în model. O astfel de regulă avută în vedere este *WhRule*, care, în varianta pentru limba engleză, se referă la depistarea antecedenților pronumelor relative.

Lista de erori obținută din rulările pe acest corpus inițial va fi folosită pentru corectarea regulilor şi recalcularea ponderilor acelora celor cu scor. Etapa următoare va consta în utilizarea unui corpus de mari dimensiuni, probabil o pare a romanului „1984”, deja adnotat parțiat la coreferințe.

Avem în vedere, totodată, realizarea unui mecanism de reglare automată a ponderilor asociate regulilor, mecanism care va folosi tehnici de învățare automată. Sperăm ca prin acest mod să realizăm un sistem hibrid simbolic-statistic, care ar trebui să aibă performanțe superioare atât unuia pur simbolic cât şi unuia pur statistic.

Nu în ultimul rând se doreşte integrarea motorului în alte proiecte care folosesc rezoluția anaforei, unul dintre acestea fiind un proiect de întrebare-răspuns pe limba română.

Referințe bibliografice

Cristea, D., Postolache, O., Dima, G.E., Barbu, C. (2002). AR-Engine – a framework for unrestricted coreference resolution. Appeared in *Proceedings of Language Resources and Evaluation Conference - LREC 2002*, Las Palmas de Gran Canaria, Spain, 29-31 May 2002, vol. VI, p.2000-2007.

Cristea D., Postolache O.D. (2005). How to deal with wicked anaphora, in António Branco, Tony McEnery and Ruslan Mitkov (editori): *Anaphora Processing: Linguistic, Cognitive and Computational Modelling*, Benjamin Publishing Books.

Moruz, A., Curteanu, N., Trandabăţ, D., Dornescu, I., Bolea, C. (2006). Parsarea predicatului (verbal / nominal) şi a clauzei (finite / nefinite) în limba română. Aplicare la parsarea FDG. În acet volum.

Postolache, O. (2004) *RARE – Robust Anaphora Resolution Engine*. Teză de disertație în Lingvistică Computațională, Facultatea de Informatică, Universitatea „Al.I.Cuza” Iaşi.

Postolache, O., Cristea, D. (2004): Designing Test-beds for General Anaphora Resolution, in Proceedings of the *Discourse Anaphora and Anaphor Resolution Colloquium* – DAARC, St. Miguel, Portugal.

Mitkov, R. (2001). *Outstanding issues in anaphora resolution*. Benjamin Publishers.

LuperFoy, S., Rich, E. (1988). An Architecture for Anaphora Resolution, *ANLP 1988*: 18-24.

INSTRUMENTE PENTRU CONSULTAREA ATLASULUI LINGVISTIC ŞI EDITAREA TEXTELOR DIALECTALE

SILVIU BEJINARIU[1], VASILE APOPEI[1], RAMONA LUCA[1], LUMINIŢA BOTOŞINEANU[2], FLORIN OLARIU[2]

[1]*Institutul de Informatică Teoretică, Academia Română, Filiala Iaşi,*
[2]*Institutul de Filologie Română „A. Philippide", Academia Română, Filiala Iaşi*

{silviub, vapopei, ramonad}@iit.tuiasi.ro, lumi_florin@personal.ro, olariuft@yahoo.com

Rezumat

Această lucrare prezintă ultimele rezultate ale cercetărilor desfăşurate în cadrul proiectului „Proiectarea şi implementarea unui sistem integrat de aplicaţii software pentru editarea textelor dialectale şi realizarea Atlasului Lingvistic Român, pe regiuni", proiect interdisciplinar al Academiei Române. Sunt prezentate noile funcţii implementate: generarea automată a indexului de cuvinte şi forme pentru dicţionarele atlasului, respectiv funcţia care permite comunicarea editorului de texte dialectale cu alte editoare de text.

1. Introducere

Proiectul intitulat *Noul Atlas lingvistic român, pe regiuni (NALR/ALRR)* a fost iniţiat în scopul radiografierii situaţiei actuale a graiurilor vorbite pe teritoriul României, anchetele făcându-se într-un număr de aproximativ 1000 de localităţi din România. Aceste atlase concretizează rezultatul cercetătorilor dialectologi din Bucureşti, Cluj-Napoca, Iaşi şi Timişoara, ele fiind importante atât pentru lingvişti, cât şi pentru istorici, geografi sau etnologi (Arvinte et al., 1987, 1997).

Scopul principal al cercetărilor actuale, în care lingvistica descoperă pertinenţa şi flexibilitatea metodelor de lucru furnizate de informatică, a fost de a crea un instrument de tip „computer aided design" care să faciliteze publicarea noilor volume ale atlasului lingvistic regional românesc. Varianta proiectării asistate de calculator a planşelor atlaselor lingvistice, orientare de mare actualitate în geolingvistica internaţională, are o serie de avantaje nete, care o recomandă spre a înlocui maniera clasică de editare: pe lângă faptul că înlătură copierea manuală, fiind mai economică sub aspectul costurilor şi al timpului de execuţie, cel mai important aspect este că ea poate constitui o sursă de informaţii stocate în format electronic pentru alte lucrări interdisciplinare.

Sistemul prezentat a stat la baza realizării planşelor pentru prospectul celui de-al 3-lea volum al *Noului Atlas lingvistic român pe regiuni. Moldova şi Bucovina*, care a fost publicat în anul 2005 sub formă de volum şi CD multimedia (Arvinte et al., 2005), (http://iit.iit.tuiasi.ro/editare_td/atlas/atlas.html).

2. Componentele atlasului lingvistic electronic

Proiectarea acestui sistem (aplicaţia ALR) a fost începută în urmă cu 6 ani şi principalele funcţii, prezentate deja cu alte prilejuri sunt următoarele:

- crearea şi întreţinerea dicţionarelor atlasului (Bejinariu et al., 2002),
- generarea automată, editarea şi tipărirea în diferite formate a planşelor de tip „hartă lingvistică", respectiv „material necartografiat" (Florea et al., 2002),
- editarea, formatarea şi tipărirea de texte dialectale (aplicaţia EditTD) (Apopei et al., 2003),
- generarea automată de pagini HTML pentru conţinutul dicţionarelor,
- consultarea sincronizată a ediţiilor mai vechi ale atlaselor.

Transcrierea fonetică specifică limbii române

Pentru a putea reda cât mai fidel toate nuanţele rostirii, transcrierea fonetică s-a dovedit a fi un instrument fiabil şi, de aici, varietatea alfabetelor fonetice utilizate de specialişti. Pe lângă transcrierea fonetică internaţională realizată cu Alfabetul Fonetic Internaţional (IPA), specialiştii au dezvoltat sisteme proprii anumitor domenii lingvistice.

În cazul limbii române, transcrierea fonetică presupune folosirea de simboluri asociate sunetelor primare: 68 de variante vocalice (17 vocale simple, fiecare dintre ele având şi câte 3 variante accentuate) şi 50 de variante consonantice (fig. 1). Simbolurilor primare le sunt asociate semne diacritice ilustrând fenomene fonetice specifice. Acestea sunt în număr de 12, organizate în 5 grupe, în cazul vocalelor, respectiv în număr de 9, organizate tot în 5 grupe, în cazul consoanelor (fig. 2).

simple	diacritice					
a	ä	ă	â	ȧ (*)	å	
e	ë	ĕ				
i			î			ĩ
o	ö	ŏ				
u	ü		û			

b, c,ć,ĉ,č,c̆, d,đ,ḍ,ȡ, f, g,ĝ,ġ,ğ,g̃, h,ɧ,χ,γ, j, k, l,ŧ,
m,ṃ, n,ṇ,ṉ,ŋ, p, r,ŕ,ɾ,ρ, s,ś,ŝ,ş,σ, t,ϑ,ţ, v, w, z,ź,ẑ, y

Figura 1. Simbolurile primare folosite în transcrierea fonetică a limbii române
Sunetul vocalic marcat cu (*) nu permite aplicarea de fenomene fonetice

Datorită numărului foarte mare de caractere primare, s-a decis folosirea modului Unicode pentru codificarea caracterelor. În cazul vocalelor, un calcul simplu arată că există un număr de 359 combinaţii de fenomene fonetice. De aici rezultă necesitatea de a proiecta un număr de 359 de fonturi, respectiv 359*17*4=24412 simboluri grafice, doar pentru vocale.

Fenomenele fonetice pot fi plasate deasupra, sub, dar şi lateral faţă de simbolul de bază. În plus, unele simboluri pot fi plasate deasupra, sau dreapta-sus în raport cu simbolul precedent din text.

Fenomene asociate vocalelor	
Grupă	**Fenomen**
Durată	Scurtime
	Semilungime
	Lungime
Nazalizare	Seminazalizare
	Nazalizare
Ocluzie glotală	Coup de glotte
Deschidere	Închidere
	Semideschidere
	Deschidere
	Deschidere mare
Afonizare	Semiafonizare
	Afonizare

Fenomene asociate consoanelor	
Grupă	**Fenomen**
Durată	Semilungime
	Lungime
Palatalizare	Semipalatalizare
	Palatalizare
	Palatalizare mare
Explozie	Explozie
Caracter silabic	Caracter silabic
Afonizare	Semiafonizare
	Afonizare

Figura 2. Fenomenele fonetice folosite în transcrierea fonetică

Din acest motiv, am decis realizarea unui sistem de generare on-line a imaginii simbolurilor cărora le sunt aplicate fenomene fonetice, prin sinteza imaginilor componente. Pentru editarea textelor folosind transcrierea fonetică specifică limbii române, utilizatorul trebuie să introducă din tastatură simbolul de bază, aplicarea fenomenelor realizându-se prin selectarea acestora direct de pe bara de instrumente. Acest mod de desenare a simbolurilor a permis reducerea numărului de fonturi folosite la numai 2, ambele derivate din fontul „Arial".

3. Generarea automată a indexului de cuvinte şi forme

Unul dintre instrumentele care s-au dovedit a fi necesare cercetătorilor lingvişti este generatorul automat al indexului de cuvinte şi forme. Generarea unui astfel de index presupune identificarea aparițiilor unui sunet sau grup de sunete în dicționarul de transcrieri fonetice.

Prima etapă a acestui proces constă în stabilirea parametrilor de căutare:

- **filtrul „cuvânt"** – stabileşte dacă căutarea se realizează între transcrierile fonetice ale unui anumit cuvânt de bază, sau în întreg dicționarul;
- **filtrul „punct de anchetă"** – permite restrângerea căutării la un singur punct de anchetă;
- **căutare în…** – permite stabilirea câmpurilor din dicționar în care se face căutarea;
- **mod căutare** – este folosit pentru a specifica modul în care se realizează căutarea, ca „text" sau ca „transcriere fonetică";
- **forma de căutat** – permite utilizatorului să editeze textul ale cărui apariții doreşte să le identifice.

Caracterele conținute în textele transcrise fonetic sunt însoțite de două grupe de atribute: pe de o parte, fenomenele fonetice, pe de altă parte, atribute care specifică modul de poziționare şi desenare. În funcție de parametrii specificați înainte de generarea indexului, aceste atribute sunt luate sau nu în considerare în cursul procesului de căutare.

Procesul de căutare este finalizat prin sintetizarea informaţiei privitoare la ocurenţele formei căutate sub forma unei liste (fig. 3), care poate fi tipărită sau poate fi folosită pentru identificarea poziţionării acestora în dicţionar.

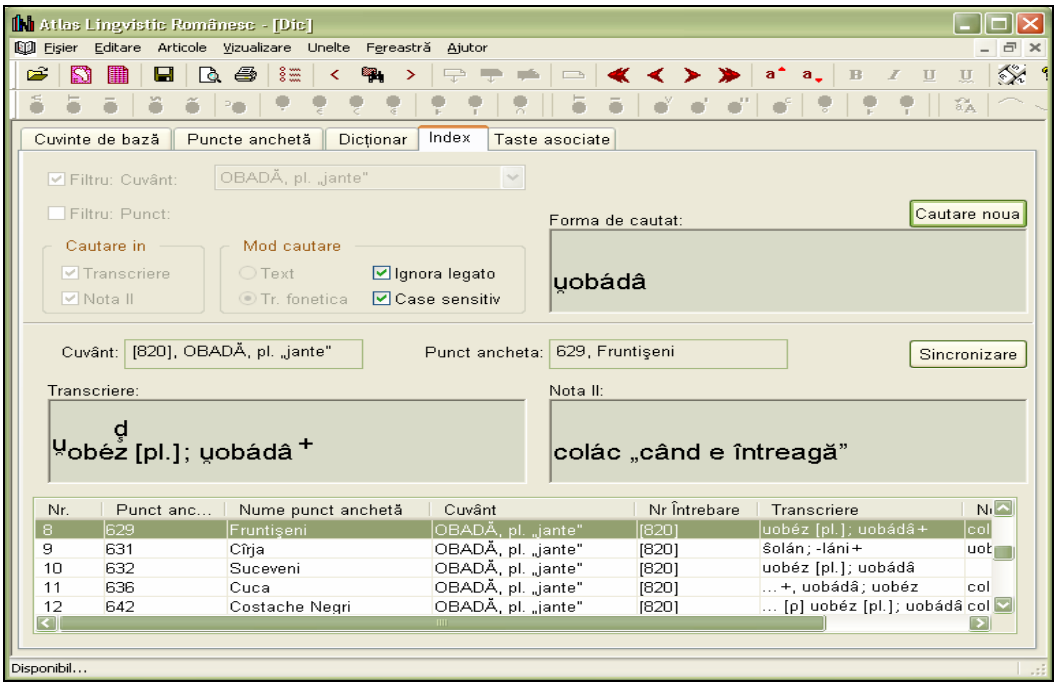


Figura 3. Indexul de forme, generat de aplicaţia ALR

4. Comunicarea editorului de texte dialectale cu alte editoare de text

Una dintre probleme apărute în cursul proiectării sistemului pentru editarea textelor dialectale a fost aceea de a oferi posibilitatea transferului de fişiere între aplicaţia EditTD şi alte editoare de text, având în vedere faptul că există deja în alte proiecte, texte editate folosind alte modalităţi de operare, care folosesc fonturi proiectate special în acest scop. În general, au fost folosite două tipuri de fonturi:

- fonturi conţinând imaginile caracterelor împreună cu un fenomen fonetic specific fontului respectiv;
- fonturi prezentând doar fenomenele fonetice sub formă de caractere separate, introduse după caracterul căruia îi sunt aplicate.

Pentru realizarea conversiei între textele editate cu alte editoare şi aplicaţia EditTD a fost ales ca intermediar formatul RTF (Rich Text Format), recunoscut pe mai multe platforme şi acceptat de editoarele din familia Microsoft.

Realizarea efectivă a importului fişierelor de tip RTF în formatul specific aplicaţiilor ALR şi EditTD presupune parcurgerea următoarelor etape:

- selecţia fişierului de intrare,
- definirea tabelei de conversie (selecţia acesteia în cazul în ea există deja),
- stabilirea regulilor de conversie globale, la nivel de font,
- stabilirea regulilor de conversie la nivel de detaliu (caracter),
- conversia propriu-zisă,

- analiza rezultatului conversiei şi eventuale corecţii.

Tabela de conversie RTF este o colecţie de reguli prin care fiecărei perechi de forma ***{caracter_iniţial, font_iniţial}*** i se asociază o structură de forma

{ caracter, font, fenomen_fonetic, atribut_poziţie }

Sistemul permite specificarea regulilor de conversie:

- la nivel global, în cazul când caracterelor din fontul iniţial le sunt asociate aceleaşi caractere din fontul de bază, cărora le sunt aplicate unul sau mai multe fenomene fonetice,
- la nivel de detaliu, pentru indicarea excepţiilor existente între regulile definite la nivel global sau pentru indicarea unor noi reguli.

Definirea unei reguli de traducere presupune selectarea fontului iniţial şi a caracterului iniţial folosit în regula de traducere, selectarea fontului şi a caracterului în care se face traducerea, selectarea fenomenelor fonetice specifice ce urmează a fi aplicate în momentul traducerii, precum şi poziţia caracterului în cazul în care aceasta este modificată în momentul traducerii.

După definirea tabelei de traducere se poate trece la conversia efectivă a fişierului RTF (fig. 4).

Menţionăm că procesul de traducere din formatul RTF în formatul intern al aplicaţiei pentru editarea textelor dialectale este realizat cu păstrarea atributelor de formatare a caracterelor şi paragrafelor. În plus, sunt tratate special caracterele diacritice specifice limbii române, deoarece aplicaţia noastră foloseşte codificarea Unicode a caracterelor, în timp ce în formatul RTF codificarea caracterelor se face pe 8 biţi, fiind specificat însă setul de caractere regional care trebuie folosit.

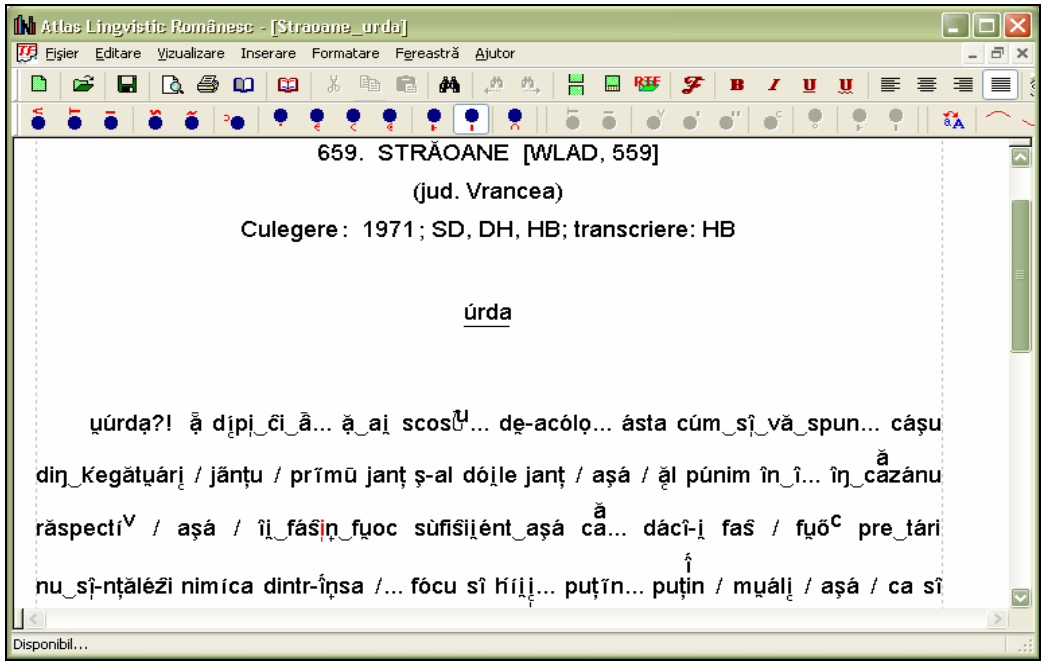


Figura 4. Rezultatul aplicării procedurii de conversie

5. Concluzii

Sistemul proiectat pentru modelarea *Atlasului Lingvistic Român pe Regiuni* este funcțional şi a fost folosit pentru publicarea prospectului celui de al 3-lea volum al *Atlasului Lingvistic Român pe Regiuni, Moldova şi Bucovina*. Acest prospect a fost publicat sub formă de volum tipărit şi CD multimedia, fiind publicat în acelaşi timp şi pe internet, la adresa http://iit.iit.tuiasi.ro/editare_td/atlas/atlas.html.

Noile funcții implementate, generarea automată a indexului de cuvinte şi forme, respectiv funcția care permite comunicarea editorului de texte dialectale cu alte editoare text, se dovedesc foarte utile în cercetare pentru lingvişti şi dialectologi.

În concluzie, putem spune că sistemul realizat reprezintă la ora actuală o încununare a unor eforturi care au început acum mai bine de 130 de ani (primul sistem de transcriere fonetică aplicat limbii române a fost pus la punct de Émile Picot, în anul 1873).

Aplicarea tehnologiei informatice la transcrierea fonetică a etno- şi socio-textelor uşurează extrem de mult procesul editării permițând în acelaşi timp realizarea unor colecții de informații lingvistice în format electronic, ce pot fi folosite în cadrul altor cercetări interdisciplinare.

Referințe bibliografice

Arvinte, V., Dumistrăcel, S., Florea, I.A., Nuță, I., Turculeț, A. (1997). *Noul Atlas lingvistic român, pe regiuni. Moldova şi Bucovina*, Bucureşti, Editura Academiei, vol. I, 1987; vol. II.

Arvinte, V., Dumistrăcel, S., Florea, I.A., Nuță, I., Turculeț, A. (2005). *Noul Atlas lingvistic român, pe regiuni. Moldova şi Bucovina*, III. *Prospect*, Iaşi, Editura Universității „Alexandru Ioan Cuza".

Florea, I.A., Apopei, V., Olariu, F.T., Bejinariu, S. (2002). „Editarea asistată de calculator a atlaselor lingvistice şi a textelor dialectale", în *Identitatea limbii şi literaturii române în perspectiva globalizării*, Editura Trinitas, Iaşi, 2002, p. 211-232

Bejinariu, S., Apopei, V., Roman, M. (2002). *Mediu pentru editarea transcrierilor fonetice în Limba Română. Realizarea Atlasului Lingvistic Român pe Regiuni*, *Limba Română în Societatea Informațională, Societatea Cunoaşterii*, Editura Expert, Bucureşti, p. 423-440.

Apopei, V., Rotaru, F., Bejinariu, S., Olariu, F. (2003). Electronic Linguistic Atlases, *Proceedings of the International Conference on Information and Knowledge Engineering IKE'03*, June 23, Las Vegas, Nevada, USA, Volume 2, ISBN 1-932415-08-4, p. 628-633

[http://iit.iit.tuiasi.ro/editare_td/atlas/atlas.html] *Noul Atlas Lingvistic Român pe Regiuni, Moldova şi Bucovina III.*

GENERARE DE CONCORDANȚE PENTRU DICȚIONARUL LIMBAJULUI POETIC EMINESCIAN

MIHAELA BRUT[1], DUMITRU IRIMIA[2], OANA PANAIT[2]

[1]*Facultatea de Informatică, Universitatea "Al.I.Cuza" Iaşi*

[2]*Facultatea de Litere, Universitatea "Al.I.Cuza" Iaşi*

mihaela @info.uaic.ro, tirimia@uaic.ro, oana_pan@yahoo.fr

Rezumat

În urma unui efort susținut de câțiva ani buni de zile, dicționarul limbajului poetic eminescian a fost finalizat, fiind deja disponibil specialiștilor şi tuturor iubitorilor de poezie. La dezvoltarea lui a fost utilizată aplicația „Concordanțe eminesciene", dezvoltată de un colectiv din Cluj la începutul anilor 90. Lipsită de scalabilitate din cauza tehnologiei depăşite pe care o utilizează, această aplicație a fost totuşi singurul instrument disponibil pentru procesarea limbii române de care s-a putut folosi colectivul care a lucrat la dezvoltarea dicționarului.

Articolul de față îşi propune să prezinte mecanismul intern de funcționare a acestei aplicații, paşii de procesare a textului eminescian, încercând să ofere o propunere de rescriere a ei apelându-se la tehnologiile Web-ului semantic.

1. Introducere

Dicționarul limbajului poetic eminescian este un proiect inițiat în anii '90 de Facultatea de Litere a Universității „Al. I. Cuza" din Iaşi, vizând două componente majore:
- *Concordanțele poeziilor eminesciene* - urmându-se modelul din lexicografia poetică europeană;
- *Semne şi sensuri poetice* - având-se drept model *Dictionnaire des Symboles* (Chevalier & Gheerbrant, 1994)

În privința primei componente a proiectului, până acum au fost finalizate şi tipărite *Concordanțele poeziilor antume* (Irimia, 2004) şi *Concordanțele poeziilor postume* (Irimia, 2006), de un real folos celor care doresc să se apropie de poezia eminesciană, să înțeleagă şi urmărească modul în care cuvintele limbii române au fost încărcate de poeticitate, de noi sensuri şi semnificații. O pătrundere de profunzime a forței semantice a limbajului poetic eminescian nu poate fi atinsă, însă, fără a urmări întreaga creație a poetului, de aceea dezvoltarea pe viitor a concordanțelor din teatru, proză literară, publicistică, critică literară/teatrală /muzicală, corespondență, însemnări manuscrise ar fi de bun augur pentru critica literară şi pentru iubitorii de poezie. În alte literaturi sunt demult disponibile concordanțele complete ale operei poeților reprezentativi: G.Leopardi, E. Montale, G. Pascoli, D'Annuzio, G. Ungaretti (Italia); William Blake, Lord George Gordon Byron, John Butler Yeats (Anglia); Federico Garcia-Lorca (Spania); Emily Dickinson (SUA).

Vom prezenta în continuare instrumentele informatice utilizate în generarea primelor două seturi de concordanțe eminesciene, accentuând deficiențele acestor instrumente şi prezentând şi o propunere de rescriere şi îmbunătățire a lor.

2. CONCORD şi SILEX

Pentru obținerea concordanțelor eminesciene au fost utilizate două instrumente informatice dezvoltate de colective aparținând Universității "Babeş-Bolyai" din Cluj:
- *CONCORD – Sistem de lematizare automată şi generare a concordanțelor*, coordonat de prof. dr. Sanda Cherata (Cherata, 1996);
- *SILEX – Sistem lexical informatizat*, sub coordonarea cerc. *Teodor VUŞCAN*; acest sistem a fost utilizat pentru realizarea analizei morfologice automatizate a cuvintelor din poezia eminesciană, fiind de fapt integrat în sistemul CONCORD (Vuşcan, 1996).

Ambele sisteme sunt proiectate utilizându-se sistemul de gestiune a bazelor de date Foxpro 2.6. A fost luată ca reper *Poezia* eminesciană - edițiile Perpessicius (1952) şi D.Murărașu (1970-1972) -, care a fost transformată într-o bază de date FoxPro numită ME.POE, conținând câte o înregistrare pentru fiecare poezie. Structura acestei baze de date include codul autorului, codul volumului, codul ciclului de poezii, codul subciclului, codul poeziei, titlul poeziei, subtitlul acesteia, dedicația scrisă de Eminescu pe marginea poeziei, *motto*-ul acesteia, notele de final ale poeziei, precum şi textul integral al poeziei (inclus într-un câmp de tip memo):

Fiecare poezie este lematizată separat, lemele sunt analizate sintactic utilizându-se SILEX, iar rezultatul acestor operații este inclus în câte o bază de date asociată fiecărei poezii. Într-o astfel de bază de date, este alocată câte o înregistrare fiecărui cuvânt din poezie, fiind incluse: codul volumului, al ciclului,al subciclului, al poeziei, numărul versului în cadrul poeziei, numărul liniei pe care este afişat acest vers, numărul cuvântului în cadrul versului, codul contextului (numărul versului asociat cuvântului - diferă în cazul cuvintelor afişate pe rândul următor al versului din care fac parte, în cadrul versurilor mai lungi), lema de care aparține cuvântul curent, categoria gramaticală, atributul eventual al acestei categorii (de exemplu, în cazul categoriei *adjectiv* poate exista atributul *posesiv* etc.), variantele eventuale ale lemei (de exemplu, pentru seară - sară), numărul caracterului din cadrul versului de la care începe cuvântul şi numărul caracterului la care acesta se sfârşeşte.

Se poate observa cantitatea imensă de informație redundantă, repetată la fiecare cuvânt în parte. În cazul în care ar fi fost utilizată o structură XML pentru stocarea poeziilor eminesciene, fiecare dintre informații ar fi trebuit furnizată o singură dată, deoarece informațiile ce țin, de exemplu, de un volum ar fi fost furnizate ca sub-elemente ale elementului `<Cod_vol>` etc.

Utilizându-se lematizarea şi analiza sintactică automatizată a poeziilor în maniera descrisă mai sus, se poate efectua o primă generare a concordanțelor, utilizându-se opțiunile *Gen_Conc* → *Integrare Volum*, apoi *Gen_Conc* → *Integrare Operă* puse la dispoziție de CONCORD.

Din cauza faptului că unele leme nu sunt recunoscute corect de analizor, lingviştii trebuie să efectueze o primă corectură pe hârtie a concordanțelor generate, în special în privința analizei gramaticale şi a recunoaşterii corecte a unei leme în formele flexionare în care aceasta apare în poezii. Aceste corecturi sunt apoi operate electronic utilizând opțiunea *Lem_Poem* → *Corect_Lem* din cadrul CONCORD. Pentru operarea corecturii asupra unei leme, trebuie furnizat codul volumului, codul poeziei, numărul versului în care apare lema respectivă (aceste informații se găsesc pe pagina listată unde s-a efectuat corectura), dar şi numărul cuvântului ce urmează a fi corectat în cadrul versului (în acest caz, numărarea făcându-se manual şi fiind dificilă mai ales în cazul cuvintelor aflate după poziția a zecea în cadrul versului).

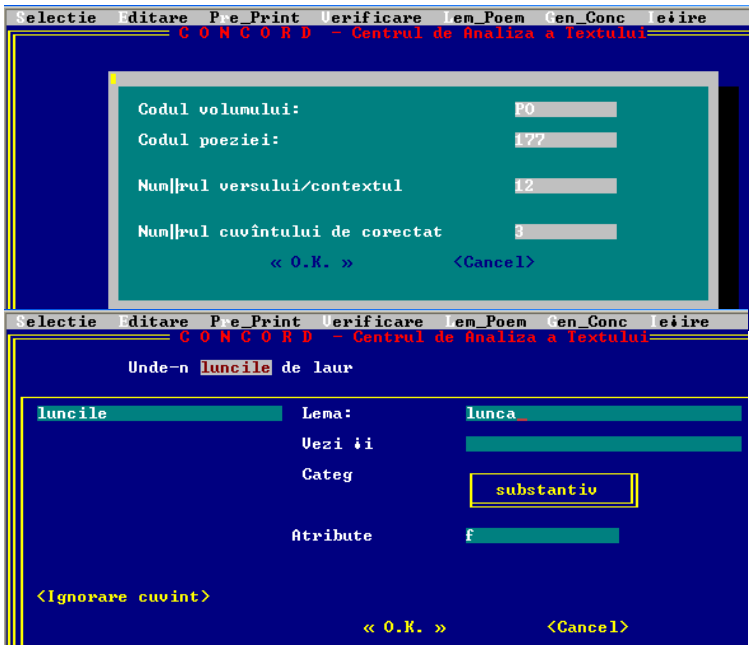


Figura 1: Operarea corecturii asupra unei leme în cadrul CONCORD

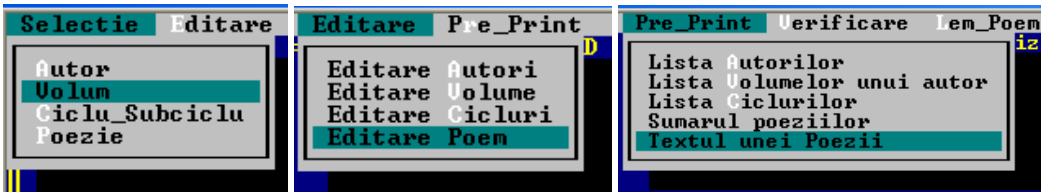


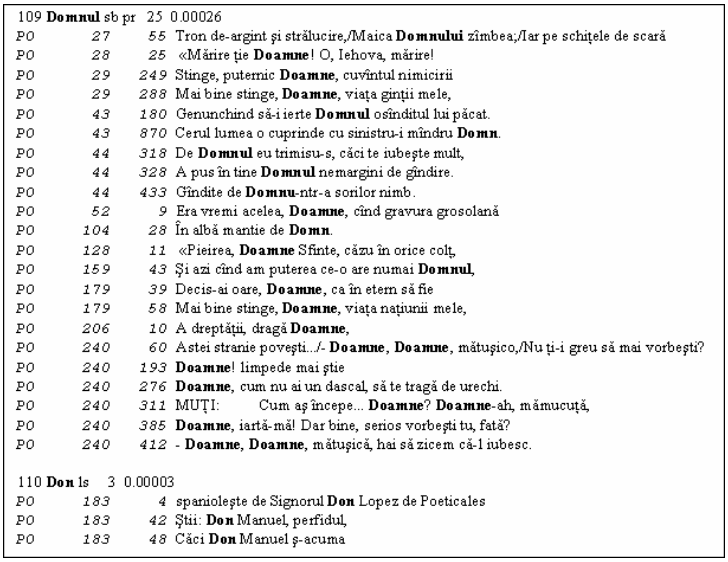
Figura 2: Facilități oferite de CONCORD

CONCORD pune la dispoziție câteva facilități suplimentare ce pot fi utile pentru a uşura munca centrală de operare a corecturilor. De exemplu, opțiunea *Selecție* → *Volum* permite selectarea unui volum, al cărui cod va apărea automat în caseta Codul *volumului* din fereastra de corectare a unei leme, nemaitrebuind să fie introdus manual la fiecare lemă în parte. Pentru a efectua corecturi asupra textului unei poezii (în cazul în care au fost observate erori), în locul modificării manuale a bazei de date corespunzătoare poate fi utilizată opțiunea *Editare* → *Editare Poem*. Pentru a vizualiza într-o manieră formatată textul unei poezii există facilitatea *Editare* → *Editare Poem*.

După operarea corecturii asupra tuturor lemelor, se realizează o nouă generare a concordanţelor, utilizându-se aceleaşi opţiuni *Gen_Conc* → *Integrare Volum*, apoi *Gen_Conc* → *Integrare Operă* oferite de CONCORD. Rezultatul acestor operaţii este alcătuit din următoarele fişiere:
- Me_C.tlm → conţinând concordanţele întregii poezii eminesciene;
- Me_ct_C.txt → lista lemelor, ordonate pe clase lexico-gramaticale;
- Me_fr_C.txt → lemele ordonate descrescător după frecvenţa de utilizare;
- Me_Lm_C.txt → lemele ordonate alfabetic.

Deoarece aceste fişiere sunt de tip text, au fost folosite anumite codificări, anumite convenţii de reprezentare a caracterelor româneşti, a caracterelor speciale sau a indicaţiilor de formatare. În plus, lemele a căror categorie gramaticală nu a fost recunoscută de către program au fost marcate cu menţiunea „??” pentru a fi analizate cu atenţie la corectura manuală.

Codificările utilizate în fişierele prezentate mai sus sunt interpretate în cadrul aplicaţiei *MS Word*, utilizând două macrocomenzi:
- *ConvConc* → aplicată fişierului Me_C.tlm şi având drept rezultat documentul Word formatat incluzând concordanţele întregului volum.
- *Conv_Car_Conc* → aplicată fiecăruia dintre fişierele Me_ct_C.txt, Me_fr_C.txt, Me_Lm_C.txt pentru a se obţine listele cu diverse ordonări ale lemelor sub formă tot de document Word formatat.



109 **Domnul** sb pr 25 0.00026
P0 27 55 Tron de-argint şi strălucire,/Maica **Domnului** zîmbea;/Iar pe schiţele de scară
P0 28 25 «Mărire ţie **Doamne**! O, Iehova, mărire!
P0 29 249 Stinge, puternic **Doamne**, cuvîntul nimicirii
P0 29 288 Mai bine stinge, **Doamne**, viaţa ginţii mele,
P0 43 180 Genunchind să-i ierte **Domnul** osînditul lui păcat.
P0 43 870 Cerul lumea o cuprinde cu sinistru-i mîndru **Domn**.
P0 44 318 De **Domnul** eu trimisu-s, căci te iubeşte mult,
P0 44 328 A pus în tine **Domnul** nemărgini de gîndire.
P0 44 433 Gîndite de **Domnu**-ntr-a sorilor nimb.
P0 52 9 Era vremi acelea, **Doamne**, cînd gravura grosolană
P0 104 28 În albă mantie de **Domn**.
P0 128 11 «Pieirea, **Doamne** Sfinte, căzu în orice colţ,
P0 159 43 Şi azi cînd am puterea ce-o are numai **Domnul**,
P0 179 39 Decis-ai oare, **Doamne**, ca în etern să fie
P0 179 58 Mai bine stinge, **Doamne**, viaţa naţiunii mele,
P0 206 10 A dreptăţii, dragă **Doamne**,
P0 240 60 Astei stranie poveşti.../- **Doamne**, **Doamne**, mătuşico,/Nu ţi-i greu să mai vorbeşti?
P0 240 193 **Doamne**! limpede mai ştie
P0 240 276 **Doamne**, cum nu ai un dascal, să te tragă de urechi.
P0 240 311 MUŢI: Cum aş începe... **Doamne**? **Doamne**-ah, mămucuţă,
P0 240 385 **Doamne**, iartă-mă! Dar bine, serios vorbeşti tu, fată?
P0 240 412 - **Doamne**, **Doamne**, mătuşică, hai să zicem că-l iubesc.

110 **Don** ls 3 0.00003
P0 183 4 spaniolește de Signorul **Don** Lopez de Poeticales
P0 183 42 Ştii: **Don** Manuel, perfidul,
P0 183 48 Căci **Don** Manuel ş-acuma

Figura 3: Extrase din documentele de concordanţe

Documentele formatate obţinute sunt din nou parcurse atent de către lingvişti. Dacă mai apar corecturi, acestea trebuie din nou operate în cadrul CONCORD, fiind necesară o nouă generare a concordanţelor conform paşilor expuşi mai sus. În cazul în care lingviştii îşi dau acordul în privinţa formei finale a concordanţelor, cele patru documente Word formatate trebuie integrate într-unul singur, ce urmează a fi aranjat în pagină conform cerinţelor specificate de editura la care va fi tipărit volumul de concordanţe.

3. Concluzii

Programele SILEX şi CONCORD au oferit suportul necesar pentru obținerea concordanțelor poeziei eminesciene, însă ar fi necesară o reabordare a acestor sisteme din perspectiva tehnologiilor moderne. Principalele deficiențe ale celor două aplicații, constatate pe parcursul utilizării efective a lor în cadrul *Dicționarul limbajului poetic eminescian* sunt:

- Lipsa de documentare a codului FoxPro 2.6, astfel încât efectuarea unor modificări pe codul sursă al aplicațiilor este foarte dificilă;

- Stocarea redundantă a informațiilor, ilustrată de structura bazelor de date ce conțin poezia eminesciană;

Rang	Lema	Cl Atrib.	Nr. ap	Frecv.
90	Roma	sb pr	44	0.00023
	barbă	sb f	44	0.00023
	casă	sb f	44	0.00023
	copac	sb m	44	0.00023
	laur	sb m	44	0.00023
	masă	sb f	44	0.00023
	mei	sb m	44	0.00023
	patimă	sb f	44	0.00023
	pânză	sb f	44	0.00023
	preot	sb m	44	0.00023
	râs	sb n	44	0.00023
	suspin	sb n	44	0.00023
	veac	sb n	44	0.00023
	vin	sb n	44	0.00023
91	amar	sb n	42	0.00022
	arbore	sb m	42	0.00022
	bucurie	sb f	42	0.00022
	fluviu	sb n	42	0.00022
	iarnă	sb f	42	0.00022
	joc	sb n	42	0.00022
	miere	sb f	42	0.00022
	mijloc	sb n	42	0.00022
	mister	sb n	42	0.00022
	plan	sb n	42	0.00022
	rând	sb n	42	0.00022
	sărutare	sb f	42	0.00022
	sicriu	sb n	42	0.00022
	sunet	sb n	42	0.00022
	uitare	sb f	42	0.00022
	vers	sb n	42	0.00022
	zid	sb n	42	0.00022
92	Ana	sb pr	40	0.00020

Rang	Lema	Cl Atrib.	Nr. ap	Frecv.
1	în	pp	6558	0.03360
2	i	cj	6432	0.03295
3	fi	vb p	5864	0.03004
4	de	pp	7322	0.03751
5	cu	pp	3558	0.01823
6	eu	pn pers	3546	0.01817
7	al	ar pos	3280	0.01680
8	tu	pn pers	3200	0.01639
9	pe	pp	2806	0.01437
10	avea	vb a	2752	0.01410
11	să	cj	2628	0.01346
12	ea	pn pers	2566	0.01315
13	sine	pn refl	2532	0.01297
14	nu	av	2406	0.01233
15	ce	pn relat	2400	0.01229
16	un	ar m	1856	0.00951
17	un	ar f	1840	0.00943
18	din	pp	1742	0.00892
19	el	pn pers	1722	0.00882
20	la	pp	1374	0.00704
21	ca	av	1336	0.00684
22	lume	sb f	1236	0.00633
23	ei	pn pers	1116	0.00572
24	cÇ	cj	1076	0.00551
25	tău	aj pos	1042	0.00534
26	când	av	934	0.00478
27	mai	av	848	0.00434
28	ochi	sb m	828	0.00424
29	vedea	vb	790	0.00405
30	viață	sb f	786	0.00403
31	prin	pp	758	0.00388
32	vrea	vb	754	0.00386
33	cum	av	740	0.00379
34	lui	ar m	712	0.00365
35	meu	aj pos	666	0.00341

Figura 4: Generarea concordanțelor în format text utilizând CONCORD

- Fluxul de operații necesare obținerii concordanțelor este unul inflexibil şi mare consumator de timp. Astfel, pentru a vizualiza o corectură efectuată în cadrul CONCORD trebuie repetați toți paşii de generare a concordanțelor, vechile instanțe de documente conținând concordanțele fiind neutilizate.

Precum am evidențiat şi mai sus, structurarea poeziei eminesciane în format XML ar elimina deficiența menționării redundante a informațiilor. Utilizarea unui sistem de lematizare cu marcatori XML ar face disponibilă analiza gramaticală a lemelor şi altor aplicații de procesare a textelor româneşti. Sistemul de generare a concordanțelor ar putea utiliza, în acest caz, o procesare DOM, beneficiind de bibliotecile de funcții de procesare existente deja în mai multe limbaje de programare.

Actualmente există mai multe instrumente de generare a concordanţelor[1]. Problema majoră este aceea că fiecare astfel de instrument impune anumite restricţii legate de formatul fişierelor ce vor fi procesate. O posibilă soluţie de reproiectare a CONCORD ar presupune selectarea unui astfel de sistem (sau proiectarea de la zero), integrarea unui analizor gramatical, precum şi reproiectarea sistemului de reprezentare a datelor, în cazul în care cel original nu corespunde formatului XML în care avem la dispoziţie opera poetică pe care dorim să o procesăm.

Mulţumiri. Autorii sunt recunoscători colectivului de cercetători de la Universitatea "Babeş-Bolyai" din Cluj care le-a pus la dispoziţie sistemul CONCORD pentru procesarea concordanţelor eminesciene.

Referinţe bibliografice

Cherata, S., (1996). CONCORD: Sistem de realizare a concordanţelor textelor poetice româneşti. *Limbaj şi Tehnologie*, Dan Tufiş editor, Ed. Academiei Române, Bucureşti, 1996, pp. 215-220.

Chevalier, J., Gheerbrant, A. (1994). *Dictionnaire des symboles : Mythes, rêves, coutumes, gestes, formes, figures, couleurs, nombres*, Robert Laffont Edition.

Irimia, D. coord (2004). *Dicţionarul limbajului poetic eminescian. Concordanţele poeziilor antume*, vol. I-II, Editura Hyperion, Botoşani.

Irimia, D. coord (2006). *Dicţionarul limbajului poetic eminescian. Concordanţele poeziilor postume*, vol. I-IV, Editura Univ. „Alexandru Ioan Cuza" Iaşi, 2006.

Irimia, D. coord (2005). *Dicţionarul limbajului poetic eminescian. Semne şi sensuri poetice. I. Câmpul semantica ARTE*, Editura Univ. „Alexandru Ioan Cuza" Iaşi.

Vuşcan, T., (1996). SILEX - sistem lexico-morfologic computerizat pentru limba română. *Limbaj şi Tehnologie*, Dan Tufiş editor,Ed. Academiei Române, Bucureşti, 1996, pp. 209-214.

[1] Exemple de instrumente de generare a concordanţelor: *AntConc*, Waseda University, Japonia: http://www.antlab.sci.waseda.ac.jp/antconc_index.html; *Concordance*, R.J.C. Watt of Dundee University: http://www.dundee.ac.uk/english/wics/wics.htm ; *Monoconc*, Athelstan: http://www.athel.com/mono.html#monopro; *Wordsmith*, Mike Scott, Oxford University: http://www.oup.com/elt/catalogue/isbn/6890?cc=gb

CREAREA UNUI GENERATOR MORFOLOGIC PENTRU VERBELE DIN LIMBA ROMÂNĂ

ANTONINA BÎRLĂDEANU, NATALIA BURCIU

Facultatea de Calculatoare, Informatică şi Microelectronică, Universitatea Tehnică a Moldovei, Chişinău

{toni_birlad, natusicb}@yahoo.com

Rezumat

Resursele computerizate pentru limba română reprezintă un suport de bază pentru dezvoltarea instrumentelor automate şi a aplicaţiilor lingvistice dedicate procesării informaţiei lingvistice specifice pentru gramatica, fonetica, şi lexicul limbii române. În acest articol prezentăm rezultatul unui studiu efectuat asupra morfologiei verbelor din limba română, precum şi etapele elaborării unui generator morfologic în baza rezultatelor obţinute în urma acestui studiu. Rezultatele obţinute de acest generator pot fi folosite pentru corectarea greşelilor de flexiune, sau chiar pentru prevenirea acestora, în cadrul traducerii automate în limba română, sau în alte aplicaţii lingvistice.

1. Introducere

Una din cercetările efectuate asupra morfologiei limbii române a fost crearea unui model de formalizare a morfologiei limbii române. În cadrul acestui proiect, cercetătorii şi-au propus ca scop realizarea unui analizor (corector şi generator) morfologic, şi realizarea unei baze de date ce conţine numai atributele specifice morfologiei. E de menţionat că atributele introduse au permis crearea unei prime versiuni de corector morfologic, cu posibilităţi de realizare a unui analizor morfologic complet, şi pe de altă parte a unui generator de paradigme concretizat într-un program de învăţare automată a conjugării verbelor. În cadrul acestui proiect s-au definit clasele flexionare, s-au determinat clasele flexionare la verb, şi clasele flexionare pentru nume, şi s-a determinat codificarea rădăcinii cuvintelor şi structurile de date (Peev et al., 1997).

O altă aplicaţie ce studiază structura morfologică a cuvintelor este aplicaţia Anmor. Această aplicaţie reprezintă un mediu de dezvoltare/actualizare pentru modelul morfologic paradigmatic al limbii române, iar preocupările sale esenţiale sunt asigurarea corectitudinii şi completitudinii datelor. Componentele principale sunt: un verificator de erori sintactice şi de inconsistenţe, un editor al dicţionarului, şi un asistent în procesul de îmbogăţire cu noi cuvinte a bazei de date. (Cosman, 2002). Studierea acestor aplicaţii lingvistice a şi stat la baza creării generatorului nostru morfologic.

2. Generatorul morfologic

Scopul acestui proiect a fost dezvoltarea unui generator morfologic bazat pe reguli pentru verbele din limba română. Am ales limba română fiindcă este una din limbile flexionare, care face ca obiectivul pe care ni l-am pus să fie mai dificil. Scopul principal a fost crearea regulilor pentru partea de vorbire flexionară verbul, prin atribuirea acestuia a caracteristicilor sale formale sub formă de atribut-valoare, pentru ca mai apoi aceste caracteristici să poată fi folosite pentru generatorul nostru morfologic. Inițial am început prin crearea regulilor pentru verbe, iar aceasta s-a dovedit a fi o sarcină foarte complexă, având în vedere că nu au fost de găsit reguli deja existente pentru această parte de vorbire.

Ideea de bază a acestui generator morfologic a fost crearea unor reguli de formare a verbelor, depinzând de atributele şi valorile acestora: conjugarea (conjugarea I, II, III, IV), modul (Indicativ, Conjunctiv, Condițional-Optativ, Infinitiv, Gerunziu, Participiu, Supin), timpul (prezent, imperfect, perfectul compus, perfectul simplu, mai mult ca perfectul, viitorul (viitorul simplu, viitorul anterior)), persoana (I, II, III), precum şi numărul (singular, plural), iar rezultatul (output-ul) acestui generator să fie salvat într-un fişier *.txt, pentru ca mai apoi acesta să poată fi folosit pentru alte aplicații lingvistice.

În Tabelul 1 sunt prezentate un set de reguli create folosind formalismul atribut-valoare, pentru atributul *Indicativ* şi valoarea *Mai-mult-ca-perfect*. Au fost create manual circa 1700 reguli, dintre care 280 de reguli pentru verbele de bază şi mai mult de 1400 pentru verbele neregulate (verbele auxiliare, verbele modale, etc.). Făcând o paralelă la numărul de reguli pentru alte părți de vorbire, e de menționat că acestea constituie aproximativ 200 de reguli pentru părțile de vorbire substantiv şi adjectiv.

Tabel 1. Crearea regulilor pentru atributul *Indicativ* şi valoarea Mai-mult-ca-perfect

Verbul la Infinitiv	*Conjugarea*	*Terminația conjugării*	*Persoana*	*Numărul*	*Verbul*	*Regulile create*
a învăța	**I**	**V-a**	1	Sg	învăț**asem**	**V+sem**
			2	Sg	învăț**aşeşi**	**V+şeşi**
			3	Sg	învăț**ase**	**V+se**
			1	Pl	învăț**aserăm**	**V+serăm**
			2	Pl	învăț**aserăți**	**V+serăți**
			3	Pl	învăț**aseră**	**V+seră**
a dormi	**IV**	**V-i, V-î**	1	Sg	dorm**isem**	**V+sem**
			2	Sg	dorm**iseşi**	**V+şeşi**
			3	Sg	dorm**ise**	**V+se**
			1	Pl	dorm**iserăm**	**V+serăm**
			2	Pl	dorm**iserăți**	**V+serăți**
			3	Pl	Dorm**iseră**	**V+seră**

3. Descrierea algoritmului de funcționare a generatorului morphologic

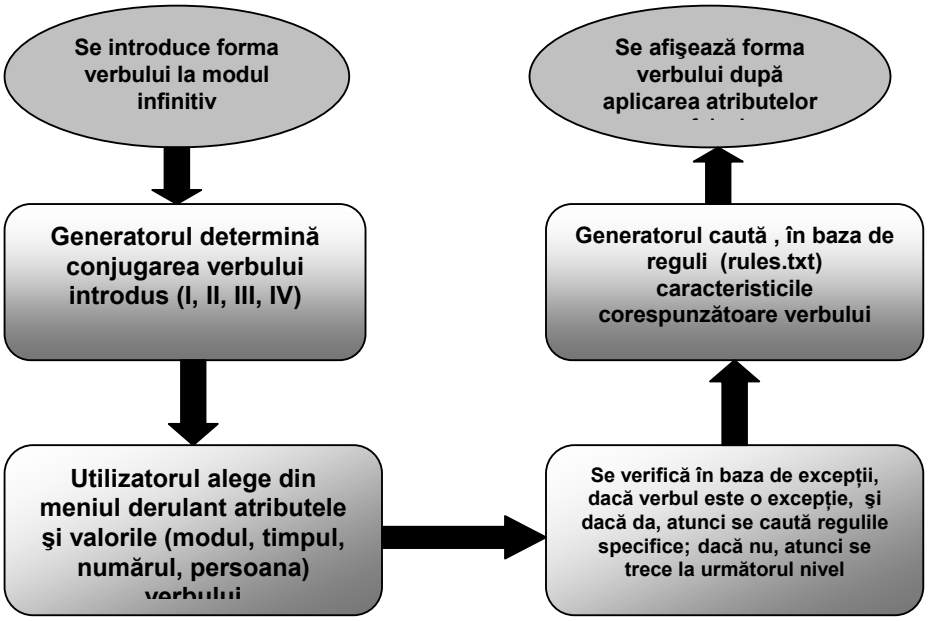


Figura 1.Algoritmul de funcționare a generatorului morfologic

În figura 1 este prezentat schematic algoritmul de funcționare a generatorului morfologic. Sistemul începe prin introducerea verbului la modul infinitiv. După care urmează determinarea conjugării verbului, pentru a putea trece la următoarea etapă.

În cazul programului nostru dispunem de mai multe fişiere care conțin informații morfologice, de tipul conjugările verbelor, regulile de formare a acestora, excepțiile (de exemplu, verbele auxiliare, etc.). După determinarea conjugării se cere alegerea caracteristicilor morfologice ale verbului, din meniul derulant al programului.

Apoi generatorul verifică în fişierul de excepții dacă verbul introdus este o excepție, şi dacă da, atunci se extrag şi se adaugă regulile proprii acestui verb-excepție, iar dacă nu, atunci se adaugă regulile ce depind de caracteristicile (atribute – valori) cerute de către utilizator.

În final se afişează rezultatul generării morfologice şi acest rezultat este salvat într-un fişier *.txt, pentru ca aceste rezultate să poată fi folosite în alte aplicații lingvistice.

Codul

```
II Indicativ Perfectul Simplu 2 Plural V - ea + uraţi|
II Indicativ Perfectul Simplu 3 Plural V - ea + ură  |
ex01II Indicativ Perfectul Simplu 1 Singular V - ea + ui |
ex01II Indicativ Perfectul Simplu 2 Singular V - ea + uşi|
```

În figura 2 este prezentată o formă în care sunt incluse atributele şi valorile verbelor, precum şi rezultatele generării acestora.

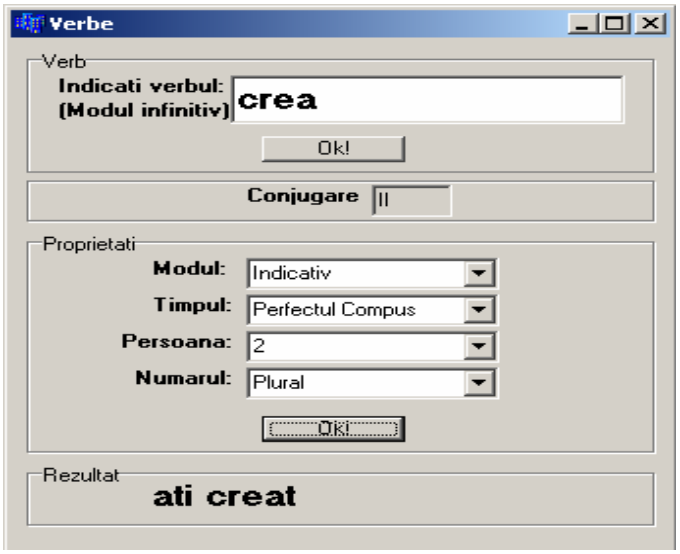


Figure 2. Afişări

4. Concluzii şi cercetări ulterioare

În acest proiect am creat prin mijloace lingvistice şi computaţionale un generator morfologic pentru o categorie flexionară a limbii române, şi anume, verbul. În cadrul acesteia au fost create reguli de generare a verbelor, bazându-ne pe regulile gramaticale de formare a verbului în limba română. Au fost create aproximativ 1700 de reguli de generare. Odată cu integrarea în Uniunea Europeană este necesară crearea mult mai multor aplicaţii lingvistice, aceasta reprezentând una din motivaţiile proiectului nostru. În viitor planificăm să adăugăm noi părţi de vorbire pentru acest generator morfologic, şi să rezolvăm problema diacriticelor.

Referinţe bibliografice

Dicţionarul ortografic, ortoepic şi morfologic al limbii române (2000). Editura Academiei Române, Bucureşti.

Popescu, Ş. (1997). *Gramatica practică a Limbii Române*, Editura Lider, Bucureşti.

Boatcă, M., Crihană, M. (1996). *Manual preparator de Gramatică a Limbii Române*, Editura Mondan, Moldova.

Peev, L., Bibolar, L., Jodal, E. (1997). Un Model De Formalizare A Morfologiei Limbii Române. In: Dan Tufiş, Poul Andersen (eds.). *Recent Advances in Romanian Language Technology*. ISBN 973–27–0626–0, Editura Academiei Române.

Cosman C. M. (2002). *Morfologia paradigmatică a limbii române. Mediu de dezvoltare / actualizare*. Lucrare de licenţă. Facultatea de Informatică, Universitatea Al.I. Cuza Iaşi.

PARSAREA PREDICATULUI (VERBAL / NOMINAL) ŞI A CLAUZEI (FINITE / NEFINITE) ÎN LIMBA ROMÂNĂ. APLICARE LA PARSAREA FDG

ALEX MORUZ[1,2], NECULAI CURTEANU[1], DIANA TRANDABĂŢ[1, 2], IUSTIN DORNESCU[1, 2], CECILIA BOLEA[1]

[1]*Institutul de Informatică Teoretică, Academia Română, Filiala Iaşi*

[2]*Facultatea de Informatică, Universitatea "Al.I.Cuza" Iaşi*

{curteanu, mmoruz, dtrandabat}@iit.tuiasi.ro

Rezumat

Lucrarea prezintă abordări şi experimente de parsare FDG pentru limba română. Sunt puse în evidenţă metode de parsare a structurilor globale (inter-clauzale) şi locale (intra-clauzale), cu accent pe parsarea grupului verbal.

1. Introducere

Articolul de faţă îşi propune să schiţeze folosirea de strategii şi programe de procesare a limbajului natural (LN), deja existente în cadrul colectivelor de cercetare din Iaşi şi Bucureşti, pentru proiectarea unui parser FDG (*Functional Dependency Grammar*) la nivel de frază (*sentence*), pentru limba română. Se pleacă de la strategia SCD (*Segmentare-Coeziune-Dependenţă*) de segmentare şi parsare a clauzei (şi sintagmelor subclauzale) (Curteanu *et al.*, 2005), intrarea în program fiind reprezentată de text (multiplu) adnotat la categorii morfologice şi marcheri SCD. Conform algoritmului de segmentare-parsare SCD, o fază esenţială în stabilirea corectă a clauzelor finite este *determinarea predicatelor finite*, verbale sau nominale.

Să precizăm de la început că în abordarea noastră folosim termenul de „*finit*" pentru toate *formele verbale* ce corespund unui *mod personal*, în timp ce termenul „*nefinit*" este atribuit formelor verbale ce corespund „modurilor" nepersonale, cunoscute şi ca *forme absolute* ale verbului (*infinitiv*, *gerunziu*, *participiu*, *supin*). În acest sens, clauza (sau propoziţia) al cărei predicat are ca nucleu semantic un verb *finit* (formă predicativ/verbală), o vom numi *clauză finită*. Astfel, *clauza finită* dintr-o frază este definită ca fiind întinderea de text aflată între doi marcheri SCD de nivel clauzal (sau de discurs), care conţine (exact) un predicat finit (cu nucleu semantic predicativ/personal). *Clauza nefinită* corespunde formelor verbale *nefinite*, predicativ/nepersonale ale categoriilor lexicale majore V (verb), N (substantiv), şi A (Adjectiv-Adverb), care posedă trăsătura de *predicaţionalitate* (*deverbalitate*). Parsarea predicatului finit (verbal sau nominal) revine la determinarea *grupului verbal finit* (Verbal Group, VG). VG mai este cunoscut în literatura de specialitate şi sub denumirea de *Complex Verbal* (Monachesi, 2005), (Barbu, 1999). Facem observaţia că determinarea VG (finit) de natură copulativă, (notat TASG – *Tense Auxiliary SubGroup* în (Curteanu & Trandabăţ, 2006)), este o etapă esenţială în parsarea predicatului nominal.

Lanţul operaţiunilor de parsare este următorul: la intrare, textul este adnotat morfologic şi apoi la marcheri SCD, rezultatul acestui proces fiind un text multiplu (*heavy*) adnotat. Pe adnotarea morfologică a acestui text se realizează parsarea VG (Curteanu *et al.*, 2006). Pentru determinarea clauzelor din componenţa unei fraze, cât şi a relaţiilor inter-clauzale dintre acestea, este folosit un program de segmentare-parsare la clauză bazat pe algoritmul de parsare SCD (Curteanu *et al.*, 2005). În urma execuţiei acestui program se obţine un arbore de dependenţă a clauzelor (*arbore clauzal*) din cadrul unei fraze date. Pentru realizarea arborilor de dependenţă intra-clauzali propunem două soluţii complementare: o *abordare deterministă*, bazată pe reguli, rezultată din algoritmul de parsare SCD intra-clauzală (Moruz, 2006), şi o *abordare statistică*, bazată pe algoritmi de învăţare automată. Scopul este de a obţine un parser mai performant prin *combinarea* unor algoritmi de parsare diferiţi (atât statistici cât şi bazaţi pe reguli). În acest moment se află în desfăşurare un proiect care are ca scop crearea unui corpus de fraze adnotate la dependenţe funcţionale pentru limba română.

2. Etape şi Instrumente

2.1. Parsarea VG

Primul pas în adnotarea FDG a unei fraze date este determinarea predicatelor din textul de intrare. Această operaţie este necesară nu numai pentru segmentarea frazei în clauze, ci şi pentru găsirea proprietăţilor *grupurilor verbale* (VGs) în vederea determinării corecte a dependenţelor intra-clauzale (de exemplu, în diateza pasivă subiectul gramatical devine *obiect semantic*, iar complementul direct devine *agent*). În urma parsării predicatelor, datele obţinute sunt următoarele:

(a) *Nucleul semantic* al fiecărui VG. Această componentă a predicatului este importantă deoarece face diferenţa dintre *predicatele nominale* şi *predicatele verbale*. În cazul în care nucleul semantic este de tip *predicaţional* (*deverbal*), VG reprezintă un *predicat verbal*; în cazul în care avem un nucleu verbal de tip *copulativ*, VG reprezintă nucleul sintactic al unui *predicat nominal*. În acest al doilea caz, *nucleul semantic* al predicatului nu mai este verbul copulativ din VG ci argumentul acestuia, *numele predicativ*. În "*Ion pleacă acasă.*", nucleul semantic este verbul "*pleacă*", în timp ce predicatul propoziţiei "*Ion a fost student.*" este nominal, având ca nucleu semantic substantivul nepredicaţional "*student*", pe când nucleul VG este copulativul "*fost*".

(b) *Diateza formal-sintactică* (*de suprafaţă*) a predicatului. Aceasta este de fapt diateza gramaticală clasică. În multe cazuri, diateza sintactică şi cea semantică nu coincid, astfel argumentele *directe* (ce corespund valenţei) ale verbului predicaţional nu sunt aranjate corect în lista SUBCAT a argumentelor. De exemplu, diateza sintactică a predicatului în propoziţia "*Maşina se spală.*" este cea *reflexivă*, dar *diateza* sa *semantică* este cea *pasivă*. Procesul de determinare a VGs împreună cu proprietăţile lor este descris în (Curteanu *et al.*, 2006). În urma găsirii nucleului semantic şi a diatezei sintactice, dacă nucleul este predicaţional (echivalând cu situaţia că nu este verb copulativ), se coboară în lexicon prin proiecţia *FX-bar inversă* în vederea determinării *diatezei semantice* a VG, după care se revine de la nivelul lexiconului la nivelul textului de suprafaţă cu noua diateză şi cu restricţiile de *linking* asociate acesteia prin

mecanismul *proiecției FX-bar directe*. Procesul de proiecție FX-bar inversă a VG către lexicon, prin care se determină nucleul semantic al VG, şi proiecția FX-bar directă a nucleului semantic al VG către VG, şi al VG către clauza finită, este descris detaliat în (Curteanu, Trandabăț, 2006). Câteva exemple de transformare a *diatezei formal-sintactice* în *diateză lexical-semantică* (de observat că *diateza sintactică* pentru toate aceste exemple este cea reflexivă):
(e1) *Ion şi Maria* ***se ştiu*** *de mici copii*. (diateză semantică = *reciprocă*);
(e2) ***Se ştie*** *vinovat de moartea mamei sale*. (diateză semantică = *reflexivă*);
(e3) ***Se ştie*** *că pisicile fugăresc şoriceii*. (diateză semantică = *impersonală*).

2.2. Parsarea Clauzelor

Parsarea clauzelor (unei fraze) este realizată prin intermediul unui program de separare a unităților clauzale şi subclauzale bazat pe strategia de parsare SCD (*Segmentare-Coeziune-Dependență*) (Curteanu *et al.*, 2005). Separarea unităților clauzale este realizată pe baza claselor de marcheri SCD, a ierarhiei de tip graf a acestor clase, şi a proprietății de *predicaționalitate* (*deverbalitate*) a categoriilor lexicale majore N (Noun), V (Verb), şi A (Adjectiv-Adverb). De aici rezultă necesitatea determinării marcherilor SCD şi a predicatelor din frază ca o preprocesare în vederea aplicării algoritmului SCD. Pentru obținerea *arborelui de dependență clauzal* este necesară, pe lângă parsarea clauzelor, şi determinarea relațiilor de dependență inter-clauzale. Aceasta se face cu ajutorul marcherilor de tip SCD, reprezentați într-o bază de date în care este specificat şi tipul de relație (inter-clauzală) pe care o introduc.

2.3. Relații de Dependență pentru Limba Română

În Tabelul 1 sunt exemplificate o parte din relațiile de dependență FDG intra-clauzală ce au fost determinate pentru limba română. Cuvintele scrise cu litere îngroşate reprezintă *fii* în relațiile descrise, iar categoria „X” - orice categorie morfologică.

Tabel 1: Exemple de relații de dependență funcțională FDG pentru limba română

NUCLEU	FIU	Cuv. Urm.	RELAȚIE	ABREV.	EXEMPLU
Substantiv	Prepoziție	Nominal	Atribut substantival	a.subst	*Praf* **de** *puşcă*
Substantiv	Articol	X	Determinant	det	**Un** *om*
Substantiv	Verb	X	Atribut verbal	a.verb	*Omul care* **merge**
Substantiv	Adjectival	X	Atribut adjectival	a.adj	*Om* **tânăr**
Adjectiv	Adverb	X	Comparativ	comp.	**Mai** *mare*
Verb	Nominal	X	Subiect	Sbj	**Ion** *merge*
Verb	Nominal	X	Complement direct	c.d.	**Îl** *văd*
Verb	Verb aux	X	Auxiliar	aux	**Am** *mers*
Verb	Nominal, adj.	X	Nume predicativ	n.pred	*Sunt* **tânăr**
Verb	Negație	X	Negație	neg	**Nu** *stau*
Adverb	Adverb	X	Comparativ	comp	**Mai** *repede*
Prepoziție	X	X	Rel. prepozițională	prep.	*De* **vorbă**
Coordonator	X	X	Rel. de coordonare	coord	**Mare** *şi* **tare**

2.4. Parsarea FDG Bazată pe Reguli

Pentru rezolvarea problemei dependenţelor funcţionale FDG au fost folosite două *abordări complementare*, una bazată *pe reguli* şi una *statistică*, bazată *pe învăţare automată*. Programul de parsare FDG (*Functional Dependency Grammar*) porneşte de la o formalizare a gramaticilor de dependenţă descrisă de (Järvinen, Tapanainen, 1997). Algoritmului foloseşte reguli (de fapt, *expresii regulate*) în vederea stabilirii de dependenţe; pentru rezolvarea problemei dependenţelor la distanţă s-a folosit paradigma *Island Parsing*. Island Parsing este o strategie de parsare multidirecţională, utilizată atât în cadrul prelucrării limbajului natural cât şi în alte domenii în care robusteţea este importantă sau resursele de procesare sunt limitate, ataşată gramaticilor independente de context. Este o strategie bidirecţională, în sensul că elemente de parsare incomplete, care corespund părţii drepte ale unei reguli de producţie independentă de context, pot fi extinse în ambele direcţii.

Regulile de parsare au fost determinate experimental, în urma studierii unei serii ample de exemple pentru limba română (Moruz; 2006); din acest motiv, expresiile regulate astfel determinate nu pot fi utilizate pentru parsarea la dependenţe funcţionale a textelor în alte limbi. În ***parsarea bazată pe reguli***, paşii urmaţi pentru a parsa la dependenţe funcţionale FDG un text sunt următorii: **(1)** *segmentarea textului* primit la intrare în clauze şi unităţi lexicale (acest pas este realizat pe baza strategiei de segmentare-parsare SCD a textului); **(2)** *delimitarea constituenţilor* (extragerea de grupuri verbale, nominale, adjectivale, etc.); **(3)** *determinarea elementelor relaţionale* de nivel inter-clauzal şi de discurs, şi ataşarea lor la elementele subordonate. Delimitarea constituenţilor şi determinarea elementelor relaţionale împreună cu legăturile lor funcţionale se realizează cu ajutorul expresiilor regulate. Expresiile regulate în sine nu sunt suficient de particulare pentru a putea determina o structură arborescentă neambiguă. În scopul scăderii ambiguităţii au fost create reguli de procesare contextuală (cum ar fi *acordul*), ce cresc eficienţa determinării structurilor (*ruleNounAdj(Tree t*1*, Tree t*2*)* – dacă rădăcina arborelui *t*1 este *substantiv* şi rădăcina arborelui *t*2 este *adjectivală*, iar cele două rădăcini sunt în acord morfo-sintactic, atunci *t*2 devine subarbore pentru *t*1).

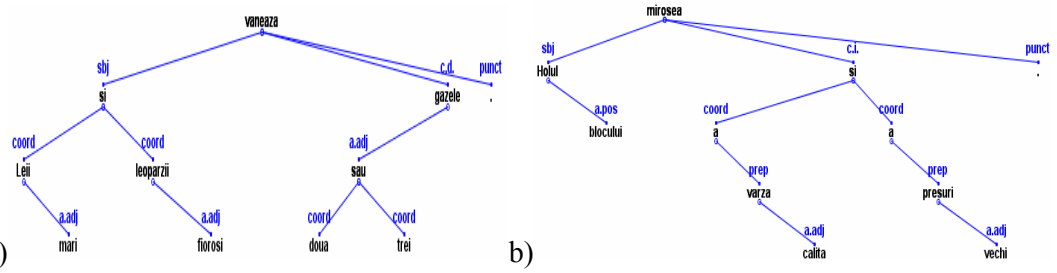
2.5. Parsarea Statistică

Raţiunea principală în ***abordarea statistică*** este că putem folosi o serie de algoritmi de parsare deja dezvoltaţi pentru alte limbi. Proiectarea unei gramatici formale pentru limba română presupune un efort foarte mare în crearea unui corpus adnotat pe baza căruia să se facă antrenarea modelului, ţinând cont de particularităţile fiecărui formalism. Întrucât pentru limba română nu avem la dispoziţie un corpus de dependenţe funcţionale FDG, scopul este de a folosi parsere statistice existente pentru dezvoltarea unui astfel de corpus. Propunem o dezvoltare *iterativ-incrementală* a unui corpus, după cum urmează: la început adnotăm manual un număr relativ mic de fraze; pe baza acestui micro-corpus, antrenăm trei parsere statistice şi rulăm cele trei modele obţinute pentru adnotarea automată a unui set nou de fraze. Acestea vor fi *doar corectate* de adnotatori umani iar corpusul astfel obţinut va fi folosit pentru reantrenarea parserelor; procesul e reluat până când întreg corpusul este adnotat. Clasa parserelor *deplasare-reducere*

foloseşte un algoritm simplu: parcurge textul cu o fereastră conţinând un număr fix de cuvinte. Alte metode de adnotare FDG statistică sunt *arborii parţiali de cost maxim* şi *meta-parserele*. *Arborele parţial de cost maxim* porneşte prin a construi un graf orientat complet care are ca noduri cuvintele din frază, iar ca arce numărul de apariţii în corpus al unei muchii similare. Algoritmul caută un arbore parţial de cost maxim, iar acest arbore reprezintă cea mai probabilă parsare FDG a frazei. *Meta-parserele* sunt modele de combinare a mai multor *parsări*, şi *nu* a mai multor parsere propriu-zise. Deşi, intuitiv, o metodă ierarhică de a combina parsările în mod diferenţiat, în funcţie de diverse criterii, ar trebui sa ducă la o creştere a preciziei, astfel de metode au dezavantajul costului computaţional destul de ridicat.

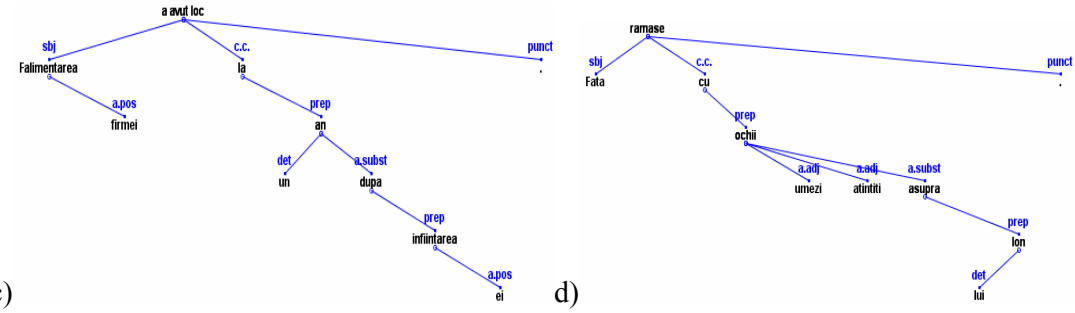
3. *Performanţe*

Datorită absenţei unui corpus adnotat la arbori de dependenţă funcţională pentru limba română, programul nu a putut fi încă testat suficient. Din acest motiv nu se pot oferi estimări realiste în legătură cu precizia ataşată procesului de adnotare la dependenţe funcţionale. În urma verificării manuale a unei părţi din arborii obţinuţi, rezultatele conduc către o precizie destul de mare pentru frazele formate dintr-o singură propoziţie.



a) b)

Figura 1: Exemple de parsare FDG: a) „*Leii mari şi leoparzii fioroşi vânează două sau trei gazele.*”; b) ”*Holul blocului mirosea a varză călită şi a preşuri vechi.*”;



c) d)

Figura 2: c) ”*Falimentarea firmei a avut loc la un an după înfiinţarea ei.*”; d) ”*Fata rămase cu ochii umezi aţintiţi asupra lui Ion.*”;

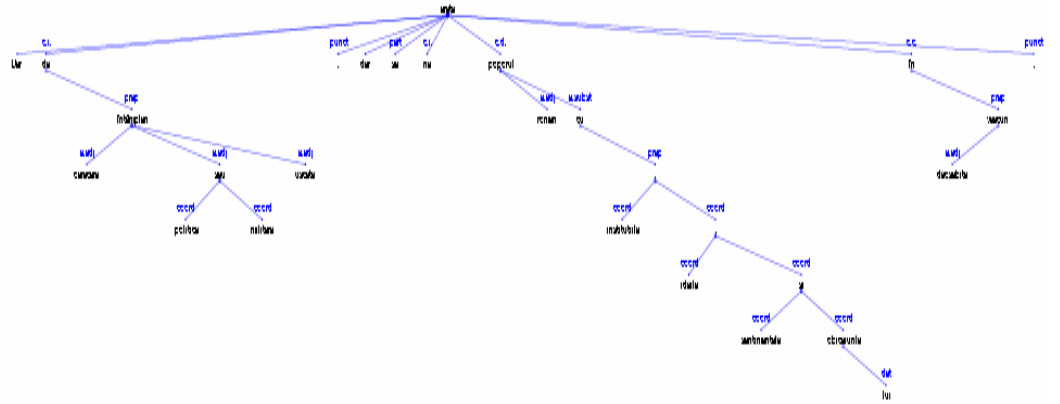


Figura 3: e) (Exemplu de parsare eronată datorată separării incorecte a frazei în clauze) „*Dar de oarecare întâmplări politice sau militare uscate, dar să ne arate poporul român cu instituţiile, ideile, sentimentele şi obiceiurile lui în deosebite veacuri.*”

Referinţe bibliografice

Barbu, A.M. (1999). *The Verbal Complex*. *Studii şi Cercetări Lingvistice*, L, no.1, Bucureşti, p. 39-84.

Curteanu, N., Zlavog, E., Bolea, C. (2005). Sentence-Based and Discourse Segmentation / Parsing with SCD Linguistic Strategy. In *'Intelligent Systems' Conference*, H.-N. Teodorescu *et al.* (Eds), Performantica Press, Iaşi, pp. 153-168.

Curteanu, N. (2006). Local and Global Parsing with Functional X-bar Theory and SCD Linguistic Strategy (I + II.). *Computer Science Journal of Moldova*, Chişinău, vol. 14, nr. 1-2 (40-41).

Curteanu, N., Moruz, M., Trandabăţ, D., Bolea, C., Dornescu, I. (2006). The Structure and Parsing of Romanian Verbal Group and Predicate. *Advances in Intelligent Systems and Technologies ECIT2006 – 4th European Conference on Intelligent Systems and Technologies*, Iaşi, Romania, Sept. 2006, pp. 93-105.

Curteanu, N., Trandabăţ, D. (2006) Functional (F)X-bar Projections for Local and Global Text Structures. The Anatomy of Predication. *Revue Roumaine de Linguistique*, Bucharest.

Järvinen, T., Tapanainen, P. (1997). *A Dependency Parser for English*. Technical reports, No. TR-1, Department of General Linguistics, University of Helsinky.

Monachesi, P. (2005). *The Verbal Complex in Romance. A Case Study in Grammatical Interfaces*. Oxford University Press, Oxford Studies in Theoretical Linguistics.

Moruz, M. A. (2006). *Dezvoltarea unui adnotator FDG pentru limba română*, Lucrare de licenţă, Univ. „Al. I. Cuza”, Facultatea de Informatică, Iaşi, 2006.

PRELUCRAREA RESURSELOR ROMÂNEŞTI ÎN CADRUL PROIECTULUI LT4EL

IONUŢ PISTOL[1], ADRIAN IFTENE[1], DIANA TRANDABĂŢ[1, 2], DAN CRISTEA[1, 2], CORINA FORĂSCU[1, 3]

[1] Facultatea de Informatică, Universitatea "Al. I. Cuza", Iaşi
[2] Institutul de Informatică Teoretică, Academia Română, Filiala Iaşi
[3] Institutul de Cercetări pentru Inteligenţă Artificială, Academia Română, Bucureşti

{ipistol, adiftene, dtrandabat, dcristea, corinfor}@info.uaic.ro

Rezumat

Proiectul LT4eL are ca scop realizarea unei tehnologii multilingve, utilizate în cadrul unui sistem de eLearning, care să faciliteze operaţiile de creare a obiectelor de învăţare de către profesori şi de regăsire a lor de către studenţi, inclusiv prin criterii de natură semantică. Până în momentul de faţă, în cadrul proiectului a fost creat un corpus semnificativ (peste 600.000 cuvinte) de texte româneşti în diferite formate de adnotare, plecând de la textul original şi mergând până la un format XML ce pune în evidenţă informaţii morfo-sintactice, cuvinte cheie şi definiţii. În lucrare sunt prezentate etapele de prelucrare ale documentelor corpusului şi modulele de prelucrare utilizate.

1. Introducere

Colectarea şi prelucrarea unui corpus românesc semnificativ ca dimensiune şi nivel de adnotare este una din cele mai importante etape în aducerea tehnologiilor şi resurselor lingvistice pentru limba română la un nivel comparativ cu limbile vest-europene. În cadrul proiectelor iniţiate în aceasta direcţie poate fi inclus şi LT4eL[1] (Monachesi et. al., 2006), proiect susţinut de Comunitatea Europeană prin departamentul *Information Society and Media Directorate, Learning and Cultural Heritage*. În acest proiect sunt implicate universităţi reprezentând 9 limbi europene (engleza, germana, olandeza, portugheza, poloneza, ceha, bulgara, malteza şi româna). Limba română este reprezentată în proiect prin autori, membri ai grupului de cercetare în Tehnologiile Limbajului Uman[2] din cadrul Facultăţii de Informatică). Scopul principal al proiectului este dezvoltarea de instrumente şi resurse lingvistice concepute pentru folosirea în învăţarea asistată de calculator (eLearning).

O primă etapă a LT4eL, pentru care responsabilă a fost echipa română, a constat în conceperea şi realizarea unui mediu de adăugare, acces şi actualizare a tuturor corpusurilor, resurselor lingvistice şi sistemelor de prelucrare ce urmau a fi dezvoltate în cadrul proiectului. În acest scop s-a creat un portal[3] prin intermediul căruia au fost colectate de la începutul proiectului (decembrie 2005) resurse, totalizând aproape 9

[1] Eng: Language Technologies for E-Learning, pagina principală la http://www.lt4el.eu.
[2] http://consilr.info.uaic.ro/research/
[3] http://consilr.info.uaic.ro/uploads_lt4el/

milioane de cuvinte, şi 30 aplicaţii dezvoltate integral în cadrul proiectului sau adaptate necesităţilor proiectului.

În secţiunea a doua a acestei lucrări va fi prezentat procesul general de prelucrare a corpusului românesc, pentru ca în secţiunile 3 şi 4 să fie descrise mai pe larg două dintre etape, semnificative prin noutatea modulelor de procesare dezvoltate. Secţiunea a 5-a conţine o descriere a etapelor următoare ale proiectului, precum şi a resurselor şi instrumentelor ce urmează să fie create pentru limba română.

2. Etapele de prelucrare a corpusului românesc

Etapa iniţială de formare a corpusului românesc a constat în colectarea de documente din 14 domenii convenite în cadrul proiectului, domenii ce ţin de predarea informaticii, pedagogie şi organizarea studiului universitar. Un prim nivel de adnotare propus a fost XML[4] cu marcaje de formatare a documentului, format definit de un DTD[5] general. Numai marcajele de formatare care pot facilita prelucrări ulterioare, cum ar fi marcarea cuvintelor cheie, au fost păstrate. După transformarea întregului corpus în format XML, etapa a doua a fost cea de adnotare lingvistică în care s-a realizat: segmentarea în unităţi lexicale, propoziţii, fraze şi paragrafe, marcarea informaţiilor morfo-sintactice, a formelor de bază (leme) ale cuvintelor flexionate, marcarea grupurilor nominale.

A treia etapă a constat în marcarea manuală a cuvintelor cheie şi a definiţiilor din corpus, relevante pentru domeniul general al proiectului. Drept cuvinte cheie au fost marcate cuvintele şi expresiile considerate de adnotator ca fiind relevante în contextul conţinutului şi al scopului lucrării respective. Toate aceste cuvinte cheie au fost traduse din celelalte 8 limbi în engleză, centralizate şi sortate, obţinându-se astfel colecţia lexicală, ca prim pas în construirea unei ontologii generale pentru domeniile de interes ale LT4eL. Formatul convenit pentru realizarea ontologiei a fost DOLCE[6]. Într-o etapă ulterioară, forma finală a acestei ontologii şi lexiconul aferent vor fi mapate integral la Princeton WordNet[7], lucru ce va permite integrarea ei cu alte ontologii de domenii, cum ar fi SUMO, dar şi utilizarea ei în cadrul altor proiecte.

În prezent, corpusul proiectului conţine peste 600.000 cuvinte, în variantele de adnotare descrise mai sus, până la marcajele de cuvinte cheie şi definiţii.

3. Conversia corpusului la formatul XML

Documentele primare ale proiectului au trebuit aduse de la formatele iniţiale la un format unitar XML în codificare UTF-8 (ce permite notarea diacriticelor în toate limbile proiectului), înainte de abordarea etapelor adnotării lingvistice şi a celei semantice. Transformarea întregului corpus LT4eL la un format XML standard, definit de o specificare DTD, s-a făcut, în parte, utilizând o serie de convertoare disponibile, în parte, implementând trei noi convertoare. Formatul html s-a ales ca nivel intermediar în

[4] Extensible Markup Language (http://www.w3.org/XML/)
[5] Document Type Definition
[6] http://www.loa-cnr.it/DOLCE.html
[7] http://wordnet.princeton.edu/

transformarea între formatele inițiale doc şi pdf şi formatul final XML. Motivul a fost că html păstrează în bună măsură indicațiile de formatare ale documentelor inițiale.

Prezentăm mai jos, pe scurt, doar câteva din problemele întâlnite la conversia corpusului din formatul html la formatul XML ce respectă restricțiile convenite în cadrul LT4eL:

- Unele cuvinte consecutive apar alipite: cauza trebuie pusă fie pe seama conversiei la html (text pe două coloane, probleme date de justify, cuvinte despărțite de newline), fie pe seama convertorului la xml. Cuvintele ce nu sunt despărțite textual, ci doar prin interpretarea etichetelor html, sunt principalele cauza ale apariției acestei erori. Dacă textul are numeroase elemente de formatare (fonturi multiple, culori, siluri), aceste probleme sunt frecvente.
- Unele cuvinte apar cu spații în interior: în general, aceasta este o consecință a conversiei de la pdf la html şi apare în textele pe două coloane, cu *justify* sau consecință a despărțirii în silabe. Rezolvarea acestei probleme necesită resurse lingvistice pentru limba respectivă, pentru evitarea creării de erori opuse, adică apariția unor cuvinte alipite.
- Unele caractere au coduri greşite în xml: fie din cauza html-ului de origine unde caracterul este notat printr-un alt cod decât UTF-8, fie pentru că un singur caracter are uneori mai multe codificări (de obicei UTF-8 şi UTF-16), sau pentru că acelaşi caracter are coduri diferite în seturi diferite de caractere UTF-8 (cum ar fi Latin-1 şi Arabic). O soluție optimă, cum ar fi crearea unui instrument automat de conversie din orice codare la UTF-8, pare a fi imposibil de realizat (de exemplu UTF-16 *unpaired*[8] nu poate fi convertit la UTF-8).
- Uneori apar spații multiple în loc de unul singur. Corectarea prin înlocuirea spațiilor multiple cu unul singur poate uneori strica formatarea documentului.
- Uneori convertorul păstrează în XML atributele unor etichete eliminate: problema apare în special la etichetele ce apar doar deschise şi la atributele ale căror valoare apare fără ghilimele. Soluția are în vedere, în principal, rezolvarea acestor două situații.

Pe lângă problemele menționate anterior mai apar câteva, semnificative, ce sunt datorate procesului de conversie. În primul rând, conversia corectă a textelor puternic formatate (tabele, formule, imagini grupate) este o problemă uneori extrem de dificil de rezolvat, lucru afirmat şi în documentațiile convertoarelor de firmă (*MS Word*, *Adobe*, *pdf2html*). O soluție viabilă pentru acest tip de documente, a căror formatare poate fi uneori relevantă pentru adnotarea lingvistică, ar impune o prelucrare adițională, în vederea transformării obiectelor problematice într-un format mai uşor de prelucrat. Acest lucru poate implica şi un efort manual din partea adnotatorului, posibil facilitat de dezvoltarea unui mediu vizual de editare. Există de asemenea documente doc şi pdf care provin din conversia unor documente scanate. Acestea vor rămâne practic imposibil de convertit la XML cu păstrarea integrală a conținutului, chiar şi după utilizarea unui software de tip OCR (Optical Character Recognizer) performant. O altă

[8] http://unicode.org/unicode/faq/

problemă ce implică o prelucrarea adiţională este cea a documentelor ce conţin, în original, secvenţe de adnotare html/XML a textului (ca cele din manuale, de exemplu) şi care se pot confunda cu metadatele XML ale formatului final.

4. Adnotarea lingvistică

Adnotarea lingvistică (îmbogăţirea corpusului cu informaţii sintactice şi morfologice) urmează etapei de transformare a corpusului în format XML. Într-o primă etapă s-a făcut o evaluare a instrumentelor de prelucrare disponibile în vederea preluării în proiect a unora dintre ele, apoi s-a luat decizia implementării unora noi. În final, pentru adnotarea lingvistică au fost folosite:

- *tokenizator* (marcator de unităţi lexicale de bază), dezvoltat de echipa UAIC;
- *POS-tagger* (adnotator morfo-sintactic), adaptat după o implementare ICIA[9];
- *lemmatizer* (marcator de rădăcini morfologice neflexionate), realizat la ICIA;
- *NP-chunker* (marcator de grupuri nominale), dezvoltat de echipa UAIC, utilizând un corpus adnotat manual pentru a genera un set de reguli, ce au fost apoi revizuite în parte înainte de a fi utilizate de marcatorul de grupuri nominale.

Cele patru module de procesare de mai sus obţin rezultate foarte bune pentru limba română. *Tokenizator*-ul şi *POS-tagger*-ul obţin scoruri *F-measure*[10] de aproximativ 98% (Tufiş şi Dragomirescu, 2004), *Lemmatizer*-ul obţine un scor *F-measure* de aproximativ 95%, iar *NP-chunker*-ul aproximativ 75%. Evaluările pentru instrumentele ICIA au fost preluate din documentaţia aplicaţiilor, iar pentru cele dezvoltate la UAIC au fost calculate automat utilizând un corpus standard adnotat manual. Acest corpus (ca şi o primă variantă a modulelor respective) a fost dezvoltat în cadrul unui proiect anterior, ce a presupus dezvoltarea unui sistem de întrebare-răspuns (Iftene et al, 2006) în context bilingv română-engleză. Corpusul utilizat ca standard în evaluare este romanul ‘1984’ de G.Orwell în varianta tradusă în limba română şi cuprinde peste 100.000 unităţi lexicale.

Pentru combinarea şi simplificarea adnotărilor şi aducerea la formatul proiectului a fost folosită o aplicaţie dezvoltată de echipa UAIC, ce va sta la baza unui viitor sistem de adnotare automată ce va combina module de prelucrare în scopul obţinerii automate a unor formate complexe.

5. Etape viitoare

Etapele viitoare ale proiectului LT4eL constau, în primul rând, în implementarea şi antrenarea unui sistem de recunoaştere automată a cuvintelor cheie şi a definiţiilor. Cuvintele cheie şi definiţiile din texte vor fi principalele elemente de legătură între obiectele de învăţare (eng: *Learning Objects*) şi vor permite integrarea şi combinarea acestor resurse în scopul creării unui mediu dinamic de generare de materiale pentru învăţarea asistată de calculator. Pentru recunoaşterea cuvintelor cheie va fi folosită o

[9] Institutul de Cercetări în Inteligenţă Artificială din cadrul Academiei Române – Bucureşti.

[10] Calculat ca 2*P*R/(P+R), unde P (precizia) = numărul de obiecte corect identificate de program raportat la numărul de obiecte identificate de program şi R (recall) = numărul de obiecte corect identificate de program raportat la numărul de obiecte existente.

adaptare a algoritmului *tf-idf* (Salton şi Buckely, 1988), cu antrenare pe corpusul adnotat manual la cuvinte cheie existent (aproape 25 mii de cuvinte cheie marcate manual, din care 7772 în corpusul românesc). Pentru recunoaşterea automată a definiţiilor s-a elaborat o primă variantă a unei gramatici ce utilizează informaţiile morfo-sintactice şi de formatare existente în corpus pentru a identifica şi marca graniţele definiţiilor din text.

Îmbogăţirea corpusului românesc LT4eL este de asemenea una din direcţiile de lucru, scopul echipei româneşti fiind acela de a utiliza tehnologiile şi formalismele dezvoltate în cadrul LT4eL pentru a colecta şi prelucra un corpus mult mai mare decât cel necesar pentru proiect, ce ar putea fi folosit mai târziu în cadrul altor proiecte de cercetare în care este implicată limba română, cum ar fi dezvoltarea unui sistem de traducere automată, a unui rezumator, a unui parser de discurs etc.

Transformarea ierarhiei de formate într-un mediu de prelucrare automată a documentelor este un alt scop avut în vedere de echipa românească, în afara cerinţelor stricte ale proiectului. Proiectarea unui astfel de mediu, cu posibilitatea îmbogăţirii sale ulterioare cu noi module de prelucrare şi resurse lingvistice poate facilita dezvoltarea unor sisteme complexe de procesare lingvistică, cu aplicaţii ce-şi pot găsi utilitatea şi în prelucrările dedicate altor limbi decât limbii române. Formalismele teoretice pentru dezvoltarea acestui mediu sunt în curs de îmbunătăţire, după ce a fost elaborată o primă propunere (Cristea et. al, 2006). Se speră ca până la terminarea proiectului LT4eL să fie finalizată o variantă a acestui mediu care să acopere cel puţin procesările necesare LT4eL.

Referinţe bibliografice

Iftene A., Pistol I., Trandabăţ D., Puşcaşu G., Forăscu C., Cristea D. (2006) Sisteme de Întrebare Răspuns pentru Limba Română. În acest volum.

Monachesi P., Lemnitzer L., Simov K. (2006). Language Technology for eLearning to appear in *Proceedings of EC-TEL 2006*, Springer LNCS (http://www.ectel06.org/index.html).

Tufiş D., Dragomirescu L. (2004). Tiered Tagging Revisited. In *Proceedings of the 4th LREC Conference*, Lisabona, 2004, pp. 39-42.

Salton G., Buckley C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management* 24(5): 513–523.

Cristea D., Forăscu C., Pistol I. (2006). Requirements-Driven Automatic Configuration of Natural Language Applications, *Proceedings of the 3rd International Workshop on Natural Language Understanding and Cognitive Science NLUCS-2006*, ICEIS 2005, 23 - 27 May 2006, Paphos, Cyprus.

SISTEM DE INSTRUIRE ASISTATĂ DE CALCULATOR PENTRU MORFOLOGIA LIMBII ROMÂNE

ELENA BOIAN, CONSTANTIN CIUBOTARU, SVETLANA COJOCARU, GALINA MAGARIU, TATIANA VERLAN, IURI ROGOJIN

Institutul de Matematică şi Informatică, Academia de Ştiinţe a Moldovei

{lena, chebotar, sveta, gmagariu, tverlan, rogozhin}@math.md

Rezumat

În lucrare[1] este descrisă concepţia de elaborare a sistemului de instruire asistată de calculator pentru limba română. Pentru implementarea acestei concepţii s-au utilizat mijloacele sistemului Claroline, o platformă Open Source. Concepţia propusă presupune studierea morfologiei limbii române pentru trei categorii de utilizatori cu diferite niveluri de pregătire.

1. Introducere

Dezvoltarea continuă a tehnologiilor informaţionale în sfera educaţiei se manifestă printr-o vastă orientare spre utilizarea produselor program deschise, tendinţă manifestată şi în domeniul sistemelor de instruire asistată de calculator (SIAC). În lucrare sunt expuse principiile şi particularităţile de proiectare şi realizare a unui SIAC pentru morfologia limbii române, sistem bazat pe o arhitectură ierarhică orientată pe categorii de utilizatori şi integrată cu resursele lingvistice reutilizabile plasate pe Internet.

Sistemul de instruire presupune nu numai cursul materialului pentru studiere (inclusiv şi exerciţii), ci şi mijloace de prezentare şi expunere a materialului de către profesor, modalitatea de organizare a comunicării între student şi profesor, tehnicile de însuşire a materialului de către student.

Elaboratorii SIAC, fiind în postură de profesor, selectează materialul teoretic şi exerciţiile necesare, ţinând cont de metodica şi principiile de prezentare a obiectului şi de nivelul de pregătire a studentului; în funcţie de student, elaborează o modalitate optimă pentru însuşirea materialului propus. Se propune o secvenţă de activităţi – alternarea materialului teoretic şi a exerciţiilor de autotestare, a testelor controlate prin intermediul sistemului de instruire, de asemenea, a testelor gestionate de profesor în procesul de comunicare cu studentul (în special, pentru exerciţiile creative).

Aşadar, un sistem de instruire nu este numai o simplă colecţie şi nu numai o colecţie de material teoretic ce ţine de disciplina pentru instruire, cum s-ar părea la prima vedere. Acesta este o noţiune mai vastă şi mai profundă şi, în cazul nostru, reprezintă un sistem consistent, multilateral şi valoros.

SIAC propus presupune elaborarea:

[1] Lucrarea este efectuată în cadrul proiectului RoLTech INTAS Ref. Nr. 05-104-7633

- materialului teoretic propriu zis,
- exerciţiilor pentru formarea deprinderilor de utilizare a materialului teoretic,
- metodicii de predare a materialului de către profesor,
- modalităţii de comunicare între student şi profesor,
- metodelor de asimilare a materialului de către student.

2. Concepţia de construire a cursurilor pentru studierea limbii române

În baza discuţiilor cu persoanele nevorbitoare de limba română, care au frecventat cursurile de studiere a limbii române, a cercetării literaturii de specialitate existente în domeniul gramaticii (Acsan et al., 2004), (Bărbuţă, 1998), (Bărbuţă et al., 2000, 2003), (Bertea, 1996), (Cazacu & Vrabie, 2006), (Cruceru & Teodorescu, 2005), (Dumeniuc & Matcaş, 1989), (Gramatica limbii române, 1963), (Irimia, 1997), (Irimia, 2004), (Limba română contemporană, 1999), (Molan et al., 1995, 1996), (Nastasenco, 1996), (Pop, 2000), (Zaiuncikovski & Repina, 1989) şi a metodicii de predare a limbii române (Consiliul Naţional pentru Curriculum, 2002), (Ionescu, 2001), (Ivănuş, 1997), (Nuţă, 2000), a fost elaborată o concepţie de construire a cursurilor asistate de studiere a limbii române.

Pentru implementarea acestei concepţii s-au utilizat mijloacele sistemului Claroline[2].

Concepţia propusă presupune studierea morfologiei limbii române pentru trei categorii de utilizatori:

- **prima categorie** – persoanele care nu cunosc gramatica şi au un vocabular sărac;
- **categoria a doua** – persoanele care înţeleg limba vorbită, au cunoştinţe gramaticale nesistematizate;
- **categoria a treia** – persoanele care cunosc limba română, dar vor să-şi aprofundeze cunoştinţele referitor la gramatică, particularităţi lingvistice, derivare a cuvintelor etc.

2.1. Prima categorie – nivelul 1

Scopurile lecţiilor de nivelul I:

- dezvoltarea limbii vorbite, începând cu fraze elementare şi expresii uzuale;
- îmbogăţirea treptată a vocabularului şi complicarea frazelor;
- studierea paralelă a categoriilor gramaticale la nivel elementar. Aceste categorii sunt predate treptat în funcţie de necesitatea utilizării lor în vorbire.

Aşadar, studierea gramaticii nu este prioritară şi nu reprezintă scopul principal al lecţiilor de nivelul I.

În concordanţă cu scopurile lecţiilor a fost elaborată structura lecţiilor de nivelul I. O lecţie conţine următoarele compartimente:

[2] http://www.claroline.net

- *cuprinsul,* care ne permite prin referinţele interactive să accesăm compartimentul selectat în cadrul lecţiei;
- *vocabularul,* care conţine cuvintele noi din cadrul lecţiei. Fiecare cuvânt este însoţit de sugestii cu privire la: informaţia despre partea de vorbire, (de exemplu, pentru substantiv se indică genul, formele derivate pentru numărul plural al substantivelor, pentru verb – conjugarea, sufixul verbului etc.); traducerea în limba maternă a studentului; exemple de utilizare a cuvântului indicat în fraze uzuale. De asemenea, se specifică două referinţe interactive: la un fişier sonor cu pronunţia corectă a cuvântului şi la resursele din Internet cu o informaţie mai amplă despre acest cuvânt;
- *gramatica,* ce conţine descrierea regulilor gramaticale la un nivel elementar pentru lecţia prezentată. Această informaţie trebuie să fie în limba română şi, la cererea studentului, să fie accesibilă în limba vorbită de student (de exemplu, în limba rusă, engleză etc.);
- *textul,* care este alcătuit în baza cuvintelor şi a gramaticii din lecţiile precedente cu utilizarea cuvintelor şi gramaticii din lecţia curentă. Fiecare propoziţie din text este însoţită cu două referinţe interactive: la fişierul sonor cu pronunţia corectă a frazei şi la fişierul cu traducerea în limba studentului;
- *modele comunicative, dicţionar de contexte minime, proverbe, fragmente din folclor* etc.;
- *exerciţii* elaborate în baza textului. Ele au ca scop întărirea deprinderilor de utilizare a noului vocabular şi a gramaticii în cadrul lecţiei. Pentru fiecare lecţie se propune elaborarea a trei tipuri de exerciţii:
  - pentru autotestare, care se propun imediat după însuşirea materialului teoretic (de exemplu, un fişier aparte cu sarcinile exerciţiilor şi răspunsurile corecte).
  - elaborate de profesor cu ajutorul mijloacelor sistemului Claroline;
  - care utilizează „corespondenţa" profesorului cu studentul.

În acest mod, în cadrul acestor lecţii se pune accent atât pe pronunţarea cuvintelor şi a frazelor uzuale, cât şi pe îmbogăţirea vocabularului. În acest scop se utilizează procedee interactive, asistate de mijloace multimedia.

2.2. Categoria a doua – nivelul 2

Scopul de bază al cursului de nivelul II este studierea sistematizată a gramaticii, în special, a morfologiei limbii române. Se presupune că informaţia despre categoriile gramaticale va fi consistentă, adică, dacă se studiază substantivul, pe parcursul a câtorva lecţii consecutive se descrie informaţia despre toate categoriile gramaticale ce ţin de substantiv: definiţia substantivului, gen, număr, caz, declinarea substantivului şi alte particularităţi ale lui. Ca şi în cadrul cursului de nivelul I, se va propune material teoretic, exemple de utilizare şi exerciţii practice. Titlurile lecţiilor se vor specifica în conformitate cu materialul gramatical expus. Materialul teoretic se prezintă cu utilizarea categoriilor şi a terminologiei uzuale. Se presupune că studentul cunoaşte această terminologie în virtutea nivelului său de pregătire. Pentru termenii de bază utilizaţi la lecţia curentă se vor folosi sugestii pentru reamintirea definiţiilor acestor termeni.

Pentru o sistematizare şi o structurizare efectivă a materialului şi, de asemenea, pentru o înţelegere mai bună, se propun scheme şi diagrame. Prezentarea regulilor gramaticale se

va explica prin utilizarea tabelelor care vor ajuta la asimilarea completă şi eficientă a materialul expus pentru instruire.

2.3. Categoria a treia – nivelul 3

Nivelul III presupune că, în principiu, studentul ştie bine limba română, cunoaşte regulile normative şi utilizarea uzuală a cuvintelor. La lecţiile de nivelul III o atenţie sporită se va acorda devierilor de la reguli, particularităţilor şi excepţiilor de la situaţiile morfologice normative (regulate). De asemenea pot fi propuse anumite reguli (situaţii) nenormative, momente speciale de utilizare a unor cuvinte şi a regulilor gramaticale. Se va explica sensul semantic şi utilizarea neologismelor. Limbajul de prezentare şi expunere a materialului se va deosebi de cel utilizat pentru descrierea lecţiilor de nivelul I şi II. Pentru cursul de nivelul III, acest limbaj este mai complicat, mai rafinat, la un nivel ştiinţific avansat. În calitate de exemple se vor utiliza citate din operele scriitorilor clasici români şi folclor. Ca şi în cazul lecţiilor de nivelul II, materialul se va expune cu utilizarea categoriilor şi terminologiei indicând sugestii pentru definirea termenilor.

3. Concluzii

În comparaţie cu alte cursuri publicate în cărţi, cursul de lecţii propus va putea fi accesat pe Internet de orice doritor de a învăţa limba română şi va conţine multe posibilităţi specifice aplicaţiilor Web. Iar în comparaţie cu materialele „pasive" în formă electronică, lecţiile se vor deosebi prin utilizarea interacţiunii studentului şi a profesorului (comunicării cu profesorul şi cu colegii din grup, schimbului de informaţie între profesor şi student) şi a mijloacelor de evaluare a cunoştinţelor.

Acest ciclu de lecţii nu-l va înlocui pe profesor, ci îl va ajuta în activitatea sa didactică cu material adiţional în predarea lecţiilor ce ţin de morfologia limbii române. Pentru studenţii ce nu cunosc limba, cursul de lecţii va deveni o sursă adiţională în procesul de studiere a limbii române, iar pentru alte categorii de utilizatori va oferi posibilităţi de aprofundare a cunoştinţelor în morfologia limbii române.

Referinţe bibliografice

Acsan, A., Cojocaru-Zavadschi, A., Cucu, L. (2004). *Limba care ne uneşte. Caiet de exerciţii. Nivelul 1*. Departamentul Relaţii Interetnice, Programul Naţiunilor Unite pentru Dezvoltare (PNUD). Chişinău.

Bărbuţă, I. (1998). *Gramatica limbii române. Scurt îndrumar*. Academia de Ştiinţe a Republicii Moldova, Inst. de Lingvistică. Chişinău, Litera, 152p. (în l. rusă)

Bărbuţă, I., Callo, T., Cojocaru-Zavadschi, A., Constantinovici, E., Cucu, L. (2003). *Limba care ne uneşte. Manual. Nivelul 1*. Departamentul Relaţii Interetnice, PNUD. Chişinău, (http://cnt.dnt.md/undp/).

Bărbuţă, I., Cicală, A., Constantinovici, E., Cotelnic, T., Dîrul A. (2000). *Gramatica uzuală a limbii române*. AŞM, Inst. de Lingvistică. Litera, Chişinău. 326 p.

Bertea, M. (1996). *Gramatica explicativă a limbii române plus Vocabular Ortografie Ortoepie.* Ed. a IV, revăz. şi adăug., Ed. Venus. Bucureşti, Î.E.P. Ştiința. Chişinău.

Cazacu, T., Vrabie, D. (2006) *Româna: eficient şi atractiv. Gramatica limbii române în scheme şi tabele.* Integritas, Chişinău. 52 p.

Consiliul Național pentru Curriculum. (2002). *Ghid metodologic pentru aplicarea programelor de limba şi literatura română – Învățământ primar şi gimnazial*. C.N.C., Bucureşti, România.

Cruceru, C., Teodorescu, V. (2005). *Gramatica limbii române*. Editura 100+1 GRAMAR, Bucureşti, 151 p.

Dumeniuc, I.Z. Matcaş, N.G. (1989). *Limba Moldovenească. Manual pentru autodidacți.* Chişinău, Lumina, 362 p. (în l. rusă)

Gramatica limbii române. (1963). Academia Republicii Române, v1, v2, Bucureşti.

Ionescu, M. (2001). *Didactica modernă*, Radu I. coord., Dacia, Cluj, România.

Irimia, D. (1997). Morfo-sintaxa verbului românesc. Universitatea „A.I.Cuza", Iaşi, România, 410 p.

Irimia, D. (2004). *Gramatica limbii române.* Polirom, ed. II, Iaşi, România.

Ivănuş, D. (1997). *Metodica predării limbii şi literaturii române în gimnaziu şi liceu*. Avrămeanca, Craiova, România.

Limba română contemporană. (1999). Îndrumar pentru persoane nevorbitoare de limba română. Chişinău, Litera, 336p. (în l. rusă)

Molan, V., Părvulescu, L., Teodorescu, I. (1995). *Limba noastră-i o comoară. Exerciții de limbă română pentru ciclu primar.* Ediția a II-a. Ed. Petrion. Bucureşti.

Molan, V., Teodorescu, I., Dobrin, E. (1996). *Gramatică, ortografie şi punctuație pentru toți copiii.* (Clasele II-IV). Editura Petrion. Bucureşti. Chişinău.

Nastasenco, O. (1996). *Gramatica limbii române în tabele*. Chişinău. VIRT. (în l. rusă)

Nuță, S. (2000). *Metodica predării limbii române la clasele primare.* Aramis, Bucureşti, România.

Pop, L. (2000). *Româna cu sau fără profesor.* Ed. a IV, revăz. şi adăug. Echinox, Cluj.

Zaiuncikovski, I.P., Repina, T.A. (1989). *Limba Română. Curs superior. Manual pentru anii de studii II–III ai facultăților filologice ale universităților*. Moscova. Visşaia şcola, (în l. rusă)

Capitolul 4

Modelare lingvistică

STRUCTURA GRUPULUI VERBAL, PREDICAȚIA LEXICALĂ ŞI REPREZENTAREA LOGICĂ A PREDICATULUI ÎN LIMBA ROMÂNĂ

NECULAI CURTEANU, DIANA TRANDABĂȚ[1, 2], ALEX MORUZ[1,2]

[1]*Institutul de Informatică Teoretică, Academia Română, Filiala Iaşi*

[2] *Facultatea de Informatică, Universitatea "Al.I.Cuza" Iaşi*

{curteanu, dtrandabat, mmoruz}@iit.tuiasi.ro

Rezumat

Articolul include o prezentare elementară a *teoriei proiecțiilor funcționale* FX-bar, introduce *predicația lexicală* în locul predicației clasice, pune în evidență principalele *substructuri (subgrupuri) ale grupului verbal*, şi propune o *reprezentare unitară* a predicatului verbal şi nominal în logică intensională / extensională.

1. Introducere

Lucrarea îşi propune să prezinte succint câteva rezultate legate de substructurile sintactice / semantice ale *grupului verbal* românesc (*verbal group*, VG) (Monachesi, 2005), (Barbu, 1999), (Dobrovie-Sorin, 1994), pornind de la instrumente şi argumente cunoscute în literatură, cărora li se adaugă mecanismele *teoriei* FX-bar a *proiecțiilor funcționale* (Curteanu, 2003-2004, 2005). Sunt schițate probleme şi soluții în cadrul teoriilor sintactice ale VG, cu accent pe semantica lexicală, interesul major fiind orientat către *parsarea* VG, *subgrupurile verbale* ale VG, *proiecțiile* FX-bar (directe şi inverse) ale VG, definirea *predicației lexicale* (în locul celei clasice), şi *reprezentarea unitară*, în logică intensională / extensională, a predicatului verbal şi nominal.

2. Proiecții FX-bar directe şi inverse, de nivel local şi global

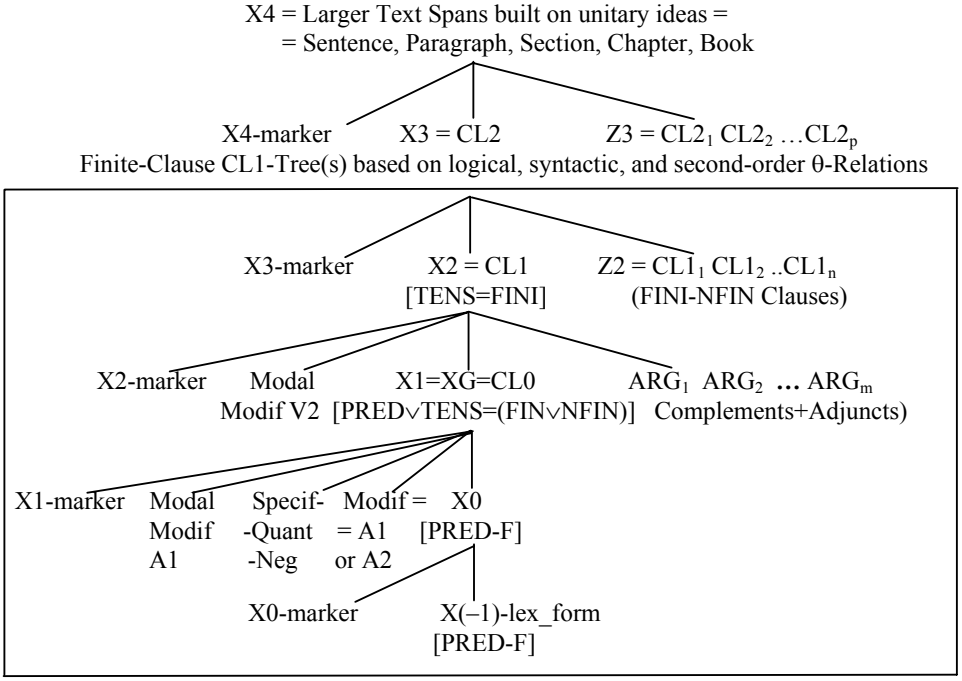
În (Curteanu, 2005) a fost (re)introdusă schema generală FX-bar din Fig. 1, ce foloseşte clasele de marcheri SCD şi ierarhia de tip graf a acestor clase, un instrument esențial pentru a reprezenta structurile sintactico-semantice de nivel propozițional (clauzal) şi pentru a stabili dependențele dintre ele. Partea inferioară a schemei, încadrată în chenar, reprezintă nivelul local, intra-clauzal, iar partea superioară indică nivelul global, inter-clauzal şi de discurs, bazat pe relații retorice RST (Mann & Thompson, 1988).

Să precizăm că în abordarea noastră folosim termenul de „*finit*" pentru toate *formele verbale* ce corespund unui *mod personal*, în timp ce termenul „*nefinit*" este atribuit formelor verbale ce corespund „modurilor" nepersonale, cunoscute şi ca *forme absolute* ale verbului (*infinitiv*, *gerunziu*, *participiu*, *supin*). În acest sens, clauza (sau propoziția) al cărei predicat are ca nucleu semantic un verb în formă predicativ/verbală (*finită*), o vom numi *clauză finită*. *Clauza nefinită* corespunde formelor verbale predicativ/nepersonale (*nefinite*), ale categoriilor lexicale majore V (verb), N

(substantiv), şi A (Adjectiv-Adverb), care posedă trăsătura de *predicaţionalitate* (*deverbalitate*). Parsarea predicatului verbal sau nominal revine la determinarea *grupului verbal* (Verbal Group, VG). Facem observaţia că determinarea VG (finit) de natură copulativă (nepredicativ/personală), notat TASG – *Tense Auxiliary SubGroup* în (Curteanu & Trandabăţ, 2006), este o etapă esenţială în parsarea predicatului nominal.

Interpretarea schemei FX-bar se face *bottom-up* pentru trecerea la o sintagmă mai complexă, dar şi *top-down* pentru stabilirea dependenţelor şi pentru adăugarea de elemente lexicale şi/sau sintagme de acelaşi nivel. Nivelul X0 este cel lexical-textual, din care se coboară la nivelul intrărilor de lexicon X(–1), unde categoriilor *lexicale majore* N (Noun), V (Verb), A (Adjectiv-Adverb) li se poate atribui *trăsătura de predicaţionalitate* (sau *deverbalitate*), notată PRED-F. Elementele de lexicon sunt (FX-bar) proiectate pe nivelul X0 după ce li se aplică X0-marcherul de *flexionare morfologică*. Prin adăugarea *on-line* de elemente lexicale noi se constituie (incremental) structurile sintagmatice de nivel X1, având ca nucleu sintactic categoriile lexicale N, V, A: *grupul nominal*, *verbal*, *adjectival-adverbial*, notate respectiv NG, VG, AG. FX-bar proiecţia lui X0 pe nivelul X1 (sau XG) poate fi însoţită de modificator (verb modal), specificator (negaţie sau cuantificator), sau auxiliare. V1 (sau VG) poate fi finit (formă predicativ/personală) sau nu, proiectându-se mai departe în schema FX-bar pe nivelul X2, alături de posibili modificatori modali şi argumente directe şi indirecte (adjuncţi), formând propoziţia (clauza finită) pe nivelul X3. Împreună cu alte clauze, se trece la proiecţia inter-clauzală. Direcţia ascendentă ilustrează mecanismul *proiecţiei* FX-bar *directe*, incluzând FX-bar proiecţia VG în clauză.

Figura 1: FX-bar schema pentru structuri de nivel local (clauzal) şi global (discurs)



Direcţia de *proiecţie* FX-bar *descendentă*, top-down în schema din Fig. 1, reprezintă *proiecţia inversă*, care include *proiecţia* FX-bar *inversă* a VG. Mecanismul este folosit pentru a stabili în mod corect proiecţia FX-bar directă a unui VG în clauză, prin

specificarea de restricţii fonologice, morfologice, sintactice şi semantice asupra nucleului semantic (eventual predicațional) şi asupra *theta*-argumentelor *directe* ale VG.

2.1. Predicația clasică şi predicația lexicală

Pentru orice intrare N, V, A, sunt reprezentate la nivelul X(–1) de lexicon: (*i*) trăsătura de *predicaționalitate*; (*ii*) *valența* (*aritatea* şi *tipul* – ordinul logic al – argumentelor); (*iii*) *diateză semantică*; (*iv*) *restricții* sintactico-semantice ale argumentelor (directe) ale nucleului semantic al VG. *Predicația clasică*, cunoscută ca fiind perechea (Subiect-Gram, Predicat), poate fi considerată doar una din fațetele VG al cărui nucleu semantic poartă *trăsătura de predicaționalitate* PRED-F (Curteanu, 2003-2004). Această pereche corespunde fie rolului tematic (*theta*-rol) de „Actor” / „Actant”, fie de „Obiect” / „Pacient”, în funcție de *diateza activă*, respectiv *pasivă* a VG. Un al treilea argument de natură *theta*-semantică poate fi „Beneficiar” / „Adresant” al predicației din VG. Predicația clasică (Subiect-Gram, Predicat) este rescrisă ca fiind perechea ($\text{SUBJ}_{Obliqueness\,=\,0}$, PREDF_verb) corespunzând *theta*-rolului de „Actor” / „Actant” sau „Obiect” / „Pacient” în lista SUBCAT a argumentelor directe cerute de valența nucleului semantic al VG (Pollard&Sag; 1994). Trebuie specificat că în mod normal există cel puțin două liste SUBCAT: $\text{SUBCAT}_{oblic\_order}$, ce conține argumentele sintactice ale verbului predicațional PREDF_verb în ordinea crescătoare a oblicității, şi $\text{SUBCAT}_{theta\_order}$ care conține argumentele în ordinea sistemică a argumentelor *theta*-semantice. De obicei, cele două liste SUBCAT sunt identice numai pentru diateza activă. La predicația clasică (Subiect-Gram, Predicat) se adăugă predicațiile din Fig. 2, justificate de comportamentul lor perfect similar față de nucleul semantic al argumentelor directe:

$$(\text{SUBJ}_{Obliqueness\,=\,0},\ \text{PREDF\_Verb}\begin{bmatrix}[VG \quad Tense\_Aspect]\\ \textbf{\textit{Semantic\_Diathesis}}(\text{SUBJ, OBJD, OBJI})\\ \qquad = (\theta(\text{SUBJ}),\ \theta(\text{OBJD}),\ \theta(\text{OBJI}))\\ Agreement(\text{SUBJ},\ Inflection\_\text{VG})\end{bmatrix})$$

$$(\text{OBJD}_{Obliqueness\,=\,1},\ \text{PREDF\_Verb}\begin{bmatrix}[VG \quad Tense\_Aspect]\\ \textbf{\textit{Semantic\_Diathesis}}(\text{SUBJ, OBJD, OBJI})\\ \qquad = (\theta(\text{SUBJ}),\ \theta(\text{OBJD}),\ \theta(\text{OBJI}))\\ Agreement(\text{OBJD},\ Clitic\text{OBJD\_VG})\end{bmatrix})$$

$$(\text{OBJI}_{Obliqueness\,=\,2},\ \text{PREDF\_Verb}\begin{bmatrix}[VG \quad Tense\_Aspect]\\ \textbf{\textit{Semantic\_Diathesis}}(\text{SUBJ, OBJD, OBJI})\\ \qquad = (\theta(\text{SUBJ}),\ \theta(\text{OBJD}),\ \theta(\text{OBJI}))\\ Agreement(\text{OBJI},\ Clitic\text{OBJI\_VG})\end{bmatrix})$$

Figura 2: Predicația lexicală extinde predicația clasică pentru toate argumentele directe

Aceste noi predicații „tradiționale” se bazează pe trăsătura de *predicaționalitate* (deverbalitate) PRED-F, ce este atribuită la nivel lexical în nucleul VG. Problema rolului 'special' al subiectului în lista SUBCAT a argumentelor este rezolvată în structura propusă prin demonstrarea similarității relației *Subiect – PREDF_verb* cu relațiile *Complement (Direct / Indirect) – PREDF_verb* şi prin folosirea funcției de *diateză semantică* (Irimia; 1997). *Acordul* dintre VG cu nucleu predicațional şi

argumentele sale directe, fie că se face prin *flexionare morfologică* (pentru subiectul gramatical SUBJ) sau prin *pronume clitice* (pentru OBJD şi OBJI), este un alt argument major ce susține *echivalența* elementelor constitutive ale *predicației lexicale*.

2.2. Diatezele formal-sintactică şi lexical-semantică

Diatezele *formal-sintactică* şi *lexical-semantică* sunt interpretări semantice diferite ale aceluiaşi şir de elemente lexicale (de suprafață) din VG. Se recunoaşte una din cele 3 diateze sintactice, se coboară la descrierile lexical-semantice din lexicon, iar nucleul semantic selectat se proiectează FX-bar într-una din cele 6 diateze semantice ale VG. Funcția *Semantic_Diathesis*(*Actor*, *Patient*, *Addressee*) primeşte ca intrare diateza sintactică a VG, reprezentată de lista de (cel mult 3) argumente directe, în ordinea oblicității $SUBCAT_{oblic\_order}$ şi returnează diateza semantică împreună cu argumentele în ordinea $SUBCAT_{theta\_order}$. Această soluție obligă *actorul*-subiect şi subiectul gramatical (cel mai puțin oblic element) să ocupe fiecare poziția corectă. Transformarea argumentelor din diateza sintactică în cea semantică este stabilită de un tabel de corespondență prezentat în (Curteanu & Tradabăț, 2006).

3. Substructurile grupului verbal VG

3.1. Natura gramaticală a componentelor grupului verbal

Necleul semantic verbal antrenează în jurul său intensificatori, modificatori, particule clitice etc. formând *grupul verbal* VG. Componentele VG sunt: (a) verbul lexical; (b) auxiliar (de timp, de pasiv); (c) verb semi-auxiliar; (d) verbe de restructurare (modale, de aspect, de mişcare); (e) adverbe speciale (*mai*, *cam prea, şi*, *tot*), fiecare având poziții sintactice şi semantice bine definite în VG; (f) clitice pronominale (ce pot apare cu sau fără *dublarea* prin argumentele directe); (g) negații. Exemple de VG-uri: *am plecat*, *nu că nu mi-l va mai şi plăti greu; nu i-ar fi trecut; n-ar trebui să putem cheltui; ar fi trebuit să nu mai poată trece*.

Dacă în literatura de specialitate este acceptat în general că *cliticele pronominale* şi *intensificatorii* au comportament de *afix*, în timp de *complementizatorii* şi *negația* au proprietăți de *cuvânt*, natura *auxiliarelor* este încă controversată. Ele sunt cuvinte morfo-sintactice după (Monachesi, 2005), afixe după (Barbu, 1999), sau clitice simple după (Zwicky, 1985).

3.2. Subgrupul auxiliarului de timp şi subgrupul verbului modal

Cel mai frecvent şi mai natural *subgrup verbal* (Verbal SubGroup, VSG) este *Subgrupul Auxiliarului de Timp* (Tense Auxiliary SubGroup, TASG): *voi fi, aş fi, am fi, sunt*. TASG poate conține, pe lângă auxiliar, *adverbe speciale* sau *negație*. TASG este *nesaturat* (*i.e.* cere un argument lexical) şi *nepredicațional* deoarece nucleul său semantic este un verb copulativ. Din acest motiv, asociem TASG cu o *funcție de atribuire* $x := y$, unde x şi y sunt variabile sau constante *intensionale* sau *extensionale.*

TASG este un subgrup al grupului *Auxiliarului de Pasiv* (Passive Tense Auxiliary Subgroup, PTASG). PTASG are acelaşi sens de atribuire copulativă, nefiind saturat, la

fel ca TASG. Exemple de subgrupuri PTASG sunt: *i-a fost dată; nu-i va fi recunoscută (diploma), trebuia să fie arestat.*

O altă substructură tipică a VG se referă la modalitate. *Subgrupul modal* (ModVSG) derivă din TASG, cu diferența că nucleul semantic pe care îl are acest subgrup este un verb modal (*a trebui*, *a putea*), cu posibile inserții de adverbe speciale sau negații. Exemple de ModVSG sunt: *Am fi putut alerga*; *Îi trebuie apă [ca] să crească*; *Nu-l mai puteam reține peste noapte*; *Ar fi putut-o vedea tot satul*; *Nu i l-ar fi putut da cu împrumut*. ModVSG este nesaturat şi cere ca argument o altă predicație, de ordin unu sau doi.

3.3. Modelare intensională / extensională a predicatului verbal şi nominal

Să considerăm următoarea serie de predicate pentru analiza intensională / extensională. **(a)** *a fost predată.* Acest VG este un predicat verbal în diateza pasivă, la timpul trecut. „*Predată*", nucleul căruia i se aplică TASG, este un predicat intensional de valență 3 (nefinit, *i.e.* formă verbală absolută). Astfel, reprezentarea pentru o propoziție de genul „*Lucrarea a fost predată.*" este $lucrarea(Y) :=_{past} predată_{passive}(x, Y, z)$, unde *Y* este o variabilă extensională, iar *x* şi *z* sunt predicate extensionale. Aici am considerat înțelesul extensional al substantivului „*lucrarea*", dar acesta poate avea şi un sens intensional-predicațional: *Lucrarea cu migală a pereților exteriori de către meşterii populari.*

(b) *a fost plecată.* Acest VG este în gramatica clasică un predicat nominal. Totuşi poate fi interpretat şi ca predicat verbal al cărui nucleu semantic este o predicație reprezentată de un verb intranzitiv (valență 1). O asemenea categorie are reprezentarea *plecată*(*x*(*X*)), unde *x* este un predicat extensional iar *X* este o variabilă extensională.

(c) *a fost frumoasă.* Acesta este un predicat nominal clar, al cărui nucleu semantic „*frumoasă*" nu mai este o categorie predicațională. Totuşi, deoarece orice modificator adjectival, predicațional sau nu, necesită (cel puțin) un argument nominal, exprimat ca un predicat extensional *x*(*X*), reprezentarea corectă este *frumoasă*(*x*(*X*)), cu *x* un predicat extensional şi *X* variabilă extensională.

(d) *a fost elevă.* Acesta este un predicat nominal clasic, nucleul semantic fiind reprezentat de predicatul extensional *elevă*(*X*), *X* variabilă (sau constantă) extensională.

(e) *va fi trădarea.* Şi acest exemplu este un predicat nominal, constând dintr-un subgrup TASG al cărui nucleu semantic „*trădarea*" este un substantiv predicațional. Reprezentarea intensională a acestei structuri este $P :=_{\text{future; act diat}} trădarea(x, y)$, unde *P* este o variabilă intensională corespunzătoare predicatului intensional „*trădarea*" iar *x* şi *y* sunt predicate extensionale corespunzând categoriei nominale predicaționale „*trădarea*", de valență 2. De exemplu, *P* ar putea reprezenta pronumele demonstrativ (şi implicit anaforă intensională) „*aceasta*" în exemplul *Aceasta a fost trădarea.*

Recapitulând această serie de exemple, se poate observa că predicatul are un subgrup TASG care se aplică unei sintagme verbale sau nominale a cărui nucleu nefinit (predicativ/nepersonal) variază astfel:

- *predată* = verb predicațional (intensional), valență 3, nesaturat, exemplul **(a)**;
- *plecată* = verb predicațional (intensional), valență 1 (intranzitiv), nesaturat, ex. **(b)**;

- *frumoasă* = adjectiv nepredicațional (extensional), nesaturat, care necesită un nucleu nominal (predicat extensional): exemplul **(c)**;
- *elevă* = substantiv nepredicațional (extensional), saturat, exemplul **(d)**;
- *trădarea* = substantiv predicațional (intensional), valență 2, nesaturat, ex. **(e)**.

Punctul de tranziție de la predicatul verbal către cel nominal este localizat în exemplele **(b)** şi **(c).** Aceste două predicate (intuite corect de gramatică ca *nominale*) sunt de fapt un *predicat verbal* şi unul *nominal*, oferind, din motive diferite, *aceeaşi reprezentare* la nivelul logicii intensionale / extensionale: *plecată* este verb predicațional-intensional, cu un singur *argument* (fiind intranzitiv), în timp ce *frumoasă* este o adjectiv nepredicațional, deci extensional; fiind însă o categorie *nesaturată*, necesită ca *nucleu nominal* tot un predicat extensional.

Referințe Bibliografice

Barbu, A.M. (1999). The Verbal Complex. *Studii şi Cercetări Lingvistice*, L, no.1, Bucureşti, p. 39-84.

Curteanu, N. (2003-2004). Contrastive Meanings of the Terms "Predicative" and "Predicational" in Various Linguistic Theories (I, II). *Computer Science Journal of Moldova*, Vol. 11, No. 4, 2003 (I); Vol. 12, No. 1, 2004 (II).

Curteanu, N. (2005). Functional FX-bar Theory Extended to Discourse (Rhetorical) Structures. In *'Intelligent Systems' Conference Volume*, H.-N. Teodorescu et al. (Editors), Performantica Press, Iaşi, pp. 169-182.

Curteanu, N., Trandabăț, D. (2006). Functional (F)X-bar Projections for Local and Global Text Structures. The Anatomy of Predication. *Revue Roumaine de Linguistique*, Bucureşti.

Dobrovie-Sorin, C. (1994). *The syntax of Romanian. Comparative Studies*. Berlin: Mouton de Gruyter.

Irimia, D. (1997). *Morfosintaxa verbului în limba română*. Ed. Univ."Al. I. Cuza" Iaşi.

Mann, W., Thompson, S. (1987). *Rhetorical Structure Theory: A Theory of Text Organization*. Research Report RS-87-190, Information Sciences Institute, University of Southern California, Marina del Rey, California, 80 pp.

Monachesi, P. (2005). *The Verbal Complex in Romance. A Case Study in Grammatical Interfaces*. Oxford University Press, Oxford Studies in Theoretical Linguistics.

Pollard,C., Sag, I. (1994). *Head-Driven Phrase Structure Grammar*. The University of Chicago Press, Chicago & London.

Zwicky, A. (1985). Clitics and Particles. *Language*, Vol. 62, No. 2, pp. 283-305.

PERSPECTIVE SEMANTICE DIN NOU: CUM ŞI SUB CE FORMĂ AVANSĂM LEXICOLOGIC SPRE *DLRI*

CRISTINA FLORESCU

Institutul de Filologie Română „Al. Philippide", Academia Română, Filiala Iaşi

cristinafl24@yahoo.fr

Rezumat

Din perspectivă lingvistică, plecându-se de la rezultatele analizelor anterioare privind tratamentul informatic al DLR, sunt propuse două modalităţi de organizare şi valorificare a unor eşantioane semnificative lexical din DLR (serie nouă) + DA (seria veche) în ideea verificării randamentului tratării în format electronic a acestui corpus lexicografic al limbii române, unic şi din punct de vedere semantic.

Faptul că limba română îşi caută intens vadul academic informatizat reprezintă unul dintre elementele evidente din arealul lingvisticii contemporane. Această afirmaţie este un loc comun. Unghiul de abordare a respectivei probleme diferă însă, deschizând un evantai de posibilităţi aflate adesea în raport de opoziţie.

Ne vom opri asupra a trei chestiuni.

I. Prima chestiune pune în corelaţie consideraţiile noastre de astăzi cu cele avansate anul trecut, despre „schiţa unui viitor lexicografic imediat", respectiv suita de propuneri şi sugestii privind realizarea *Dicţionarului limbii române informatizat şi unificat* (Florescu, 2005)[1]. În respectivul material se propune o modalitate de eşalonare pe parcursul a zece ani a complicatelor operaţii lexicografice şi informatice care sunt avute în vedere în procesul de actualizare lingvistică a vechiului DA; se consideră că întreprinderea respectivă trebuie să aibă ca punct de plecare introducerea în format electronic a DA şi DLR.

1. Cea mai mare parte a factorilor punctuali, a elementelor concrete şi a reperelor temporale din seria de soluţii cuprinse în lucrarea prezentată anul trecut, s-a modificat, unii factori – în mod previzibil, alţii – imprevizibil. Un exemplu din ultima categorie: s-a micşorat drastic numărului lingviştilor lexicografi, autori ai DLR, dispuşi să reia conform aceloraşi parametri (fie şi în format electronic) laborioasa şi tehnica redactare lexicografică, extrem de tensionată profesional în ultimii ani.

Se adaugă o problemă deloc de neglijat: transformarea DLR în text adnotat presupune în mod obligatoriu desfăşurarea unui demers filologic deosebit de minuţios, cronofag, care uzează specialistul în mod neprofitabil şi neperformant: corectarea lingvistico-lexicografică a unui număr de câteva zeci de mii de pagini tip (2000 semne/pagină).

[1] Cf. http://consilr.info.uaic.ro/ro/resources/pre/DLRI/DLRI.ppt

Din cadrul „propunerilor şi sugestiilor privind *Dicţionarul limbii române informatizat şi unificat*" a rămas cu certitudine punerea în circulaţie a iniţierii unei planificări din perspectivă lingvistică şi informatică, factor care a aplecat încă o dată benefic spiritele asupra problemei.

La ora actuală conducerea Academiei Române lucrează la un acord de colaborare dintre intituţiile filologice şi informatice, acord care va reprezenta primul pas în realizarea DLRI.

2. În limba română, corpusurile de texte existente, fie ele ediţii ale unor texte vechi, fie corpusuri ale limbii vorbite (limbaje regionale, familiare, televizuale, radiofonice, socio-profesionale, argotice, eşalonate stilistic sau pragmatic, psiholingvistic sau vericondiţional, neologice sau conservative ş. a. m. d.), toate acestea nu sunt supuse *totdeauna* unor normative / norme lingvistice prealabile. Conceptul de normă nu este utilizat aici în sens scolastic, ci în sens larg, desemnând reguli care corespund sistemului limbii, uzuri cuprinse obiectiv în anumite categorii, tipologii, paradigme, valabile şi incorporate caracteristicilor limbii române.

Se înţelege că un instrument informatizat în sensul discuţiei noastre este cu atât mai elastic (şi cu atât mai performant prin urmare) cu cât înmagazinează (adecvându-se prin prelucrare informatică) un număr cât mai mare de texte. Este scopul oricărui demers informatic aplicat unei limbi /unor limbaje.

Această aplecare informatică asupra limbii pune totdeauna problema unei prealabile analize lingvistice care să direcţioneze principial şi metodologic faptele lingvistice.

Limba română în speţă este inegal prelucrată lingvistic. Are gramatici performante, ortografie elaborat actualizată. Fără discuţie, caracterul specific al limbii române în areal romanic este marcat şi analizat. La nivel romanic, istoria limbii române, gramatica şi ortografia limbi române, prin urmare partea mai puţin elastică (dacă se acceptă acest calificativ) a limbii române este vizualizată sistemic în cea mai mare parte a ei. Este suficient să ne referim la lucrările de tipul MDA, [1,2]DOOM sau [1,2]*Gramatica limbii române.* Partea de maximă elasticitate, efervescenţă a limbii, respectiv semantismul limbii române are cele mai multe lacune analitice. Celelalte limbi romanice dispun (dincolo de o suită, cu vechime apreciabilă, de lexicoane tezaur de diferite tipuri) de tratate de semantică şi/sau de corpusuri bogate de texte eşalonate pe secole, curente culturale, scriitori, limbaje de specialitate etc.

În mod evident, DLR reprezintă la ora actuală, pentru limba română privită în totalitatea ei, cea mai bogată sursă semantică structurată cu fermitate şi normată lingvistic adecvat.

Faptul pare de la sine înţeles şi cunoscut. Lucrurile nu stau însă totdeauna astfel. Caracterul de „fără precedent" al extensiei semantice a DLR este de multe ori foarte greu acceptat.

Exemplele sunt nenumărate. Ne vom rezuma la un singur caz relativ recent.

În cadrul unei îndelungate cercetări lingvistice, cuprinse într-un volum (Florescu, 1999), s-a demonstrat faptul că în limba română există pentru verbul *a lăsa* o situaţie

specială, o idee semantică specifică, cea „a coborârii, a deplasării pe plan înclinat”[2], în cadrul tuturor celorlalte limbi romanice în care există bine dezvoltați urmaşii lat. *laxare*. Demonstrația punctuală[3] este lungă, minuțioasă şi tehnică. Au fost purtate câteva discuții pe această temă şi cu o serie de specialişti romanişti neromâni. Faptele au fost acceptate, demonstrația a convins, prin urmare.

Punctul de minimă credibilitate a fost numărul de ocurențe[4], suma de fapte lingvistico-semantice cu ajutorul cărora s-a realizat, în cadrul DLR, redactarea articolului lexicografic corespunzător analizei semantico-lingvistice a verbului românesc *a lăsa*. Pentru un lexicograf autor al DLR, a selecta, ierarhiza, închega pentru prima oară semic (la acest nivel de minuție) în jur de câteva mii de ocurențe este ceva obişnuit. Pentru un lingvist de aiurea numărul se dovedeşte incredibil.

Ceea ce demonstrază încă o dată, şi din exteriorul sistemului, că structura semantică a DLR este excepțională.

II. Al doilea aspect se referă la modalitatea concretă de unificare a celor două structuri lexicografice (DA şi DLR).

Realizarea unui grup lexical „de probă” (prin care să se actualizeze articolele DA) este obligatorie şi semnificativă.

Au existat câteva încercări meritorii în acest sens, nefăcute public. Ele însă nu au dispus de instrumentul informatic. Prin urmare, timpii de apreciere sunt în afara sferei noastre de interes.

Un prim eşantion care se dezvoltă în direcția introducerii în format electronic a *Dicționarului limbii române* a realizat, în cadrul unui grant[5], un instrument de lucru adecvat informatic şi lingvistic chestiunii numit DLRex (ale cărui caracteristici şi performanțe sunt prezentate pe larg în (Haja et al., 2005).

Am propus, ca o a doua întreprindere realizată în sensul discuției noastre, un demers de acelaşi calibru ca cel de mai sus, demers care se opreşte, pentru început, asupra unui număr de 142 de cuvinte, eşalonate în DLR şi în DA, derivate pe teren românesc cu sufixul de origine slavă (slavă veche, slavonă, bulgară şi sârbă) în *–işte*. În funcție de sursa de finanțare, cercetarea se va putea extinde şi la alte categorii de derivate.

Lingvistic problema este complicată. O categorie lingvistică este o mulțime „continuă” pe când un grup / eşantion lexicografic este discret. Trecerea de la un grup ale cărui limite sunt în mod inerent şi obiectiv vagi, la un număr bine limitat de intrări

[2] Cităm câteva structuri prototipale: *păsările se lasă pe câmp*, *X se lasă în fântână / la vale*, *s-a lăsat ceața*, *i s-au lăsat măruntaiele*.
[3] Cf. (Florescu, 1999: 46 – 57, 128 - 137)
[4] Compară, în acest sens, două studii paralele: în (Soares, 1999:77) pentru port. *deixar* (şi forma veche *leixar*) cifra citată este de aproape 4.000 ocurențe, în (Florescu, 1999: 48) pentru rom. *a lăsa* cifra citată se ridică la un număr de peste 15.000 de ocurențe; reamintim faptul că ambele cuvinte sunt echivalente lingvistic (cu etimologie comună) în areal romanic.
[5] Cf. grantul CNCSIS nr. 1415 pe 2003-2005, prezentat şi valorificat în (Haja et al., 2005).

lexicografice reprezintă o cercetare lingvistică în sine, complexă, cu multe probleme lexicologice de rezolvat.

Plecând de la faptul că acest aspect lexical (stabilirea limitelor categoriei lexicale a derivatelor în discuție) este deja rezolvat în cadrul unei cercetări lingvistice (Florescu, 2006), proiectul propune introducerea - prin scanare, OCR-izare şi trasformare în limbaj XML - în memoria computerului a unei sume de cca 200 pp. DLR şi DA - pagini echivalente articolelor care cuprind derivatele în studiu; va urma prelucrarea paginilor respective cu ajutorul DLRex, rafinarea acestuia din urmă şi realizarea unei serii de articole lexicografice actualizate din seria veche DA, conform principiilor lexicografice ale seriei noi DLR.

Existența cercetării informatice anterioare oferă actualului demers multe şanse de reuşită.

În final vom dispune de un mini-corpus lexicografic al limbii române tip DLR, adnotat în format XML, unificat lingvistic prin mijloace informatice.

III. Al treilea aspect se referă la modalitatea de valorificare a masei lexicale fără precedent, a mulțimii de citate / contexte / texte existente în seria nouă DLR.

Dacă admitem ca premiză de lucru faptul că prima parte a travaliului legat de introducerea seriei noi DLR în format electronic va fi realizată, ne găsim în fața a zece tomuri (34 de volume) DLR scanate, adnotate şi prelucrate în XML cu ajutorul DLRex-ului.

Cum poate fi valorificată acestă masă lexicală, acest corpus de micro-texte, pentru actualizarea seriei vechi DA?

Chestiunea este lexicologic şi lingvistic extrem de delicată. Reprezintă una dintre frânele care blochează eventualele aproximări şi planificări. Nu se poate avansa un proiect academic fără o relativ corectă evaluare a acestui fapt.

Părerile lingviştilor sunt extrem de împărțite: unii consideră că, în urma efortului preluării citatelor (micro-contexte) din DLR, grupul lexical aparținând seriei vechi (cuvintele din porțiunile A-C, F-J) va avea lacune lingvistice (sensuri, etimologie etc.) prea mari pentru a merita eforul respectivei preluări preluări. Alți lingvişti consideră că, în urma valorificării citatelor (micro-contexte) din DLR, „zonele albe" corespunzătoarea articolelor din DA vor fi relativ restrânse şi că trebuie făcut efortul preluării acestor citate. De exemplu, articolul lexicografic corespunzător cuvântului *casă* (redactat, prin urmare, în seria veche, în DA), prin preluarea tuturor citatelor existente în DLR în care apare cuvântul, va putea utiliza, să zicem, şi contextele în care apare sintagma *casa mare*, contexte care pot fi recuperate din DLR tom VI, Litera M, articolul MARE.

Există tentația de a spune că adevărul este undeva la mijloc. Nimic mai fals. Oricare dintre ipostazele probabile sunt posibile. Singura soluție este realizarea unui proiect parțial care să verifice în mic realitatea lingvistico-lexicografică.

Un asemenea proiect se poate opri asupra tomului ieşean XIII al literelor: V, W, X şi Y, tom care însumează 1340 pagini DLR şi 6327 de intrări (cca 5000 de cuvinte). Dintre acestea numai 325 pagini şi 1747 intrări (cca 1500 cuvinte) nu se găsesc în memoria computerului.

Scanarea şi adnotarea întregului tom XIII va fi urmată de „vărsarea" ocurenţelor / citatelor / microcontextelor - în funcţie de cuvintele relevate şi delimitate corect lingvistic şi lexicografic - în ordine alfabetică. Se înţelege că se va avea în vedere confruntarea, prin operaţii lexicologico-lexicografice specifice, cu o sumă de liste alfabetice lexicale anterioare, inclusiv acelea din DA şi din MDA.

Stabilirea listei de cuvinte în funcţie de variantele lexicale ale unui cuvânt şi de structurile omonimice va presupune eşalonarea a cel puţin trei operaţii de identificare.

1. Prima identificare va fi făcută la nivelul intrărilor.

2. Următoarea: la nivel diacronic – modalitatea de acoperire a timpilor filologi originari, vechi, moderni şi contemporani în cadrul fiecărui articol / cuvânt.

3. Ultima identificare comparativă şi corelativă va privi extensivitatea semică (sensuri, subsensuri, construcţii fixe, variaţii dialectale, gramaticale şi stilistice).

Se va realiza un instrument de lucru adecvat informatic şi lingvistic, cu ajutorul căruia informaticianul şi lexicograful vor putea face o apreciere corectă privind gradul de performaţă lingvistico-informatică a demersului respectiv.

Rezultatele acestor două intreprinderi - a) eşationul DLRI al derivatelor în *–işte* şi b) posibilităţile de acoperire semantică a nevoilor de actualizare a DA - vor limpezi o suită de probleme care, la ora actuală, în lexicografia academică românească, par insurmontabile nu atât prin imposibilitatea realizării, cât prin numărul prea mare de variante posibile.

Referinţe bibliografice

Haja, G., Dănilă, E., Forăscu, C., Aldea, B. (2005). *Dicţionarul limbii române (DLR) în format electronic. Studii privind achiziţionarea*. Editura Alfa, Iaşi.

Cristea, D. (2005). *Resurse lingvistice şi tehnologii ale limbajului natural. Cazul limbii române*. Prelegerile Academiei Române. Filiala Iaşi.

[1,2]*Gramatica limbii române.* Academia Română. Institutul de Lingvistică "Iorgu Iordan-Al.Rosetti", Bucureşti. Editura Academiei Române. Ediţia I, 1963. Ediţia II, 2005.

DA = *Dicţionarul limbii române.* Academia Română. Bucureşti. Librăriile Socec, Universul. Tom I-II, 1913-1937.

DLR = *Dicţionarul limbii române. Serie nouă.* Bucureşti. Editura Academiei Române. Tomul VI ş. u., 1965 etc.

[1,2]DOOM = *Dicţionarul ortografic, ortoepic şi morfologic al limbii române.* Ediţia I, Editura Academiei Române, Bucureşti, 1982. Ediţia a II-a, Univers Enciclopedic, Bucureşti, 2005.

Florescu, C. (1999). *Gîndire specifică şi gîndire europeană în semantismul românescului* ***a lăsa***, Iaşi, Document.

Florescu, C. (2005). Propuneri şi sugestii privind Dicţionarul limbii române unificat şi informatizat (DLRI). *Atelierul de lucru ConsILR. Resurse lingvistice şi instrumente pentru prelucrarea limbii române*, Iaşi, 3 noiembrie 2005.

Florescu, C. (2006). *Linişte şi derivatele pe teren românesc în –****işte****.* În Volumul Omagial Mioara Avram, Bucureşti, Editura Academiei, 2006 (sub tipar).

MDA = *Micul Dicţionar Academic*, Academia Română, Institutul de Lingvistică "Iorgu Iordan", Bucureşti, Univers Enciclopedic, vol. I-IV, 2001-2002.

Soares da Silva, A. (1999). *A Semântica de 'Deixar'. Uma Contribuição para a Abordagem Cognitiva em Semântica Lexical.* Braga, Fundação Calouste Gulbenkian.

MODELAREA RELAȚIILOR SEMANTICE ÎNTR-UN DICȚIONAR DE SIMBOLURI

CRISTINA CIOCÂRLĂU, MIHAELA BRUT

Facultatea de Informatică, Universitatea "Al.I.Cuza" Iaşi

{cciocarlau, mihaela}@info.uaic.ro

Rezumat

Dicționarele de simboluri sunt instrumente de lucru foarte importante pentru cei care lucrează în domenii precum critica literară, lingvistica, literatura, dar şi relații publice, marketing sau arte vizuale. Considerând poezia eminesciană un punct de plecare pentru descoperirea bogăției de semnificații ale cuvintelor din limba română, aplicația prezentată în articolul de față oferă posibilitatea întreținerii şi utilizării unui dicționar de simboluri, organizat pe arii semantice. Adăugarea unui nou cuvânt în dicționar de către administrator are ca efect evidențierea lui în toate poeziile eminesciene, utilizatorul obişnuit având posibilitatea de a-i vizualiza diversele accepțiuni şi de a-şi dezvolta propriile comentarii, definiții, adnotări. În plus, alături de definițile simbolului selectat, utilizatorului îi sunt semnalate şi simbolurile din aceeaşi arie semantică. Astfel, având acces la diversele valențe semantice ale unui cuvânt, utilizatorului îi va fi facilitată utilizarea mai nuanțată şi mai percutantă a acestuia.

Aplicația este implementată în XHTML, CSS, XML, DOM PHP, asocierea de metadate XML simbolurilor poetice constituind un prim pas spre dezvoltarea unei aplicații de Web semantic pentru simbolurile înmagazinate de limba română.

1. Introducere

Operele artistice de valoare propun o uimitoare dialectică a semnificațiilor şi contextelor, incitând la interpretări multiple, oferind astfel delectări intelectuale şi spirituale. Simbolurile poetice consacrate apar în operele marilor creatori îmbogățite semantic sau chiar resemantizate, iar miturile sunt reinterpretate sau le este extinsă semnificația. Deseori, artiştii îşi dezvoltă propriile simboluri poetice, care contribuie într-un mod insolit şi fascinant la stabilirea corespondențelor simbolice subterane ale operei lor. Privită prin prisma unei creații artistice, *constelația simbolurilor*[1] are o perspectivă personalizată. Metamorfoza semnificațiilor se dezvăluie studiind maniera în care zestrea culturală a umanității a fost asimilată şi trasfigurată de o experiență artistică particulară.

Dicționarele de simboluri sunt punctul de plecare în acest demers. Dicționare precum *Dicționarul de simboluri* al lui Jean Chevalier şi Alain Gheerbrant, sau *Elsevier's Dictionary of Symbols and Imagery* scris de Ad de Vries şi revizuit de Arthur de Vries,

[1] Termen consacrat de Gilbert Durand, *Structurile antropologice ale imaginarului*, Editura Univers Enciclopedic, Bucureşti, 1998.

ori *The Complete Dictionary of Symbols* de Jack Tresidder sunt instrumente de lucru uzuale ale criticilor literari şi artistici.

Disponibilitatea on-line a acestui tip de dicţionare ar putea facilita accesul la informaţie. Avantajul dicţionarului de simboluri propus de noi este acela că permite criticilor literari - sau altor tipuri de utilizatori interesaţi - să îşi construiască un sistem de adnotări referitoare la semnificaţiile particulare ale simbolurilor în opera eminesciană, aplicaţia fiind uşor extensibilă la operele diverşilor artişti sau la varii contexte. Astfel, paralel cu definiţiile furnizate de dicţionar pentru un anumit simbol, utilizatorul îşi gestionează un set de comentarii legate de aceste semnificaţii. În plus, pentru fiecare simbol, sunt oferite şi referinţele la simbolurile din aceaşi arie semantică.

Deoarece simbolurile sunt foarte importante în dezvoltarea unor diverse tipuri de discursuri, considerăm util dicţionarul propus de noi şi celor care lucrează în publicitate, marketing, relaţii publice, arte vizuale.

2. Funcţionalităţile oferite de dicţionarul de simboluri

Aplicaţia *Dicţionar de simboluri on-line* oferă suport pentru două tipuri de utilizatori: utilizatori obişnuiţi, care doresc să acceseze suportul oferit de dicţionar, şi administrator, care are posibilitatea gestionării globale a aplicaţiei.

Figura 1: La afişarea unei poezii sunt marcate simbolurile existente în dicţionar

Utilizatorul obişnuit are acces, după autentificare, la lista poeziilor eminesciene. La afişarea unei poezii, simbolurile care se regăsesc în dicţionar vor apărea marcate în text. Selectarea unui simbol din poezie va determina afişarea definiţiilor simbolului, aşa cum

sunt ele memorate în dicționar, precum şi enumerarea simbolurilor din aceeaşi arie semantică. În partea de jos a ecranului, utilizatorul are la dispoziție o zonă în care poate adăuga diferitele definiții ale simbolului ales, şi în acelaşi timp poate aduce modificări şi completări acestor definiții, creând un mic comentariu aplicat la contextul curent. Operația de redactare a zonei respective poate continua şi după selectarea altui simbol din poezia curentă sau dintr-o altă poezie; informația este persistentă. În plus, alături de afişarea definițiilor pentru simbolul curent selectat, utilizatorului îi sunt puse la dispoziție referințe către simbolurile din aceeaşi arie semantică, astfel încât, dacă se doreşte un studiu asupra unei problematici particulare a poeziei eminesciene, să se ajungă la toate fragmentele şi simbolurile elocvente. Această zonă de editare rămâne activă pe parcursul navigării în alte poezii, iar conținutul editat poate fi salvat, fiind accesibil ulterior doar utilizatorului curent. Pe viitor ar putea fi adăugată o opțiune de partajare între utilizatori a comentariilor făcute unor diverse simboluri, figuri de stil, poezii.

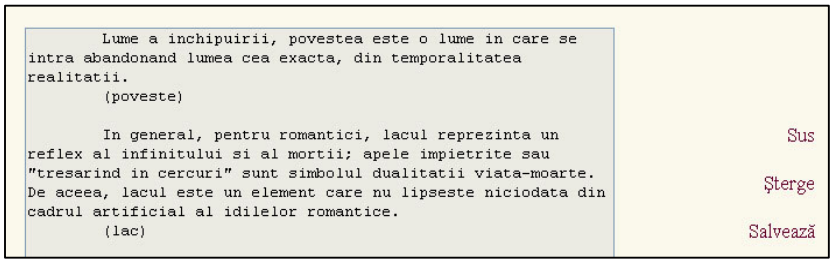


Figura 2: Zona de editare în care se pot adăuga atât definiții de simboluri, cât şi comentarii proprii

Funcția principală a administratorului aplicației este gestionarea intrărilor în dicționar. Adăugarea unui nou simbol în dicționar presupune, alături de precizarea definițiilor acestuia, indicarea simbolurilor conexe, făcând parte din aceeaşi arie semantică. Desigur, pot fi indicate drept simboluri conexe doar simbolurile existente la momentul curent în dicționar, includerea ulterioară a unui simbol ce poate avea această însuşire trebuind făcută cu indicarea simbolului curent drept conex. Ar putea fi dezvoltat pe viitor un mecanism automatizat de moştenire sau de tranzitivitate a apartenenței la aceeaşi arie semantică.

Modificarea definiției unui simbol existent este o altă funcționalitate accesibilă administratorului. Simbolul este selectat dintr-o listă derulantă, iar definiția nou completată o va substitui pe cea veche. Un procedeu similar se aplică în cazul în care administratorul doreşte îmbogățirea colecției de definiții corespunzătoare unui anumit simbol cu o nouă definiție.

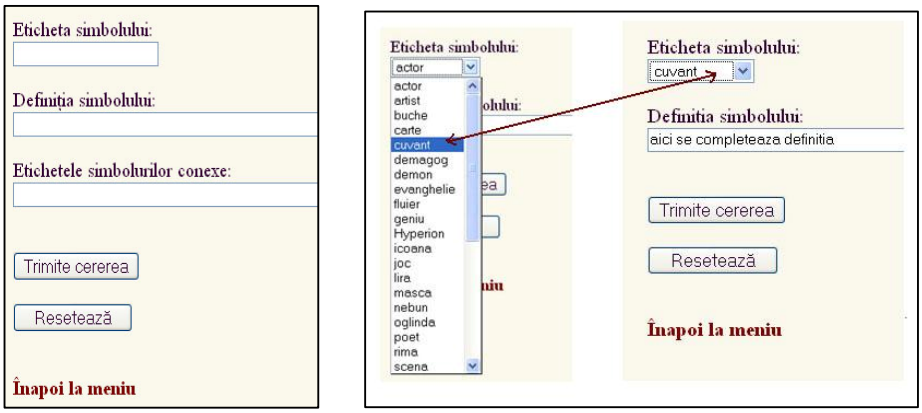


Figura 3: Adăugarea unui nou simbol în dicționar; completarea definițiilor pentru un simbol existent

După inserarea unui nou simbol în dicționar, administratorul trebuie să efectueze operația de parsare a poeziilor, pentru ca noul simbol să apară marcat şi în text. În cadrul acestei operații trebuie precizate formele flexionare ale noului simbol, pentru ca şi diversele apariții ale acestora în text să fie recunoscute şi marcate corespunzător. Unele studii ne arată că această operație poate fi efectuată de un modul de generare automată a formelor flexionare ale unui cuvânt. Administratorul mai poate face listarea simbolurilor existente, fiecare însoțit de identificatorul unic asociat în mod automat la includerea în dicționar.

3. *Tehnologiile utilizate în implementare*

Interfața utilizator a fost organizată şi formatată utilizând XHTML şi CSS2. Flexibilitatea soluției noastre este conferită de utilizarea XML pentru reprezentarea şi stocarea datelor folosite, precum şi de utilizarea implementării PHP a DOM (*Document Object Model*) pentru prelucrarea informațiilor din documentele XML.

```
− <poezii>
 − <poezie nume="adio">
    <titlu>Adio</titlu>
   + <strofa></strofa>
   − <strofa>
      <vers>De astazi dar tu fa ce vrei,</vers>
      <vers>De astazi nu-mi mai pasa</vers>
     − <vers>
        Ca cea mai dulce-ntre
        <simb id="femeie">femei</simb>
      </vers>
      <vers>Ma lasa.</vers>
    </strofa>
   + <strofa></strofa>
```

Poeziile sunt stocate în fişierul *poezii.xml*, a cărui structură este descrisă în imaginea alăturată. Acest format ne permite evitarea informației redundante, proprietate mai greu de obținut, de exemplu, în cazul memorării poeziilor într-o bază de date, în care s-ar aloca o întreagă înregistrare pentru un singur vers sau chiar un singur cuvânt, trebuind să fie precizat - la fiecare înregistrare - codul poeziei, al volumului, al autorului etc.

În cadrul fiecărei poezii sunt adnotate cuvintele care constituie intrări în dicționar. Dicționarul în sine este reprezentat tot în format XML, fiecare simbol având un identificator unic generat automat la crearea acestuia. De asemenea, fiecare simbol are indicate mai multe simboluri conexe, prin intermediul identificatorilor acestora.

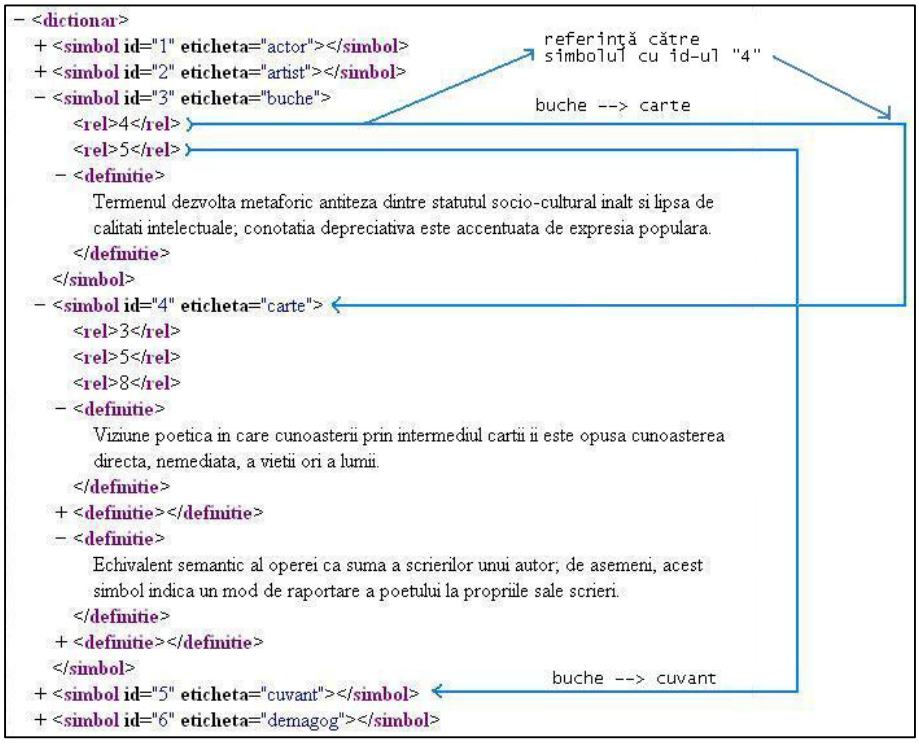


Figura 4: informațiile în format XML din dicționarul de simboluri

4. Concluzii şi direcții viitoare

Prima extindere a aplicației ar trebui să aibă în vedere integrarea simbolurilor din dicționar într-o ontologie de simboluri, oarecum complementară ontologiilor de concepte generale de tip *WordNet*, *SUO* sau *SUMO*. Acest lucru este viabil deoarece fiecare simbol are o serie de accepțiuni, înregistrate de dicționarele de simboluri, evident diferite de accepțiunea comună a cuvântului care desemnează respectivul simbol. Pe acest schelet ontologic pot fi grefate accepțiunile particulare dobândite de fiecare simbol în contextul unei opere sau a unui poem, şi pot fi stabilite de asemenea diverse relații (de exemplu, de apartenență la un câmp semantic comun) între două sau mai multe simboluri.

Deci, urmând ideile din (Niles & Pease, 2003) şi (Huang, 2004), ar putea fi făcută trecerea de la un simplu text la un lexicon (de exemplu, în genul unui dicționar de simboluri obişnuit) şi apoi la o ontologie dacă s-ar ține cont de delimitarea între:

- *Ontologie generală*: ontologie de nivel înalt partajată de toate domeniile (precum SUMO şi Wordnet)
- *Ontologie specifică*: pentru un domeniu, perioadă istorică, autor, eventual câmp semantic etc.

Un dicționar general de simboluri ar putea constitui o ontologie generală, iar adaptările şi modificările de semnificații ale simbolurilor în contextul unei opere paticulare, sau a

unui scriitor particular ori a unui curent literar particular, ar putea fi structurate într-o ontologie specifică.

Aplicația *Dicționar de simboluri on-line* se dorește a fi extinsă și prin adăugarea spre parsare a altor opere literare, aparținând unor autori diferiți, sau chiar a fragmentelor diverse, aflate în afara creației artistice. Dicționarul poate fi dezvoltat într-o altă direcție prin asocierea și integrarea de simboluri grafice alături de simbolurile textuale actuale. Pentru utilizatori ar putea fi utilă și gestionarea unui istoric al comentariilor, în care să fie marcate simbolurile și fragmentele de plecare pentru fiecare comentariu, efectuându-se astfel cu ușurință corelarea simbolurilor/fragmentelor cu adnotările.

Deși reprezintă doar un început, aplicația *Dicționar de simboluri on-line* oferă în acest moment un cadru general de lucru pentru exegeții operei eminesciene. În perspectiva oferită de diversele direcții de dezvoltare, putem vorbi despre această aplicație ca despre o unealtă care vine în sprijinul specialiștilor din domeniul literaturii și chiar al lingvisticii.

Referințe bibliografice

Abdoullaev, A.Sh. (2006). *Ontology, Semantic Technology, and Knowledge Society: World Wide Intelligent Web*, http://www.eis.com.cy/

Chevalier, A., Gheerbrant, A. (1998). *Dicționar de simboluri*, Ed. Artemis.

Huang, Chu Ren (2004). Text-based Construction and Comparison of Domain Ontology: A Study Based on Classical Poetry, *Proceedings of the 18th Pacific ASIA Conference on Language, Information and Computation*.

Niles, I & Pease A. (2001). Towards A Standard Upper Ontology. In *Proceedings of FOIS 2001*, October 17-19, Ogunquit, Maine, USA.

Niles, I & Pease A. (2003). Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology, *Proceedings of the IEEE International Conference on Information and Knowledge Engineering*, 412-416.

*** *SUMO* (*Suggested Upper Merged Ontology*) http://ontology.teknowledge.com/

*** *SUO - Standard Upper Ontology*, The IEEE Standard Ontology: http://suo.ieee.org

*** *Wordnet. A lexical database for the English language*: http://wordnet.princeton.edu/

DREPTUL DE PUBLICARE PE WEB

NOEMI BOMHER

Facultatea de Litere, Universitatea "Al.I.Cuza" Iaşi

noemibom@yahoo.com

Rezumat

În contextul societății informaționale se observă, chiar la o sumară cercetare a site-urilor româneşti de pe Internet, prezența extrem de redusă a resurselor on-line de literatură română, privite din perspectiva istoriei literarare, a teoriei literaturii, a teoriilor predării-învățării de înalt nivel ştiințific etc. Situația se impune a fi semnalată în vederea găsirii soluțiilor de ameliorare, în limitele legii dreptului de autor în vigoare, printr-o strategie stabilită, la nivel instituțional, de către cei în măsură să o facă.

1. Introducere

Problema dreptului de difuzare a informației prin mijloace electronice nu este nouă. Ea a fost dezbătută, la nivel european, iar liniile generale ale unei legislații specifice au fost stabilite (Huét, Mails, 1989).

Chestiunea pe care dorim să o punem, în cele ce urmează, nu se referă neapărat la restricțiile legislative impuse de legea dreptului de autor şi implicațiile acestora în difuzarea *on-line* a textelor literare româneşti, ci la faptul că, în limitele acestei legi, prezența autorilor studiați de istoria literaturii române, la nivel universitar, este nepermis de săracă.

În condițiile de învățământ actuale, în care numărul studenților de la facultățile umaniste, în speță de la facultățile de litere, este în creştere, posibilitatea de documentare oferită de bibliotecile tradiționale este insuficientă. Chiar la nivel preuniversitar, accesarea resurselor electronice de către tinerii învățăcei este tot mai frecventă şi este cunoscut faptul că, urmare a cererii mari, s-au creat site-uri de tipul „www.referate on-line.ro" unde sunt postate lucrări, cu autori cvasianonimi, redactate după criterii diferite, în domenii diferite, fără însă a avea girul unui specialist.

Un program educativ, dirijat la nivel național de un for academic, ar trebui să aibă în vedere realizarea unor lucrări-model, de referință atât pentru profesori cât şi pentru elevi, cu prevederea explicită a interdicției reproducerii / însuşirii materialelor în nume personal.

2. Educația în societatea informațională

2.1. Societatea informațională şi schimbările impuse de aceasta

Ideea coordonării instituționalizate a promovării literaturii române prin mijloace electronice se impune cu atât mai mult cu cât există în derulare, la nivel guvernamental, un amplu program de informatizare a instituțiilor şi a domeniilor de desfăşurare a activităților economice şi ştiințifice. Or, aceste direcții de implementare a informatizării instrumentelor şi resurselor ar trebui gândite în ansamblu, în funcție de implicațiile sociale ale fenomenului.

Societatea informațională nu este o societate nouă, ruptă de societatea existentă. Ea reprezintă un stadiu evoluat al acesteia din urmă. Ceea ce nu trebuie să se uite este faptul că fundamentele stabilității sociale rămân aceleaşi: democrație, civilitate, solidaritate, egalitate, muncă. Doar forma de expresie, modul de exercitare a acestor drepturi şi intensitatea diferitelor aspecte ale vieții democratice pot fi altfel decât cele de până în prezent, dar nu contradictorii. Este adevărat că internetul schimbă tradițiile şi obiceiurile de a gândi şi de a acționa, cere repunerea în discuție a multor probleme, impune renaşterea cunoştințelor şi a practicilor. Există instituții pe care noua conjunctură comunicațională le îngrijorează: anumite administrații statale se tem de tulburări; poliția şi justiția se simt depăşite de proceduri legale care îi împiedică să urmărească eficace criminalii care se joacă cu frontierele şi cu distanțele doar printr-un *click* de *mouse*…, instituțiile fiscale doresc să impună tranzacții care să restrângă dezvoltarea comerțului electronic etc., etc.

În mod firesc, dreptul de comunicare la distanță interactivă ar trebui să ofere răspunsuri satisfăcătoare unor interese principial opuse: libertatea comerțului şi dreptul consumatorului; dreptul de autor şi acela de utilizator; libertatea expresiei şi libertatea de a limita această libertate; obligația de a asigura securitatea persoanelor, a bunurilor şi dreptul la viața privată şi deci la protejarea / criptarea mesajelor.

Dinamismul societății contemporane creşte ca rezultat al dezvoltării tehnologiilor informatice şi de telecomunicație. Legăturile devin mai strânse, schimbările sunt rapide şi inovațiile apar succesiv, cu o mare frecvență. Această dinamică provoacă o stare de stupefacție sistemului legislativ, pentru că, tradițional, legea nu este un instrument care să se poată adapta suficient de repede, „din mers", rigorilor proceselor dinamice prin care trecem şi schimbărilor tehnologice rapide. Dezvoltarea care afectează cel mai puternic fundațiile legii o reprezintă evanescența sau dispariția frontierelor teritoriale, procesele de dematerializare şi declinul intervențiilor umane nemijlocite.

Principiile fundamentale care asigură o protecție legală cetățenilor rămân importante şi un obiectiv capital constă în asigurarea că aceste principii sunt respectate în noua societate. Trebuie să subliniem faptul că noile tehnologii nu sunt un scop social în sine şi că legislatorii nu trebuie să fie biruiți de această nouă tehnologie.

O societate în care distanțele şi granițele dispar, presupune modificarea viziunii asupra multor domenii: lucrul „la distanță" schimbă relațiile între lucrători, între lucrător şi angajator; informațiile publicate pe net de diversele ministere modifică relațiile între

guvern şi cetăţeni; comerţul *on-line* creează alte relaţii între producători şi consumatori; serviciile bancare prin internet facilitează relaţia client – bancă ş.a.m.d.

2.2. Educaţia – domeniu fundamental al oricărei societăţi – în contextul informatizării

Unul dintre cele mai însemnate domenii la care societatea informaţională actuală trebuie să mediteze atent este acela educativ.

Învăţământul la distanţă (IDD) propus de instituţiile de tip universitar este rezultatul modificărilor impuse de noul sistem de comunicare.

Pornind de la materialele documentare realizate pentru această formă de învăţământ, de la situaţia de criză a fondului de carte pentru bibliotecile tradiţionale, de la nevoia tot mai mare de facilitare a accesului la informaţie, la sursele textuale propriu-zise şi bazându-ne pe o experienţă de peste treizeci de ani în domeniul predării istoriei şi a teoriei literare în cadrul Facultăţii de Litere a Universităţii „Alexandru Ioan Cuza", credem că trebuie susţinută necesitatea creării unui departament guvernamental care să coordoneze, prin intermediul Academiei şi al instituţiilor de învăţământ superior, sistemul informaţional de pe internet, adecvând la noul sistem legile privitoare la învăţământ, legile privitoare la publicarea operelor literare pe web, la modul de promovare a acestora, cu o concentrare specială asupra sistemelor de legi referitoare la traducerea / publicarea textelor literare pe internet, percum şi la răspândirea acestora.

În domeniul literaturii, se impune crearea unor programe de nivel naţional care să finanţeze proiecte ce ar trebui să aiba ca finalitate publicarea operelor fundamentale ale literaturii române (de la primele scrieri literare româneşti până la literatura interbelică) în ediţii critice definitive (în parte existente, în bună parte urmând a fi realizate de specialişti filologi), instrumente indispensabile cercetării literare fundamentale.

Desigur, acest demers nu trebuie să impieteze legislaţia în vigoare şi este necesară menţinerea drepturilor privitoare la taxa de timbru ş.c.l., astfel încât instituţii tradiţionale, precum Uniunea Scriitorilor din România, să nu aibă de suferit. În orice caz, sistemul informatic permite găsirea unor soluţii de acces condiţionat la textele publicate pe net.

Paşi în această direcţie au fost făcuţi, ca urmare a unor iniţiative izolate ale unor entuziaşti cercetători ori ale unor edituri. Ne referim aici la editarea integrală, în format electronic, a *Operelor* lui Mihai Eminescu şi ale celor semnate de I. L. Caragiale. Dar aceste demersuri sunt izolate şi accesibile în mod restrictiv doar celor care îşi permit achiziţionarea respectivelor CD-uri. Or, se ştie, specialiştii în literaturi clasice sunt favorizaţi din acest punct de vedere. Ediţii adnotate la cuvânt, corpusuri paralele de texte latine şi greceşti sunt la îndemâna utilizatorilor de net, fără să fie impuse nici un fel de restricţionări. Iar acesta este doar un exemplu. Marile biblioteci europene au început crearea unor imense baze de texte în vederea conservării fondului de carte veche şi rară, dar şi a unor ediţii epuizate ale operelor marilor autori ai literaturii lumii. Fiecare dintre aceste instituţii au stabilit modalităţi proprii de facilitare a relaţiei cititor – carte, în

funcţie de legislaţiile naţionale şi de cele europene, în funcţie, mai ales, de politicile specifice de promovare a propriilor culturi şi limbi.

Publicarea textelor literare pe net ar facilita, în bună măsură, crearea unor proiecte de traducere a acestora în limbile de circulaţie internaţională mai uşor realizabile şi profitabile din multe puncte de vedere. O bibliotecă electronică a literaturii române – în original şi tradusă – uşor de consultat, eliminând distanţele şi rigorile impuse de contractul cititor-bibliotecă obişnuită, ar face mai vie relaţia carte-receptor, indiferent de graniţele fizice şi cele subiective.

O cultură „minoră", ca să citãm epitetul folosit de Lucian Blaga, precum este a noastră nu poate avea decât de câştigat de pe urmă utilizării constructive a noilor mijloace de conservare şi de promovare a limbii şi literaturii.

Pe lângă publicarea textelor literare propriu-zise – a celor mai vechi de 25 de ani şi a celor contemporane, cu îngăduinţa autorilor interesaţi de această formă de promovare a propriei creaţii – este necesară şi publicarea instrumentelor necesare predării / învăţării. Ne referim aici, pe de o parte, la dicţionarele de specialitate, şi, pe de altă parte, la cursuri şi la volume de istorie, critică literară, teorie şi poetică, la manualele (atât de multe!) alternative, compendii şi antologii comentate ş.a.

3. Închinare seniorilor

Un câştig îl constituie deocamdată publicarea, şi în format electronic, a unor periodice de cultură consacrate (d. ex. „România literară", „Dilema veche", „22", de la Bucureşti, „Timpul" de la Iaşi) într-un proces ce nu dezechilibrează economic aceste reviste, susţinute şi de Ministerul Culturii şi Cultelor. Dar, în orice ţară civilizată, cultura este finanţată şi promovată de instituţii guvernamentale naţionale. Sunt de semnalat, de asemenea, publicaţiile cu orientare literară, ori cenacluri literare româneşti (d. ex. www.clubliterar.com; www.agero-stuttgart.de; www.onlinegallery.ro) care există numai în spaţiul virtual al calculatorului, indiferent de notaţia finală de după punct (.com, .de, .ro), fără ca acest statut „imaterial" să determine vreo scădere a frecventării şi, deci, a cunoaşterii lor.

Ne înclinăm dinaintea seniorilor ce pot decide soarta literaturii române în societatea informaţională rugându-i să ia aminte, *sine ira et studio*, la mersul vremurilor şi la destinul propriei noastre culturi şi să-şi amintească faptul că portretul nostru în ansamblul colajului global singuri ni-l facem, singuri îl putem expune, dacă vrem ca el să ne reprezinte şi să nu arate asemenea unei caricaturi glumeţe, în cel mai blând dintre cazuri.

Referinţe bibliografice

Huét, J., Mails, H. (1989). *Dreptul informaticii şi telecomunicaţiilor. Situaţia întrebărilor. Texte şi jurisprudenţă, studii şi comentarii*, Paris, Litec, 1011 p.

MODELARE CU ONTOLOGII ŞI ADNOTĂRI

RADU CIBOTARU

Facultatea de Informatică, Universitatea "Al.I.Cuza", Iaşi

deceneu@info.uaic.ro

Rezumat

Lucrarea prezintă un studiu de caz privind crearea, popularea şi utilizarea unei ontologii. Studiul se rezumă la un domeniu restrâns şi anume, la camere digitale, descrie structura ierahică de clase şi relaţiile dintre proprietăţi ale ontologiei. Prezintă o modalitate de populare automată a ontologiei, bazata pe şabloane. Face un studiu comparativ a metodei descrise cu o altă metodă bazată pe adnotări. Lucrarea este concentrată asupra modului de organizare şi structurare a informaţiilor.

1. Introducere

Diverse studii au demonstrat necesitatea utilizărilor sistemelor de adnotare în modelarea relaţiilor semantice. Sistemele de adnotare s-au dovedit utile deoarece un document adnotat a fost stocat şi utilizat ulterior de acelaşi sistem sauun altul, fapt ce a permis reutilizarea şi extinderea sistemului. Un astfel de sistem este Gate (Cunningham et al., 2006). Problema principală ce apare la comunicarea între două aplicaţii diferite prin intermediul documentelor adnotate este compatibilitatea adnotărilor. În general un document adnotat are o structură XML, auto-descriptivă. Datorită acestei flexibilităţi fiecare sistem îşi organizează datele într-un format propriu. De aici apare şi problema incompatibilităţii adnotărilor. Petru ca două aplicaţii, ce utilizează sisteme de adnotare diferite, să poată comunica este necesară translatarea unui document dintr-un sistem de adnotare în celălalt sau în altul intermediar. Această problemă se accentuează când adnotările marchează informaţii semantice prin intermediul referinţelor.

Odată ce sistemul creşte în dimensiune şi trebuie să răspundă la mai multe cerinţe este necesară o organizare riguroasă. Cu ajutorul unei structuri ierarhice se pot defini concepte, entităţi şi relaţii între entităţi (Perez et al., 2005). Această soluţie rezolvă unele probleme, dar introduce altele noi:

- se confruntă în continuare cu problema translatării adnotărilor;
- necesită studiul şi înţelegerea sistemului creat;
- devine tot mai complicată şi dificil de utilizat, pe măsură ce se introduc concepte noi.

Comunicarea cu documente adnotate seamănă cu serviciile web, ce se confruntă cu aceeaşi problemă: a incompatibilităţii celor doi participanţi la comunicare. În cazul serviciilor web se foloseşte o soluţie orientată obiect. S-au creat librării ce convertesc documentele XML în obiecte specifice limbajului de programare[1]. Astfel aplicaţiile au

[1] http://java.sun.com/webservices

posibilitatea să lucreze obiectual făcându-se abstracție de formatul mesajului din comunicare. Această tehnică are mai multe avantaje:

- introduce un standard pentru formatul mesajelor utilizate în comunicare, rezolvând problema translatării;
- se lucrează la nivel de limbaj de programare fără a introduce concepte noi;
- nivelul de dificultate se limitează la cel al limbajului utilizat;

Pentru a face față noilor cerințe impuse de aplicații pentru fişierele de configurare în format XML[2], s-a impus standard-ul OWL[3]. Fără a face o introducere în OWL, este de remarcat faptul că se apropie tot mai mult de conceptul „orientat obiect", prin toate aspectele sale: clasa, moştenire, polimorfism. Prin faptul că se introduce conceptul de clasă şi instanță, devine posibilă separarea informației de structură. Iar prin apropierea de conceptele orientate-obiect, s-a facilitat crearea de librării pentru procesarea documentelor OWL în diferite limbaje de programare. O implementare reuşită este oferită gratis de HP, numele proiectului este – Jena. Utilizarea şi exemplificarea conceptului orientat-obiect, reprezintă scopul de bază al lucrării. Pe lângă avantajele enunțate la serviciile web, la OWL remarcăm:

- spațiile de nume, ce permit re-utilizarea în alte contexte;
- sistem de inferente ce permite interogarea în mod standard a ontologiei;

2. Arhitectura şi funcționalitatea sistemului

Aplicația „Domain Ontology" a fost creată pentru a exemplifica practic utilizarea ontologiilor în locul adnotărilor. Sursa datelor pentru popularea ontologiei e un număr restrâns de situri ale producătorilor de camere digitale (http://www.kodak.com, http://www.canon.co.uk). Datele sunt restrânse numai la camere digitale, pentru a uşura analiza şi a permite compararea rezultatelor, proiectul a fost divizat în două părți:

- crearea structurii ontologiei (s-a utilizat Protege (Horridge et al., 2004));
- popularea automată a ontologiei cu informațiile extrase de pe situri web;

Fiecare parte conține anumite etape de dezvoltare.

Partea I:

- analiza datelor tehnice legate de camere digitale;
- organizarea şi gruparea datelor tehnice în clase şi subclase;
- crearea ierarhiei de clase şi a proprietăților claselor;
- detalierea claselor de bază cu subclase particulare datelor tehnice;
- crearea de relații între clase şi ierarhii de clase;
- popularea manuală a unui şablon aferent unei pagini web cu detalii tehnice la o singură cameră, pentru a permite analiza rezultatelor;
- repetarea procesului de populare a câte unui şablon pentru fiecare producător;

Partea II:

- crearea unui procesor general pentru o pagină web;
- crearea extensiilor la procesorul general, pentru fiecare site în parte;
- crearea unui crawler general pentru un singur site;

[2] http://www.w3.org/XML
[3] http://www.w3.org/2004/OWL

- crearea extensiilor la crawler-ul general pentru fiecare site în parte;
- crearea punții de comunicare între crawler şi procesor;
- crearea punții de comunicare între procesor şi modulul de populare a ontologiei;
- generarea modelului obiectual pentru ontologia creată;
- implementarea funcției abstracte „createIndividual" din clasa abstractă de bază a modelului generat, pentru a conecta acest model generat cu cel de populare.

3. Descrierea sistemului

Din analiza specificațiilor proiectului se distinge uşor modularizarea şi elementele de comunicare între module.

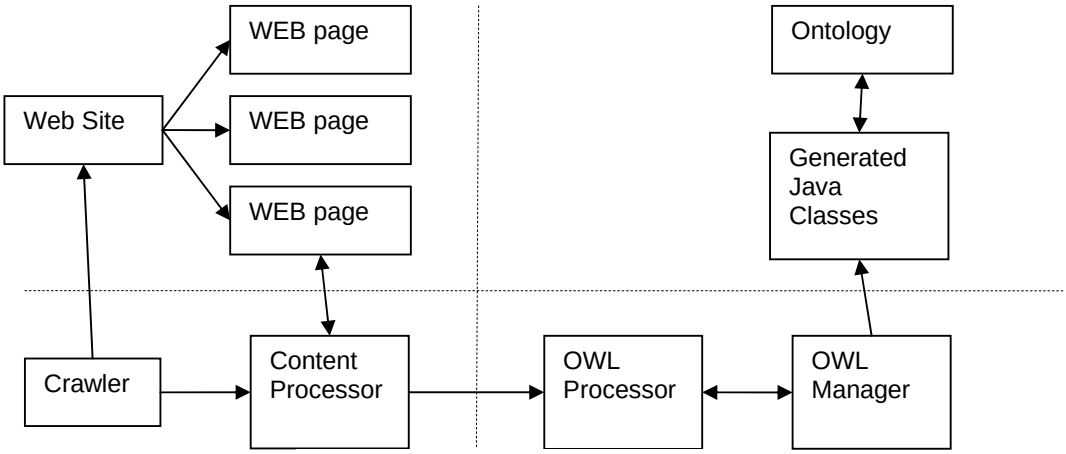


Figura 1: Arhitectura aplicației

Scopul principal impus acestui proiect a fost re-utilizarea şi posibilitatea extinderii ulterioare. Arhitectura generală a aplicației este prezentată în figura 1. S-a luat în calcul organizarea tip plug-in, încât să fie posibilă adăugarea unui nou procesor pentru un alt site fără a modifica funcționalitatea celorlalte module. Iar scoaterea unui modul specific unui site să nu afecteze funcționalitatea aplicației. Clasele aferente funcționalității generale sunt scrise astfel încât conțin operații aplicabile oricărei surse de date (site producător). Astfel, se reduce numărul operațiilor efectuate de către clasele particulare unui singur producător şi totodată se micşorează efortul dezvoltării unui nou modul (pentru alt producător). Clasele particulare unui producător, efectuează operații specifice acelui producător. În figura 1 se observă fiecare nivel; situl producătorului şi modulele de cautare şi extragere a conținutului, ontologia şi modulele de populare automată a ontologiei. Avantajul major adus de aplicație e faptul că modulul de populare rămâne acelaşi indiferent de modulul utillizat pentru procesarea conținului web. Acest lucru a fost posibil prin introducerea unui nivel abstract pentru procesarea conținutului web ce nu intră în detalii aferente sitului web.

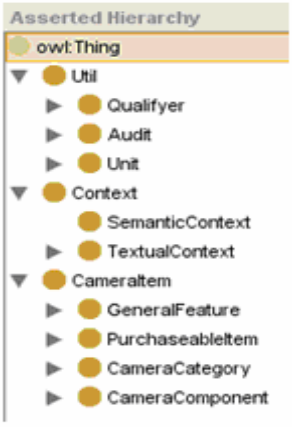


Figura 2: Clasele de bază ale ontologiei

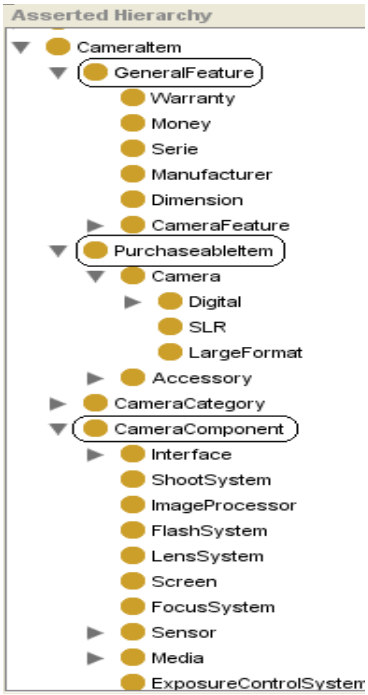


Figura 3 : Detalierea claselor de bază cu clase particulare

Ontologia a fost creată pentru a acoperi toate aspectele tehnice ale unei camere digitale. Datele tehnice au fost grupate în categorii generale (figura 2) şi detaliate cu altele particulare (figura 3). Toate aspectele tehnice au o singură clasă de bază: „*CameraItem*" (figura 2). Acestă clasă a fost creată pentru a permite relaţionarea diferitor aspecte tehnice între ele, precum şi compunerea unui aspect tehnic din mai multe detalii (figura 3).

Clasa „*TextualContext*" (figura 2), subclasă a clasei „*Context*" este folosită pentru a stoca şabloane pentru fiecare producător în parte. Mecanismul e simplu: pentru fiecare producător se alege un produs şi pentru acel produs se crează manual un şablon. Şablonul conţine instante ale claselor din ontologie ce sunt folosite în procesul de populare automată a ontologiei. Avantajul utilizării acestui sistem este faptul că se crează un singur şablon pentru un singur produs şi pe baza acestui şablon se procesează întreg situl producătorului.

Fiecare instanţă a clase „*TextualContext*" stochează câte un şablon pentru un singur site al unui producător de camere digitale. Iar proprietatea „*cameraItemContext*" relaţionează şablonul cu instanţele ce conţin informaţiile tehnice din ontologie. Fiecare instanţă, a unei informaţii tehnice, este marcată cu o proprietate „label" utilizată în momentul populării ontologiei pentru a identifica tipul informaţiei tehnice conţinută în pagina web. Pe baza acestei proprietăţi se face legătura între conţinutul paginii web şi şablon, pentru a identifica tipul informaţiei. Iar şablonul este utilizat în continuare pentru a identifica tipul clasei pentru care se va crea un individual, ce va stoca informaţia din pagina web. Aici se poate observa un aspect important al ontologiilor, şi anume, de a modela atât datele cât şi structura lor, precum şi relaţiile dintre ele.

4. Rularea sistemului

Procesul de rulare se desfăşoară în mai multe etape:

- configurarea directoarelor de resurse şi a ontologiei cu şabloane;
- crearea şi popularea manuală a unui şablon la o pagină web pentru un singur produs de la un producător de camere digitale şi repetarea operaţiei pentru fiecare alt producător;
- lansarea în execuţie a aplicaţiei;

În procesul de completare a şabloanelor se crează câte o instanţă a clasei „*TextualContext*" pentru fiecare producător. Pentru fiecare site se selectează câte un produs pentru popularea şablonului. Pentru fiecare informaţie tehnică se crează câte o instanţă a unei clase din ontologie. Tipul instanţei este ales de utilizator pe baza categoriei din care face parte informaţia tehnică, această categorie este folosită pentru a popula proprietatea „*label*" a instanţei create, iar informaţia propriu zisă este folosită

pentru a completa proprietăţile instanţei. Instanţa nou creată şi populată cu informaţii este ataşată proprietăţii „*hasTextualContext"* pentru a face legătura cu şablonul.

Aplicaţia încarcă întâi ontologia cu clase, iar la pasul doi încarcă ontologia cu şabloane. Lista de şabloane este inspectată pentru a instanţia crawlerul (figura 1) specific sitului producătorului. Acest crawler are singura funcţie de a naviga pe site şi a extrage paginile cu toate produsele oferite de producător, a le corecta şi a le transforma întru-un document XML valid. Paginile sunt transmise unui procesor (figura 1) particular producătorului ce are doar funcţia de a extrage informaţiile din pagină şi a le transmite într-un format unic, procesorului pentru ontologie (figura 1). Acest ultim procesor are rolul de a popula ontologia cu instanţe ale claselor din ontologie. În procesul de populare se foloseşte şablonul pentru a identifica tipul instanţei ce trebuie creată, iar crearea relaţiilor între individuali se face apelând metodele prin reflexie. În cazul în care relaţionarea trebuie făcută folosind o proprietate a unei clase de bază, se foloseşte inspectarea claselor de baza pentru a identifica nivelul la care se poate crea legătura.

5. Discuţii şi evaluări

Primă remarcă privitoare la rezultatele populării ontologiei este faptul că s-a urmărit crearea corectă de individuali aferenţi datelor tehnice conţinute în paginile web, nu şi a caracteristicilor unei date tehnice. Corectitudinea populării ierarhiei „CameraItem" este dată de fidelitatea corespondenţei template - site web producător. Problemele ce au apărut au fost legate de şabloane. Şablonul a fost creat pentru un singur produs, iar acel produs nu conţine neapărat toate caractersticile tehnice ce le poate avea un produs de la acel producător. Astfel pentru un anumit produs pot apare caracteristici tehnice ce nu sunt cuprinse în şablon. Soluţia ar putea fi crearea unei aplicaţii ce ar examina conţinutul sitului şi l-ar compara cu cel din şablon. În procesul de evaluare a rezultatelor s-a utilizat ontologia populată automat cu instanţe. Din această ontologie s-au analizat instanţele create automat pentru produsul ce a fost folosit în şablon. S-a constatat o precizie la populare de 100% pentru acel produs, însă această precizie nu s-a menţinut pentru toate produsele. În funcţie de producător s-a constat o precizie de peste 90% pentru produse din aceeaşi categorie şi o precizie sub 50% pentru produse din altă clasă. Acest procentaj se poate îmbunătăţi considerabil prin popularea şablonului cu informaţii ale produselor din categorii diferite, modificând doar datele de intrare. O altă problemă deosebit de gravă este faptul că siturile producătorilor se modifică în timp. Pentru rezolvarea acestei probleme, s-a recurs la utilizarea unui fişier de configurare în care sunt păstrate informaţiile sensibile aplicaţiei. Pentru extragerea informaţiei dintr-o pagină s-a folosit Xpath. Un aspect interesant este compararea ontologiei create cu o altă ontologie[4] accesibilă pe web. Primul lucru ce poate fi observat este faptul că această ontologie a fost o sursă de inspiraţie. Însă ontologia nu făcea diferenţă între categorii de obiecte, ci doar le relaţiona. Ontologia nou creată organizează mai bine datele tehnice în categorii, detaliază fiecare aspect tehnic şi introduce ierarhii de proprietăţi pentru a specifica relaţiile între instanţe ale ontologiei. Ultimul aspect important este compararea ontologiei cu sistemul de adnotări. Pentru a pune în lumină toate diferenţele vom recurge la un exemplu foarte simplu extras din ontologia populată automat şi se va compara cu o

[4] http://protege.cim3.net/file/pub/ontologies/camera/camera.owl

soluție posibilă bazată pe adnotărilor.Cea mai mare problemă cu care se confruntă o soluție bazată pe adnotări este necesitatea unui editor vizual care să genereze o astfel de structură. Mai important este că, în momentul în care apare necesitatea modificării structurii documentului xml, este necesară modificarea aplicației vizuale. Deşi, la ora actuală există adnotatoare foarte bune, ele sunt capabile să realizeze marcări elementare. Iar problema cel mai des întâlnită apare la crearea relațiilor între etichetele ce marchează textul. În urma analizei se constată că toate problemele discutate sunt generate de structura documentelor xml, cu care se lucrează, precum şi de incompatiblitățile ce apar între ele. Înlocuirea formatului xml cu OWL aduce avantajul separării conținutului de structură, iar utilizarea unui editor vizual de documente OWL ar face posibilă editarea oricărui document indiferent de structura sa internă.

6. Concluzii

Lucrarea de față prezintă un caz practic de utilizare a ontologiilor şi face o comparație cu un sistem de adnotare aplicat aceleiaşi situații. S-au luat în discuție probleme actuale cu care se confruntă aplicațiile ce folosec documente adnotate, utilizate în procesarea limbajului natural. Pentru fiecare problemă s-au propus soluții utilizate în aplicații ce nu țin de domeniul lingvisticii computaționale şi anume, domeniul comercial, care e cel mai sensibil la probleme minore. Soluțiile propuse au fost prezentate în ordinea apariției lor, sau în ordinea în care s-au impus pe piață, cu avantajele sau dificultățile aduse. S-a descris arhitectura aplicației, fără a da detalii legate de implementare. S-a dat o atenție deosebită structurii ontologiei, precum şi a modului de organizare a conceptelor, fapt foarte important în dezvoltarea unei ontologii. S-a constatat că utilizarea ontologiilor duce la eliminarea multor probleme legate de modul de structurare a unui document XML, dar introduce o problemă importantă: nu vor mai putea fi folosite vechile instrumente dezvoltate. Această problemă poate fi uşor rezolvată prin crearea unei punți pentru conversia din xml în owl şi invers. Acest proiect ar putea fi punctul de plecare pentru noile cercetări în domeniul lingvisticii computaționale.

Referințe bibliografice

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., Ursu, C., Dimitrov, M., Aswani, M.N., Roberts, I. (2006) University of Sheffield, *Developing Language Processing Components with GATE Version 4 (a User Guide)*

Perez, D., Postolache, O., Alfonseca, E., Cristea, D. and Rodriguez, P. (2005): *Hierarchical XML Layers Representation for Heavily Annotated Corpora* In Proceedings of the RANLP-2005 Conference, Borovets, Bulgaria, 21-23 September 2005, pp. 380-386.

Horridge, M., Knublauch, H., Rector, A., Stevens, R., Wroe, C. (2004) A Practical Guide To Building OWL Ontologies Using The Protege-OWL Plugin and CO-ODE Tools.

CADRE PENTRU O IMPLEMENTARE PC-PATR A VERBELOR TRANZITIVE DIN LIMBA ROMÂNĂ

NADIA LUIZA HUȚULIAC

Institutul de Cercetări pentru Inteligență Artificială, Academia Română, Bucureşti

hnadia_luiza@hotmail.com

Rezumat

Ne propunem să prezentăm din perspectiva formalismului lingvistic **PATR** diferite situații gramaticale întâlnite în analiza verbelor tranzitive: structuri verbale bivalente şi trivalente; acordul trăsăturilor morfologice între complementul direct şi forma neaccentuată a pronumelui personal, care îl anticipează; introducerea informației semantice în structurile de trăsături asociate grupului verbal. Realizarea unor reguli complete de implementare folosind PC-PATR este o sarcină uriaşă, deoarece există foarte multe posibilități în limbajul real de care cercetătorul nu poate ține cont în a analiza fiecare propoziție. Pentru un grup de propoziții însă, această implementare se dovedeşte utilă deoarece arborii sintactici rezultați pot fi transformați ulterior în exemplele de traducere ale unei baze de traducere.

1. Introducere

Scopul acestei cercetări este de a furniza o implementare **PC-PATR**[1] pentru verbele tranzitive din limba română, ca punct de plecare pentru un proiect ulterior de extragere a cunoştințelor de traducere dintr-un corpus paralel, aliniat propozițional. Tehnologiile lingvistice sunt puse la dispoziție de către **SIL** International şi se caracterizează prin modularitate, deoarece instrumentul **CARLASTUDIO**, abreviere pentru *Computer-Assisted Related Language Adaptation*, foloseşte o gramatică generată de sistemul expert **PAWS** (*Parser and Writer for Syntax*) şi un lexicon creat cu ajutorul programului **ToolBox**.

2. Tehnologii lingvistice

CarlaStudio[2] **(CS)** este un program cu dublă funcționalitate: pe de o parte, permite modelarea unui limbaj anume de către lingvistul cercetător, iar pe de altă parte, implică modelul lingvistic creat în analiza textelor sau în adaptarea lor pentru un alt limbaj. Versiunea **2.9.0.4 Unicode** utilizată include trei programe importante pentru scopul cercetării noastre, fiecare cu o funcție specifică de procesare: analizorul morfologic **AMPLE**, analizorul sintactic **PC-PATR**, responsabil de dezambiguizarea

[1] http://www.sil.org/pcpatr/
[2] http://www.sil.org//

şi crearea arborilor sintactici, **JOINCOMP**, instrumentul ce recunoaşte cuvintele compuse sau locuțiunile morfologice.

Autorii sistemului **CS** (John Hatton, Andy Black, Bob Eaton, Marius Doornenbal) au gândit modelarea limbajului ca o gestionare sinergică a mai multor tipuri de informație lingvistică. Diacriticele din sistemul românesc sunt tratate la primul nivel de procesare, unde există posibilitatea de a descrie asociații de caractere dependente de limbă. Categoriile gramaticale aparțin nivelului de analiză şi sunt adaptate după recomandările sistemului expert **PAWS**, astfel încât co-există categorii tradiționale, de tipul : nume, verb, pronume, adverb, adjectiv, şi categorii adaptate- cuantificator, demonstrativ, auxiliar.

CS are la bază paradigma Analiză-Transfer-Sinteză. Analiza e focalizată pe morfologie şi fonologie. Din cauza ambiguității analizei cuvintelor independent de contextul lor, s-a urmărit modelarea construcțiilor sintactice prin folosirea ordinii permise a lexemelor în sintagme şi construcții gramaticale speciale, prin marcarea trăsăturilor de acord şi a celor de entități numite.

Una din principalele motivații legate de crearea sistemului expert **PAWS** a fost de a adăuga un instrument de dezambiguizare bazată pe sintaxă pentru instrumentele de analiză morfologică existente în **CS**. În scenariul propus de Andy Black şi Cheryl A. Black, utilizatorii folosesc un lexicon de morfeme pe care programul morfologic **AMPLE** le utilizează pentru a analiza cuvintele dintr-un text în constituenți. Rezultatul poate fi avansat apoi către **PC-PATR**, inclus în **CS**, împreună cu un fişier de gramatică.

Ultimul program utilizat în dezvoltarea implementării **PATR** a verbelor tranzitive este lexiconul **Toolbox**, responsabil de crearea unui fişier cu extensia **.lex**, apelat de **CarlaStudio** în analiza cuvintelor. Primul câmp (introdus prin \w) furnizează forma grafică a cuvântului, al doilea (\c) introduce categoria morfo-sintactică a intrării, al treilea (\g) este glosa, al patrulea (\f) conține trăsăturile de subcategorizare ale intrării lexicale. Ultimul câmp permite o evidență a actualizărilor datelor introduse de lingvist.

3. Morfologia verbului

Modurile *indicativ* şi *declarativ* sunt sinonime în contextul **PAWS**, dar opțiunea pentru lexemul *indicativ* s-a realizat prin scrierea tipului corespunzător:

```
(1)  Let indicative  be  <head infl mood indicative> = +
```

Modurile imperativ şi conjunctiv există în descrierea morfologică **PAWS**, iar condiționalul a fost introdus pentru limba română:

```
(2)  Let conditional be  <head infl mood conditional> = +
```

Pentru modul indicativ am adăugat timpurile imperfect şi mai-mult-ca-perfect, iar timpurile compuse au fost definite prin reguli şi restricții particulare:

```
(3)  Let imperfect   be  <head infl tense imperfect> = +
(4)  Let pluperfect  be<head infl tense pluperfect> = +
```

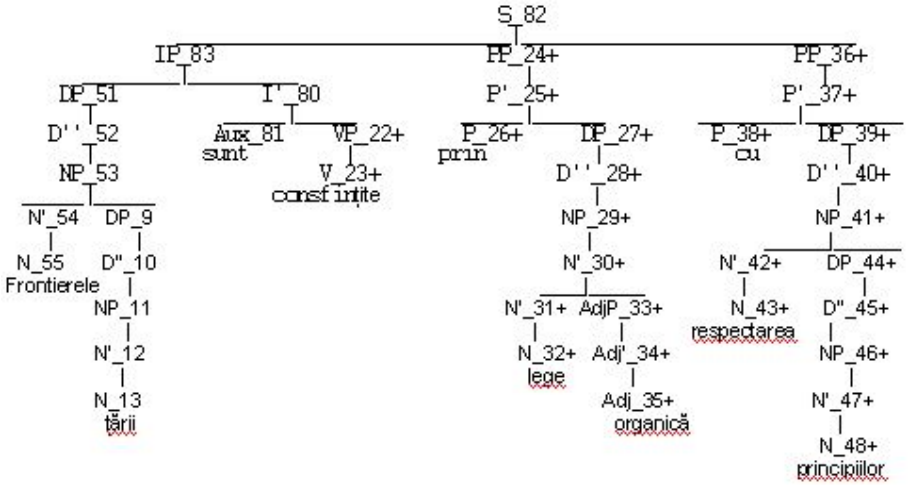
4. Structuri verbale de subcategorizare

4.1. Reguli sintagmatice de expansiune a grupului verbal

Regulile sintagmatice de expansiune a grupului verbal înlocuiesc simbolul unic al sintagmei verbale printr-unul sau mai multe simboluri, rezultând două reguli generale, diferențiate după valența verbului:

```
(A)   VP = V DP[3]
            <VP head> = <V head>
            <V head object> = <DP>
            <V head type transitive> = +
            <V head type copular> = -
            <V head type passive> = -
(B)   VP = V DP_1 DP_2
            <VP head> = <V head>
            <V head object> = <DP_1>
            <V head type ditransitive> = +
            <DP_1 head case> = accusative
            <DP_2 head case> = dative
```

În figura următoare, pot fi urmărite relațiile de dependență sintactică stabilite între nodurile fiică şi mamă din perspectiva regulilor de expansiune a grupului verbal, cu nucleu pasiv şi determinări prepoziționale:



Cazul special al anticipării obiectului direct prin formele pronominale neaccentuate este rezolvat prin introducerea a două trăsături gramaticale- **Case**, respectiv, **ObjAgr**- şi prin

[3] Ca terminologie, am păstrat structura mai extinsă **Determiner Phrase** (DP), motivul constituindu-l existența numeroaselor poziții pentru tipuri diferite de modificatori ai numelui.``

modificarea regulii de expansiune a grupului nominal. Deşi marcată prepoziţional, sintagma cu funcţia de obiect nu e considerată grup prepoziţional, ci nominal, unificându-se trăsăturile de acord şi de caz:

```
Let Case   be   <cat> = Case | for pe
                <head case> = accusative
                <head type proper> = +
Let ObjAgr  be  <cat> = ObjAgr
                 <head type proper> = + |for whole set
              of object agreement pronouns that go
              before the aux and/or verb
rule {IBar option3- ObjectAgr initial}
I' = ObjAgr  VP
     <I' head> = <VP head>
     <I' head type auxiliary> = +
     <I' head object head agr> = <ObjAgr head agr>
     <I' head object head type proper> = +
     <I' head type transitive> = +
     <I' head type prefix> = <ObjAgr head type prefix>
rule {DP option Case- prepositional marker for DO}
DP = Case DP_1
     <DP head> = <DP_1 head>
     <DP head type proper> = +
     <Case head case> = accusative
     <DP head case> = accusative
     <DP option> = Case
```

4.2. Roluri semantice

Rolurile semantice nu interesează în mod strict gramatica generată de **PAWS**, însă ele pot fi adăugate şabloanelor de descriere, cu observaţia că există o diferenţă între a indica rolul complinit de un verb printr-un obiect şi a predetermina rolurile unui substantiv într-o structură de subcategorizare.

Codificarea informaţiei semantice în structura de trăsături se realizează îndeosebi în lexicon, deoarece fiecare verb va avea un cadru de subcategorizare, iar argumentele trebuie să aibă asignate fiecare un rol semantic potrivit. Aceste trăsături introduse pentru intrările lexicale vor avea specificate tipuri care să le modifice în structuri de trăsături ce se vor unifica apoi corect prin regulile sintagmatice.

Iniţial, gramatica nu includea şi roluri semantice, iar descrierea tipului de tranzitivitate nu prezenta o structură specifică de subcategorizare. Pentru nominale, soluţia cea mai

satisfăcătoare este să se marcheze ce roluri nu pot fi îndeplinite, de exemplu- rolul de AGENT sau de EXPERIENCER pentru obiecte inanimate:

```
|templates for semantic roles on nominals
Let -AGENT   be  <head role AGENT> = -
                 <head type animate> = -
Let -EXP     be  <head role EXPERIENCER> = -
                 <head type animate> = -
```

Pentru o structură verbală bitranzitivă, perechea de roluri tematice AGENT- THEME este evidențiată prin următoarele declarații de descriere :

```
Let AGENT_THEME be <head subject head role AGENT> = +
                   <head object head role THEME> = +
                   <head indirectobject> = none
```

Cadrele de subcategorizare sunt foarte complexe. În general, schimbările la nivelul regulilor nu sunt recomandate pentru fişierul de gramatică, pentru că trăsăturile nou introduse vor unifica apoi cu alte trăsături, mărind numărul de analize şi de arbori de analiză şi, implicit, timpul de procesare sintactică.

5. Concluzie

Rezultatele recente din **NLP** datorează mult metodelor statistice sau celor bazate pe corpus. Traducerea automată nu face excepție de la această afirmație (Carl & Way, 2003), deoarece paradigmele traducerii bazate pe exemple şi cea statistică folosesc lexicoane şi reguli de traducere achiziționate din corpusul paralel şi compilate apoi în motorul de traducere. Conceput inițial ca o analiză PC-PATR a verbelor tranzitive, proiectul nostru va fi dezvoltat în continuare cu o dublă funcționalitate. Pe de o parte, intenționăm realizarea unor modele lingvistice pentru română şi engleză prin procesarea unui corpus paralel, aliniat propozițional. Pe de altă parte, vom utiliza modelele lingvistice PC-PATR pentru a obține cunoştințe de traducere, exprimate în unități lexicale sau sintagmatice echivalente, prin care să fie validată traducerea automată.

Referințe bibliografice

Black, C., (1997). A PC-PATR Implementation of GB Syntax. *SIL Electronic Working Papers 1997-0006*.

Carl, M., Way, A. (eds.) (2003). *Recent advances in Example-Based Machine Translation*. Kluwer Academic Publishers, Netherlands.

Index de autori