Resurse lingvistice româneşti şi tehnologii informatice aplicate limbii române

Dan Cristea♦, Dan Tufiş♣

♦ Conferenţiar doctor
Universitatea "Al. I. Cuza" Iaşi, Facultatea de Informatică
♣ Cercetător principal 1, membru corespondent al Academiei Române
Academia Română – Institutul de Inteligenţă Artificială

1. Un pericol al societăţii moderne

Ataşaţi, aşa cum suntem, ideii de globalizare şi deschidere spre Europa şi lume, nu putem însă neglija şi un pericol, datorat imixtiunii abuzive prin mediile electronice (televiziune, internet) a limbilor intens mediatizate asupra celorlalte. Ameninţate în acest mod sunt dialectele locale, deluvionate de tăvălugul lingvistic al limbilor oficiale, dar şi limbi ale unor naţiuni relativ mari dar, temporar, defavorizate economic, ce pot fi impurificate prin importuri abuzive în anumite registre tehnice sau segmente ale jargonului.

Într-un recent schimb de mesaje între membrii Comitetului Program al conferinţei filialei europene a Asociaţiei de Lingvistică Computaţională[1], ce urmează să fie găzduită anul viitor de Budapesta, Claire Gardent – preşedinta Comitetului Program propunea organizarea unui atelier de lucru cu titlul *Resurse şi instrumente lingvistice pentru limbile est-europene*. În încercarea de a face acest atelier de lucru de interes şi pentru participanţi ce vin din alte zone decât estul Europei, sintagma "limbi mai puţin studiate"[2] a putut doar cu greu fi evitată prin iterări succesive asupra titlului, în final ajungându-se la formularea "limbi de densitate mai mică"[3]. E lesne de înţeles de ce formulări de acest tip, aplicate unei arii lingvistice din care face parte şi româna sună atât de neplăcut pentru orice lingvist român, care cunoaşte îndelungata strădanie a Academiei Române pentru construirea de dicţionare, pentru definirea regulilor de gramatică românească şi pentru crearea tezaurelor lingvistice ale limbii noastre. Cu toate acestea, veninoasele sintagme sunt, din păcate, justificate, pentru că tipul de tratament pe care îl insinuează ele ca fiind insuficient este cel de natură informatică.

Societatea contemporană are acum mijloace pentru a împiedica dispariţia limbilor mici, pentru a încetini "dizolvarea" lor în limbile mari, ori pentru a stopa deteriorarea lor prin importuri abuzive. Limbile au în ele tot atâta creaţie divină ca şi speciile. O limbă este, într-un anumit fel, un "jurnal de bord" al unui popor, pentru că limba unui popor păstrează urme ale trecerii lui prin timp aşa cum o formă de relief memorează uneori, în straturile geologice, istoria unei specii. A pierde o limbă e la fel de dureros cu a pierde o specie. Dar, aşa cum selecţia naturală, cu sacrificiile ei, a întărit speciile, poate că procesul de "nivelare" a limbilor, pe care noi îl resimţim ca fiind o tragedie, trebuie de fapt privit cu detaşare. Pentru că, ar putea spune cineva, el nu reprezentă altceva decât un act de selecţie naturală cu care ne-a obişnuit evoluţia, actul numit "cel mai tare rezistă", exerciţiul eschimos de a izola în noaptea polară un bătrân neputincios şi nefolositor pentru ca cei tineri şi utili să aibă o îmbucătură în plus, să ducă mai departe făclia speciei. Aceasta a fost evoluţia firească a naturii de la facerea ei şi e posibil ca fără această "ladă de gunoi" a speciilor extincte ori a limbilor dispărute nici noi să nu fi fost atât de inteligenţi şi de bine făcuţi şi nici o limbă atât de frumoasă precum italiana, ca să luăm un exemplu, să nu fi existat.

În acest calcul liniştitor, nu putem însă să lăsăm la o parte angoasa ca nu cumva societatea modernă, prin noile canale de comunicaţie create, capabile de disiminare planetară şi instantanee, să fi denaturat într-un fel compoziţia supei primordiale în care se coace dintotdeauna evoluţia limbilor. Pentru ca o entitate colectivă (specie sau limbă) ce se supune unor legi compoziţionale (de încrucişare sau de comunicare) să evolueze natural spre o identitate "remarcabilă" (singura formă care ne interesează), e nevoie să fie respectate

[1] *European Chapter of the Asscociation for Computational Linguistics* (EACL)
[2] *less-studied languages*
[3] *low density languages* (desigur densitatea aici referindu-se la

simultan: o dimensiune minimă, care să împiedice o proliferare a raporturilor incestuoase, dacă vorbim de specie, sau o plafonare a formei şi lexicului datorată placidităţii imaginative a unei colectivităţi redusă numeric, dacă vorbim de limbă, dar, credem noi, şi o dimensiune maximă, care e totuna cu o anumită izolare geografică sau comunicaţională, pentru a împiedica o diluare a trăsăturilor genetice, dacă vorbim de specie, sau o impurificare lingvistică, dacă vorbim de o limbă naturală.

Forţarea dimensiunii maximale poate duce la pierderea identităţii unei limbi "năcăjite" prin agresiunea arogantă a limbilor tari, a celor ce fac ocolul globului în mediile electronice. În ziua întâi a globalizării, această perspectivă sumbră este cunoscută şi guvernele statelor lumii devin din ce în ce mai conştiente de necesitatea apărării identităţii lingvistice autohtone. Globalizarea nu trebuie să conducă la o amestecătură lingvistică amorfă, un fel de *new-speak* planetar, ci dimpotrivă, la cultivarea limbilor proprii prin uşurarea folosirii lor în comunicarea în mediile electronice. Ori este evident că acest lucru nu poate fi realizat fără aportul tehnologiilor informatice aplicate limbajului natural.

2. Comisia şi Consorţiul de Informatizare pentru Limba Română

În octombrie 2001, a fost creată în Academia Română, în cadrul Secţiei de Ştiinţa şi Tehnologia Informaţiei, Comisia de Informatizare pentru Limba Română (CILR). Scopul acestei comisii îl constituie apărarea identităţii limbii române prin promovarea studiilor dedicate ei dintr-o perspectivă informaţională. Ulterior, la Adunarea Generală a CILR din martie 2002 s-a hotărât împărţirea activităţii Comisiei în patru sectoare: lingvistică teoretică şi formală, prelucrarea limbajului natural, prelucrarea vorbirii şi terminologie.

Rezultat al unor întâlniri şi discuţii desfăşurate între lingvişti şi informaticieni ce au avut loc succesiv în Bucureşti, Iaşi, şi Chişinău, încă înainte de apariţia Comisiei, se prefigura necesitatea creării unui Consorţiu care să asigure un cadru organizat de comunicare între cercetătorii lingvişti şi informaticieni şi de depozitare de informaţii în domeniile de lucru promovate de Comisie. Momentul decisiv în crearea Consorţiului de Informatizare pentru Limba Română (ConsILR) poate fi considerat eliberarea unei finanţări, în primăvara anului 2002, din partea proiectului naţional Societatea Informaţională – Societatea Cunoaşterii, finanţare care a permis construirea unui sit electronic dedicat[4].

Explozia de informaţii pe Internet, necesitatea tot mai mare de a accede la ele prin mijloace inteligente utilizând limba naturală proprie, indiferent de limba de depozitare a informaţiilor, face astăzi ca domeniul Prelucrării Limbajului Natural să fie considerat prioritar în Comunitatea Europeană. Tot mai multă lume recunoaşte că Tehnologia Limbajului va deveni domeniul primordial al Tehnologiei Informaţiei în secolul nostru. Tot mai multe limbi sunt astăzi extrem de bine studiate şi reprezentate în mediile electronice. Rămânea în urmă în acest domeniu înseamnă un handicap tehnologic cu urmări care nu se limitează numai asupra limbii proprii, ce poate fi grav afectată, dar poate avea şi consecinţe economice greu de recuperat. Se manifestă, ca urmare, o necesitate stringentă de a ridica limba română, în privinţa resurselor şi a tipurilor de prelucrări automate, la nivelul altor limbi europene. Acesta este scopul principal al creării Consorţiului.

În al doilea rând, limba română dispune de fonduri de date lingvistice impresionante (dicţionare explicative, tezaur, terminologice, bilingve etc.) care, din păcate, din cauza unei politici de protecţie deficitară ori a neînţelegerii posibilităţilor de exploatare oferite de informatică, sunt puţin ori deloc cunoscute publicului larg ori cercetătorilor din lingvistică computaţională. Alte tipuri de resurse în limba română, în special corpusuri lingvistice, în stare neprelucrată ori adnotate, există într-un număr insuficient încă în comparaţie cu alte limbi europene. De crearea lor depinde accederea la tehnologia lingvistică modernă şi este neîndoios că acest proces trebuie accelerat.

În al treilea rând, doar puţini cercetători lingvişti sunt astăzi conştienţi de avantajele enorme pe care le-ar putea avea în munca lor apelând la modalităţi de prelucrare automată a

[4] Adresa acestui sit este http//consilr.info.uaic.ro.

datelor lingvistice sau slujindu-se de calculator ca de un instrument de lucru. Conştientizarea în rândul lingviştilor a posibilităţilor de utilizare a tehnologiilor informatice în cercetarea lingvistică este încă un motiv în favoarea creării Consorţiului.

În al patrulea rând, puţini cercetători care lucrează în lingvistică computaţională ori în prelucrarea limbajului natural ori a vorbirii, în România şi Republica Moldova, şi care au un fundament informatic, dispun de cunoştinţe lingvistice suficiente care să le permită o integrare la un nivel înalt a cunoştinţelor din cele două domenii. Atragerea lingviştilor de formaţie clasică în colective mixte de cercetare alături de informaticieni, cât şi promovarea studiilor post-universitare, de masterat ori doctorat în domeniul lingvisticii computaţionale şi al prelucrării limbajului natural este un alt argument în favoarea creării Consorţiului.

Activitatea ConsILR este caracterizată, în principal, de următoarele trăsături:

- partenerii la Consorţiu se pot asocia pentru alcătuirea de proiecte adresate unor instituţii finanţatoare naţionale ori internaţionale, în vederea solicitării de granturi de cercetare şi mobilitate;
- Consorţiul se doreşte să devină un forum al ideilor de cercetare teoretică şi aplicativă asupra limbii române, inclusiv pentru dezvoltarea de resurse dedicate limbii române;
- Consorţiul se preocupă de includerea pe situl său, într-un format electronic care să permită exploatarea facilă, a acelor componente ce se consideră că au o valoare covârşitoare pentru limba română (dicţionare, enciclopedii, corpusuri textuale sau fonetice, gramatici, terminologii etc.). Orientarea generală trebuie să fie una de transparenţă, militându-se pentru obţinerea acordului de acces public asupra a cât mai multe componente considerate de interes naţional şi internaţional;
- rezultatele eforturilor de cercetare, prin consensul autorilor, por fi depozitate (şi) pe situl Internet al Consorţiului, autorii contribuţiilor urmând să-şi păstreze neştirbite drepturile de autor. Membrii Consorţiului, în nume individual sau colectiv, pot utiliza componente aflate pe sit pentru a le exploata în scopuri didactice ori ştiinţifice, sau în scopuri ce se pot finaliza cu profit, după obţinerea acordului autorilor şi, dacă e cazul, reglementarea cu aceştia a manierei de participare la profit;
- Consorţiul îşi propune să identifice surse de finanţare (obţinute din contracte de cercetare/dezvoltare, exploatarea componentelor de pe sit, cotizaţii etc.), pe care să le poată folosi pentru cercetare şi dezvoltare, inclusiv pentru creşterea resurselor de calcul şi depozitare electronică a datelor, sprijinirea participării membrilor săi la conferinţe naţionale şi internaţionale, organizarea de întruniri, şcoli de scurtă durată, cursuri de utilizare a calculatoarelor pentru prelucrări lingvistice, cursuri de masterat şi doctorat în domeniul tehnologiei limbajului uman şi al lingvisticii computaţionale, reclama componentelor sitului în vederea valorificării lor, activităţi de editare etc.;
- Consorţiul va trebui să ajungă la confortul financiar care să facă posibilă editarea cel puţin a unui periodic ştiinţific propriu (inclusiv în format electronic), eventual înfiinţarea unei edituri care să publice cercetările şi realizările membrilor săi ca serii de cărţi şi CD-uri.

3. Resurse, instrumente, arhitecturi şi standarde lingvistice şi fonologice computaţionale pentru limba română

3.1 Metodologie

În versiunea finală, situl ConsILR va trebui să reflecte patru teme mari: **Resurse – R**, **Standarde – S, Instrumente – I** şi **Arhitecturi – A**. Fiecare dintre aceste teme va fi divizată în subteme corespunzând componentei **Textuale** – **T** şi **Fonologice** – **F**, rezultând subtemele:

RT - Resurse Textuale,

RF - Resurse Fonologice,

ST - Standarde Textuale,

SF - Standarde Fonologice,

IT - Instrumente Textuale,

IF - Instrumente Fonologice,

AT - Arhitecturi Textuale,

AF - Arhitecturi Fonologice

AM - Arhitecturi Mixte (aplicaţii care integrează atât componente de prelucrări textuale cât şi a vorbirii).

Tema de **Resurse** urmăreşte colectarea unui eşantion semnificativ sub aspect lingvistic (textual şi fonologic) al limbii române. Ea va ţinti colectarea şi/sau crearea de colecţii de date scrise şi vorbite care să se constituie într-un depozit informaţional care să susţină aplicaţiile de prelucrare a limbii scrise şi vorbite, sau din care să se infereze colecţii de reguli şi modele statistice pentru limba română. Ca metodologie generală se recomandă un echilibru între utilizarea corpusurilor pentru inferarea colecţiilor de reguli, structuri sintactice ori modele ale limbii şi introducerea lor de mână de către experţi lingvişti. Atragem atenţia asupra faptului, în general, recunoscut că o inferare automată din corpusuri a regulilor este rapidă, inducând descrieri consistente, dar prea generale (ceea ce se poate traduce în acceptarea unor construcţii agramaticale), în timp ce compunerea lor manuală este consumatoare de timp, poate provoca inconsistenţe, poate duce la respingerea unor construcţii uzitate dar încă neincluse în normă, pentru că sunt greu de găsit mecanisme adecvate de control, dar poate genera reguli mult mai compacte. Alinierea prin mijloace tradiţionale a normei la uz este un proces lent şi imperfect, care poate fi îmbunătăţit numai prin mijloace automate.

Tema de **Standarde** urmăreşte codificarea formatelor intermediare în lanţul de prelucrări. Prin crearea acestui set de standarde se urmăreşte compatibilizarea în intrare/ieşire a modulelor de prelucrări textuale şi fonologice. Formatul de codificare promovat este XML (http://www.w3.org/XML/), în cadrul căruia se pot adopta standarde internaţionale (parţial) acceptate precum TEI (Barnard&Ide, 1997), XCES (Ide&Romary, 2000) Parole (Ruimy *et al.*, 1998) etc. Pentru formatele pentru care nu există standarde sau practici general acceptate ,se are în vedere propunerea de noi scheme de adnotare şi documentarea lor (sintaxă şi semantică). În mod normal, detalierea tipurilor de codificări ar trebui să preceadă elaborarea instrumentelor şi a arhitecturilor, deşi un proces de rafinare a standardelor ulterior elaborării modulelor componente ale aplicaţiilor ar putea, de asemenea, fi acceptat în anumite situaţii. Ideea generală este ca, prin standardizare, diverse module ale aplicaţiilor să se poată cupla în lanţuri alternative de prelucrări. Pe de altă parte, ori de câte ori o nouă schemă de codificare se propune, ea trebuie să fie însoţită de o descriere formală (sintaxă şi semantică) care să permită conversia (semi-)automată şi deterministă ulterioară din/spre scheme de adnotare viitoare sau compatibilizarea cu scheme existente.

Tema de **Instrumente** urmăreşte realizarea unor capabilităţi de prelucrare textuală şi fonologică a limbii române grupate într-un set de module software. Acestea pot fi realizate în totalitate în cadrul Consorţiului dar şi preluate din surse externe (eventual adaptate). În principiu, instrumentele, componente software, trebuie să fie independente de limbă, utilizarea lor în tandem cu resurse româneşti atribuindu-le funcţionalitatea specifică prelucrării limbii române.

Tema de **Arhitecturi** urmăreşte realizarea unui sistem inteligent, capabil să asiste construirea de aplicaţii complexe de prelucrare a limbii române în formă scrisă, vorbită sau mixtă. Strategia de dezvoltare trebuie să sprijine utilizatorul (informat) în procesul de construire a unei aplicaţii pentru limba română prin combinaţii de module din sfera prelucrării textuale/fonologice (independente de limbă), oferită de tema de Instrumente. Alinierea la standarde atât a realizării instrumentelor cât şi a resurselor este o premisă esenţială pentru realizarea efectivă a obiectivelor Consorţiului.

Vom detalia în cele ce urmează câteva componente din domeniul textual.

3.2 Resurse Textuale (subtema RT)

Se urmăreşte crearea următoarelor resurse textuale pentru limba română:

- **RT-CorpusText**: o colecţie structurată de texte româneşti în format electronic, independent de platforma hardware sau software de prelucrare, acoperind o arie largă de registre lingvistice (ziaristic, beletristic, ştiinţific, e-mail etc.), arii geografice (Ardeal, Moldova, Banat, Moldova românească, Basarabia etc.), origini în timp (secolul trecut, limba dintre cele două războaie, limba actuală) ori origini sociale (livresc, limbajul cult, limbajul suburban ori de argou etc.). Textele incluse în această colecţie trebuie să conţină obligatoriu informaţie de identificare a criteriilor de clasificare (sursă, autor, registru, arie geografică, perioadă etc.). Pe baza acestor criterii, explicit codificate, pot fi extrase oricând corpusuri speciale (tematice) sau balansate (de referinţă), corpusuri diacronice (monitor) etc. Dimensiunea unei astfel de colecţii, pentru a fi reprezentativă, trebuie să fie cât mai mare, conţinând cel puţin 500 milioane de cuvinte ocurenţă. O tipologie a corpusurilor general acceptată de comunitatea ştiinţifică din domeniul prelucrării limbajului natural este prezentată in (Sinclair, Ball, 1995) şi comentată în (Teubert,1997).
- **RT-CorpusAnnotXXX**[5]: o colecţie de corpusuri româneşti, adnotate în conformitate cu diverse scheme. Adnotarea trebuie realizată în format XML (elementar şi din ce în ce mai complex, în diverse formate – v. şi TEI/CES). Adnotarea se va realiza fie prin proceduri semi-automate (asistată de calculator), fie complet automată. Tipul abordării depinde de natura şi granularitatea informaţiei adăugată textelor (care în fond definesc dificultatea procesului de adnotare). În acest proces vor putea fi folosite instrumente din categoria **IT-Annot** (v. subtema **IT**). Diversele componente ale acestor corpusuri vor constitui sursele principale de antrenare a componentelor temei **IL**. Este recomandabil ca diferitele componente ale resursei **RT-CorpusAnnotXXX** să utilizeze într-o măsură cât mai mare aceleaşi texte de origine, pentru a facilita în acest fel coeziunea unor lanţuri de antrenare a componentelor din **IT**. Din punct de vedere tehnic, este de asemenea de dorit ca adnotările la diferite niveluri, aplicate aceluiaşi text, să poată fi separate oricând prin filtre, lucru ce poate fi realizat uşor prin transpunerea adnotărilor în baze de date. Codificările folosite în **RT-CorpusAnnot** trebuie să fie în concordanţă cu standardele descrise în secţiunea **ST** a proiectului.
    - **RT-CorpusAnnotMSD**: corpusuri adnotate la nivel morfo-lexical;
    - **RT-CorpusAnnotNP**: corpusuri adnotate la nivelul grupurilor nominale;
    - **RT-CorpusAdnotSyn**: corpusuri adnotate la nivel sintactic, având specificate atât structurile sintactice cât şi dependenţele dintre ele; astfel de corpusuri adnotate se mai numesc şi *bănci de arbori* (*tree-banks*);
    - **RT-CorpusAnnotArg**: corpusuri adnotate la nivel semantic având specificate structurile de tip predicat-argument, cu identificarea statutului (opţional sau obligatoriu) şi a rolului fiecărui argument (agent, beneficiar, loc, mod etc); astfel de corpusuri se mai numesc şi *bănci propoziţionale* (*prop-banks*), propoziţia fiind considerată în accepţiunea ei din logica predicatelor: o expresie logică de predicate *n*-are, fiecare predicat fiind descris de o schemă semantică precizând atât numărul şi tipul argumentelor sale cât şi numele lor;
    - **RT-CorpusAnnotSeg**: corpusuri adnotate la segmente (propoziţii în cadrul frazelor, unităţi de discurs, adesea confundate);
    - **RT-CorpusAnnotAna**: corpusuri adnotate la relaţii anaforice: coreferinţa (echivalenţa referenţială), referinţe meronimice explicite (parte-întreg, obiect-

[5] Sufixul **XXX** adăugat numelui unei componente, aici cât şi mai departe, semnifică posibilitatea de împărţire a unei componente în mai multe subcomponente, din aceeaşi clasă.

proprietate, posesor-obiect posedat etc.) sau implicite (*bridge anaphora*) (ce pot fi de aceleaşi tipuri ca şi cele explicite);

- **RT-CorpusAnnotWSD**: corpusuri adnotate la sensuri ale cuvintelor, ce pot fi folosite pentru inferarea regulilor de identificare a sensurilor cuvintelor (*word sense disambiguation*);
- **RT-CorpusAnnotDS**: corpusuri adnotate la structură de discurs, de exemplu *structura relaţiilor retorice* (*Rhetorical Structure Theory - RST*).

Adesea corpusurile adnotate cuprind niveluri diferite de adnotare. De exemplu, un corpus din categoria **RT-CorpusAnnotDS** trebuie să cuprindă, cel puţin, următoarele niveluri de adnotare: morfo-lexicală, grupuri nominale, unităţi de discurs şi relaţii retorice. Clasificarea noastră trebuie înţeleasă ca identificând o listă a elementelor specifice de adnotare, încât un corpus adnotat anumit să poată aparţine simultan mai multor tipuri.

- **RT-CorpusParallel**: un corpus de texte paralele (română – o limbă străină) aliniate. Alinierea poate fi realizată la nivel de paragraf, propoziţie, sau chiar cuvânt. Un astfel de corpus este extrem de util pentru antrenarea componentelor de traducere automată (v. **IT-TradXXX**).
- **RT-RegSegm**: o colecţie de reguli/şabloane pentru segmentarea textelor româneşti. Segmentarea poate fi făcută la diverse niveluri, incluzând un nivel morfologic, un nivel sintactic şi un nivel al discursului. O componentă a acestei resurse va putea fi integrată instrumentului **IT-Tokenizer** (v. subtema **IT**) ce urmăreşte recunoaşterea articolului lexical de prelucrare primară; noţiunea de articol lexical depinde de natura aplicaţiilor ce folosesc prelucrarea limbajului natural şi, în acest sens, specificarea regulilor de segmentare trebuie să fie parametrizabilă în raport cu fineţea şi profunzimea prelucrărilor. Regulile de segmentare (cel mai adesea exprimate prin *expresii regulate*) permit recunoaşterea abrevierilor, datelor calendaristice, numerelor, valorilor financiare, formulelor, tabelelor etc. De asemenea, pe baza unor resurse lexicale specifice, pot fi descrise structuri multi-cuvânt ce urmează să fie tratate ca un singur articol lexical (*de+jur+împrejur => de_jur_împrejur*, *de-a+dreptul => de-a_dreptul*, *verde+de+Paris => verde_de_Paris*) sau dimpotrivă, forme cliticizate ce trebuie interpretate ca două sau mai multe articole lexicale (*dându-mi-le => dând+mi+le*). Un segmentator mai special este cel ce recunoaşte numele proprii şi în plus le atribuie o etichetă semantică dintr-o mulţime specificată apriori (persoană, instituţie, loc etc.)[6]. Alte colecţii de reguli/expresii regulate de segmentare urmăresc identificarea grupurilor nominale nerecursive, a propoziţiilor componente ale frazelor sau a unităţilor de discurs[7]. Drept surse pentru această resursă se pot folosi: dicţionare de abrevieri, dicţionare de cuvinte compuse, lista cliticelor limbii române, liste de nume proprii (*gazetteers*) etc.
- **RT-DictMorphLex**: dicţionar morfo-lexical conţinând colecţia cuvintelor flexionate româneşti cu descrieri morfo-lexicale ataşate incluzând obligatoriu lema, ca structuri de caracteristici nerecursive sau codificate pe câmpuri (*Morpho-Sintactic Description – MSD*). Identificarea unui model de codificare standardizată, din punct de vedere morfo-lexical, aplicabil cât mai multor limbi a fost unul din obiectivele mai multor proiecte europene (EAGLES, MULTEXT, MULTEXT-EAST) pentru limba română, o instanţiere a acestui model fiind descrisă în (Tufiş et al., 1997).
- **RT-MorphPara**: o descriere paradigmatică a morfologiei româneşti, care să cuprindă colecţia completă a paradigmelor flexionare româneşti (terminaţii, rădăcini de cuvinte şi asocieri valide rădăcini-terminaţii). Un instrument de analiză care ar lucra cu această

[6] Engl. *name entity recognizer*.

[7] Secvenţe lexicale ce se constituie în unităţi atomice ale structurii de discurs, cel mai adesea propoziţii de sine stătătoare sau propoziţii constitutive ale frazelor. Uneori însă o unitate de discurs poate fi mai largă decât strict o propoziţie din constituţia unei fraze.

componentă ar trebuie să fie capabil să „spargă" un cuvânt flexionat în rădăcină şi terminaţie prin încadrarea lui într-o paradigmă cunoscută. În cazul cuvintelor necunoscute, componenta ar trebui să „ghicească" categoria morfo-lexicală cea mai plauzibilă. Alegerea celei mai plauzibile interpretări este o problemă contextuală, de multe ori rezolvabilă printr-o analiză strict distribuţională (v. şi **RT-Guesser**).

- **RT-Guesser**: o colecţie de reguli de inferenţă (statistică) a celei mai probabile interpretări morfo-lexicale a unui cuvânt necunoscut.
- **RT-LangModel**: un model al limbii române scrise, adică o reţea de probabilităţi a grupărilor (bigrame, trigrame) de categorii morfo-sintactice, obţinută prin antrenare pe un corpus. **RT-LangModel** poate avea realizări specifice, în funcţie de registrul lingvistic, autor etc.
- **RT-WordNet**: un *wordnet* românesc. Într-un *wordnet* (Miller *et al*., 1990), seriile sinonimice (*sinset*-uri), caracterizate de un înţeles comun, sunt noduri ale unei reţele semantice. Relaţiile ce corelează nodurile reţelei sunt de natură lexicală sau semantică. Nodurile mai multor reţele semantice monolingve pot fi puse (în mod independent) în relaţii de echivalenţă conceptuală cu reprezentări independente de limbă. Folosind proprietăţile relaţiilor de echivalenţă, se pot identifica, pornind de la reprezentările conceptuale, lexicalizările acestora în diverse limbi. Realizarea *wordnet*-ului pentru limba română, aliniat la nucleul comun european, constituie un obiectiv însemnat pentru abordarea traducerii automate având limba română la un pol, cât şi pentru generarea de dicţionare electronice bilingve spre/dinspre limbi pentru care astfel de resurse lipsesc (Tufiş, 2002).
- **RT-DictColocat**: un dicţionar de colocaţii pentru limba română. El trebuie să cuprindă, pentru fiecare sens al unui verb sau substantiv derivat din verb, exemple de colocatori frecvenţi sau minimali.
- **RT-DictSubcat**: un dicţionar de subcategorizare al limbii române (pentru limba engleză v. Beth Levin). Acest dicţionar trebuie să cuprindă (în varianta sa completă) pentru fiecare sens al fiecărui verb (inclusiv derivatele sale nominale) din limba română, structura sa de subcategorizare (argumentele obligatorii, cu restricţiile sintactice şi semantice). Construcţia dicţionarului se poate automatiza într-o mare măsură în condiţiile existenţei unei corpus reprezentativ adnotat la nivel sintactic si semantic (*tree-bank, prop-bank*).
- **RT-DictParalelXXX**: un număr de dicţionare paralele pot fi realizate (semi-)automat din momentul în care va fi finalizată componenta **RT-WordNet**, aliniată cu alte limbi.

3.3 Standarde Textuale (subtema ST)

Principalele standarde aplicate resurselor textuale pe care le avem în vedere sunt:

- **ST-Txt**: este notaţia noastră pentru text neadnotat.
- **ST-CorpusAnnotXXX**: o colecţie de standarde de adnotare a corpusurilor ca şi a textelor prelucrate automat. Includerea în aceeaşi categorie a acestor două clase de documente este voită şi exprimă tendinţa de uniformizare a codificărilor ataşate corpusurilor (de obicei, interactiv, de către experţi, prin utilizarea instrumentelor de adnotare) cu cele generate automat de maşină, în urma proceselor de analiză. Trebuie avută în vedere posibilitatea ca mai multe adnotări, cu destinaţii diferite, să fie combinate pe acelaşi corpus sau text, ceea ce impune necesitatea ca un standard să reflecte doar o soluţie elementară de adnotare, iar nu o combinaţie de codificări. O astfel de concepţie ar permite compunerea liberă a adnotărilor elementare, dacă restricţiile limbajului XML sunt respectate (de exemplu neintersectarea etichetelor) - v. (Cristea *et al.*, 1998a), (Cristea *et.al.*, 1998b). Din acest motiv, unei componente din **RT-CorpusAdnotXXX** îi pot

corespunde una sau mai multe componente din **ST-CorpusAdnotXXX**. Exemple de componente din această categorie sunt:

ST-CorpusAnnotTok: formatul de după **IT-Tokenizer** (v. secţiunea **IT**). El marchează cu etichete distincte: cuvinte, clitice, semne de punctuaţie, date calendaristice, formule, tabele etc.

ST-CorpusAnnotMSDXXX: formate pentru descrieri morfo-sintactice, în care fiecărui element lexical (*token*) i se ataşează una sau mai multe interpretări morfo-sintactice. Aceste formate pot cuprinde şi marcaje pentru cuvinte necunoscute (necuprinse în dicţionar). Variante, pot fi:

ST-CorpusAnnotMSDk: una sau mai multe interpretări pentru fiecare *token*, posibil şi *token*-uri nemarcate;

ST-CorpusAnnotMSDk+: una sau mai multe interpretări pentru fiecare *token*, alsolut toate *token*-urile marcate;

ST-CorpusAnnotMSD1+: o unică interpretare pentru fiecare *token*, toate *token*-urile marcate.

ST-CorpusAnnotNP: standard de codificate a grupurilor nominale de recursivitate limitată (în sensul că nu conţin subgrupuri verbale; de exemplu în *Omul pe care l-am văzut conducând maşina este directorul firmei* se notează ca grup nominal *omul* iar nu *omul pe care l-am văzut conducând maşina*).

- **ST-RegSegmXXX**: format de codificare a regulilor de segmentare la diferite niveluri.
- **ST-DictSubcat**: formatul de codificare a dicţionarelor de subcategorizare.
- **ST-WordnetXXX**: formatul de codificare a diferitelor componente ale unei baze de date de tip *wordnet*.

3.4 Instrumente pentru prelucrări textuale (subtema IT)

Modulele temei **I** sînt componente software a căror caracteristică esenţială este independenţa de limbă. Aceasta înseamnă că aplicarea lor la limba română trebuie să se facă prin utilizarea de resurse specifice, componente ale temei **RT**. Fiecare modul realizează o funcţie dată, fiind cuplat într-un sistem prin intermediul a trei tipuri de porţi: intrări, ieşiri şi resurse (v. Figura 1).

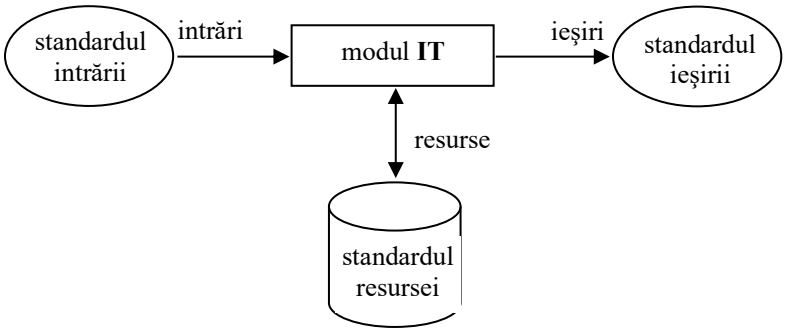

Figura 1: Interfaţa unui modul cu sistemul

Pentru simplificarea notării interfeţelor modulelor temei **I**, vom utiliza în cele ce urmează o convenţie de desemnare a acestora (semnături) de forma:

<nume modul>**(I**:<listă standarde intrări>**, R**:<listă resurse>**, O**:<listă standarde ieşiri>**)**.

Intrările şi ieşirile vor fi compatibilizate astfel prin componentele temei de standarde – **S**.

Dăm în continuare a listă a câtorva dintre cele mai semnificative tipuri de instrumente de prelucrări lingvistice:

- **IT-Tokenizer**: modúl de depistare a graniţelor dintre cuvinte, de recunoaştere a formulelor, tabelelor, cliticelor, abrevierilor, datelor calendaristice, numerelor, valorilor financiare, compuşilor şi a numelor proprii. Acest modul primeşte în intrare un text sursă (format **ST-Txt**), accesează **RT-RegSegm** şi produce în ieşire un format **ST-Tok**, sau, în convenţia noastră:

  IT-Tokenizer(I:ST-Txt, R:RT-RegSegm, O:ST-Tok).

- **IT-LookUpMSD**: modúl de depistare a interpretărilor morfo-lexicale ale cuvintelor prin căutare în dicţionar. Semnătura propusă este:

  IT-LookUpMSD(I:ST-Tok, R:RT-DictMorphLex, O:ST-CorpusAnnotMSDk).

- **IT-LookUpPar**: modúl de depistare a interpretărilor morfo-lexicale ale cuvintelor prin analiză paradigmatică. Rezultatul trebuie să fie analog lui **IT-LookUpMSD**:

  IT-LookUpPar(I:ST-CorpusAnnotTok, R:RT-MorphPara,
  O:ST-CorpusAnnotMSDk).

- **IT-Guesser**: modúl de tratare a cuvintelor necunoscute. Trebuie să furnizeze o clasă de interpretări posibile pentru un cuvânt necunoscut, important aici fiind să nu se piardă interpretări. Este de presupus că, la ieşirea din **IT-Guesser**, formatul **ST-CorpusAnnotMSD** nu mai conţine *token*-uri neetichetate. În subtema S acest format este notat **ST-CorpusAnnotMSDk+**. Semnătura propusă:

  IT-Guesser(I:ST-CorpusAnnotMSDk, R: RT-Guesser, O:ST-CorpusAnnotMSDk+).

- **IT-TaggerXXX**: un set de module de dezambiguizare morfo-sintactică. O componentă de acest tip presupune existenţa unui model al limbii **RT-LangModel** şi a unei funcţii de optimizare (un criteriu de decizie). Se pot adopta mai multe soluţii: soluţia „lacomă"(*greedy*) **IT-TaggerGre** – în care se atribuie o unică interpretare pentru fiecare *token*; soluţia celor mai bune *k* etichete (*k-best tagging*) **IT-TaggerkBT** – în care se acceptă cel mult *k* interpretări atunci când probabilităţile etichetelor posibil de atribuit unui token nu sunt suficient departajate, maniera etichetării pe niveluri (*tiered-tagging*) (Tufiş, 1999), algoritmul Welch Baum (Baum, 1972), algoritmul Brill (Brill, 1994) etc. Câteva semnături posibile sunt:

  IL-TaggerGre(I:ST-CorpusAnnotMSDk+, R:RT-LangModel,
  O:ST-CorpusAnnotMSD1+)

  IL-TaggerkBT(I:ST-CorpusAnnotMSDk+, R:RT-LangModel,
  O:ST-CorpusAnnotMSDk+).

- **IT-ChunkerXXX**: module de depistare a grupurilor nominale, prepoziţionale etc. utilizând reguli de parsare de suprafaţă (*shallow-parsing*) (Aït-Mokhtar&Chanod, 1997). O semnătură posibilă este:

  IT-NPChunkerNP(I:ST-CorpusAnnotMSD1+, R: RT-RegSegm,
  O:ST-CorpusAnnotNP).

- **IT-ChuRulesInducerXXX**: module de inferenţiere a regulilor de depistare a grupurilor simple (nerecursive) din corpus. Aceste module prelucrează componente ale corpusurilor adnotat la grupurile respective pentru a genera resurse de folosit de către instrumente de parsare. Un exemplu de modúl din această clasă îl reprezintă un inferenţiator de reguli de depistare a grupurilor nominale:

  IT-ChuRulesInducerNP(I:ST-CorpusAnnotNP, R:∅, O:ST-RegSegmNP).

- **IT-ConcordXXX**: instrumente de găsire a concordanţelor lexicale şi gramaticale într-un corpus. Lexicograful (pe post de utilizator) poate dispune de facilităţi de alegere a criteriilor de selecţie atât pentru elementul central căutat cât şi a celor folosite ca filtre

adiţionale (contexte ale elementului central). Un concordanţier lexical (**IT-ConcordLex**) lucrează pe un corpus textual (**RT-CorpusText**) pe când unul gramatical (**IL-ConcordGra**) – pe un corpus adnotat la componente sintactice (o componentă a **RT-CorpusAnnotXXX**). Cu instrumente din această clasă se pot forma dicţionare de colocaţii (**RT-DictColocat**) şi de subcategorizare (**RT-DictSubcat**). Un concordanţier gramatical ar trebui, în principiu, să permită lexicografului consultarea interactivă a caracteristicilor diverselor elemente de context. Semnături propuse:

ITConcordLex(I:ST-Txt, R:∅, O:{∅, ST-DictColocat})

ITConcordGra(I:ST-CorpusAnnotXXX, R:∅, O:{∅, ST-DictSubcat}).

Ieşirile vide ce apar în semnăturile de mai sus se referă la utilizarea programelor concordanţiere pentru obţinerea de liste.

- **IT-StatistXXX**: pachet de instrumente pentru generarea de estimări statistice pe corpusuri. Exemple de astfel de estimări sunt informaţiile mutuale (astfel, dacă frecvenţa de co-apariţii a unui verb împreună cu un grup nominal depăşeşte un anumit prag se poate genera concluzia unei utilizări tranzitive).
- **IT-TradXXX**: pachet de instrumente pentru ajutorarea traducerii în/din limba română.
- **IT-SpellXXX**: pachet de instrumente pentru corectarea ortografiei româneşti.
- **IT-DiscourseXXX**: pachet de instrumente pentru prelucrări ale discursului ş.a.m.d.

3.5 Arhitecturi pentru prelucrări textuale (subtema AT)

Maniera de conectare a unui modul prin cele trei tipuri de porţi (v. tema **I**) permite realizarea de interconexiuni de prelucrare complexe. Unul dintre acestea este cel în cascadă (*pipe-line*) , v. Figura 2.

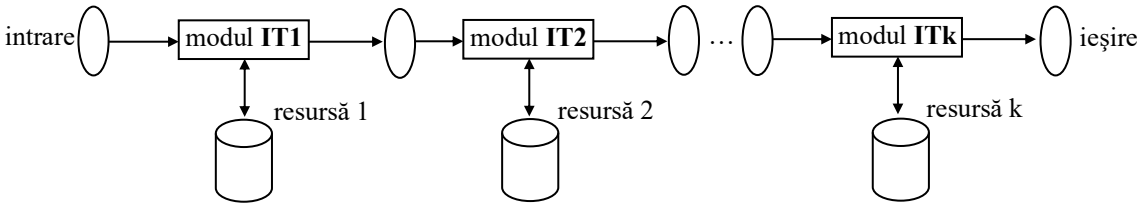

Figura 2: Arhitectură în cascadă (ovalurile reprezintă standarde)

Aceeaşi schemă de proiectare individuală a modulelor poate asambla, de asemenea, arhitecturi tablă neagră (*blackboard*), prin utilizarea unor ieşiri specializate să scrie pe o resursă temporară comună, care este tabla (v. Figura 3).

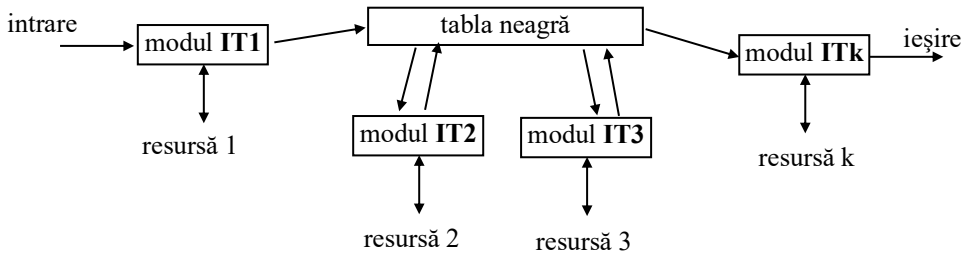

Figura 3: O arhitectură tablă neagră

Pentru realizarea de configuraţii mobile, este recomandabil ca modulele să poată lucra în două maniere: intrare completă (sincron) sau intrare incrementală (asincron). În varianta cu intrare completă, tot textul este consumat de primul modul, înainte de ca ieşirea acestuia să

constituie intrare pentru al doilea modul ş.a.m.d. (v., de exemplu, arhitectura GATE – (Tablan *et al.*, 2002)). Într-o variantă incrementală, intrarea ar putea fi fragmentată, un modul trebuind să rezolve doar o parte din intrare. Acest tip de prelucrare poate fi utilă în conceperea arhitecturilor *tablă-neagră*.

Problema asigurării portabilităţii aplicaţiilor este deosebit de importantă. Neglijarea ei poate transforma o aplicaţie de succes într-un eşec comercial prin imposibilitatea de a se adresa unei mase de utilizatori ce nu au acces la platforma de calcul pentru care a fost ea construită. Pentru asigurarea portabilităţii aplicaţiilor, recomandăm asigurarea accesului la distanţă, pentru execuţie de pe orice platformă utilizator, prin intermediul unor interfeţe de compatibilizare (v. Figura 4).

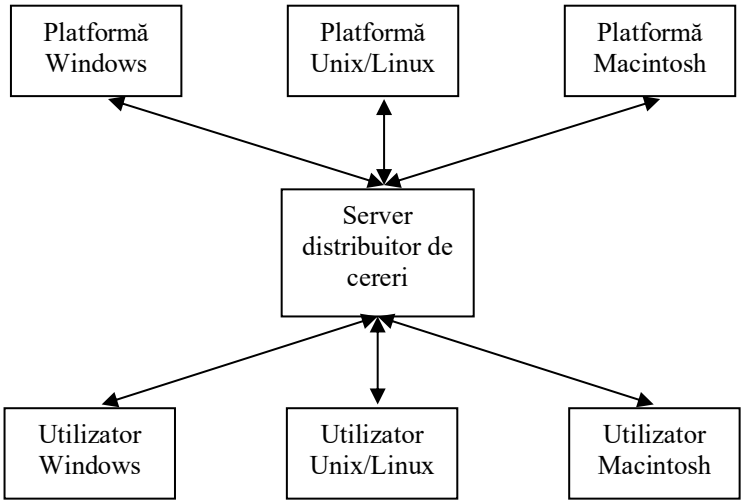

Figura 4: Arhitectură de comutare, transparentă, pentru asigurarea portabilităţii aplicaţiilor

Zona aplicativă a sitului va trebui să permită şi interogări *on-line*, inclusiv configurarea dinamică într-un timp scurt a unor aplicaţii de prelucrare a textelor sau vorbirii pentru limba română.

Referinţe bibliografice

Aït-Mokhtar, S., Chanod, J.-P. (1997). Incremental Finite-State Parsing, *Proceedings of the Fifth Conference on Applied Natural Language Processing*.

Barnard, D. and Ide, N. (1997). The Text Encoding Initiative: Flexible and Extensible Document Encoding. *Journal of the American Society for Information Science.*

Baum L.E. (1972). An Inequality and Associated Maximization Technique in Statistical Estimation of Probabilistic Fuctions of Markov Processes, *Inequalities*, 3, 1-8.

Brill, E. (1994). Some Advances In Rule-Based Part of Speech Tagging. *Proceedings of the AAAI.*

Cristea,D.; Ide,N.; Romary,L. (1998a): Marking-up multiple views of a Text: Discourse and Reference. *Proceedings of the First International Conference on Language Resources and Evaluation*, Granada, Spain.

Cristea,D.; Crăciun,O; Ursu,C. (1998b): GLOSS-A Visual Interactive Tools for Discourse Annotation. *Proceedings of the Workshop on Annotation Tools, ESSLLI'98*, Saarbruecken, Germany.

Ide, N., Bonhomme, P., Romary, L. (2000). XCES: An XML-based Standard for Linguistic Corpora.. *Proceedings of the Second Language Resources and Evaluation Conference* (LREC), Athens, Greece.

Miller G.A., Beckwidth R., Fellbaum C., Gross D., Miller K.J. (1990) Introduction to WordNet: An On-Line Lexical Database, *International Journal of Lexicography*, Vol. 3, No. 4 (winter 1990).

Ruimy, N.; Corazzari, O.; Gola, E.; Spanu, A.; Calzolari, N.; Zampolli, A. (1998). LE-PAROLE project: The Italian Syntactic Lexicon, *Proceedings of EURALEX'98*, Université de Liège, Proceedings Volume I.

Sinclair, J., Ball, J. (1995). Text typology *(External Criteria)*. Eagles Report

Tablan, V.; Ursu, C.; Bontcheva, K.; Cunningham, H.; Maynard, D.; Hamza, O.; McEnery, T.; Baker, P.; Leisher. M. (2002). A Unicode-based Environment for Creation and use of Language Resources. *Proceedings of the Third International Conference on Language Resources and Evaluation*, Las Palmas, Spain.

Teubert W. (1997) Language Resources for Language Technology. În Tufiş D., Andersen P. *Recent Advances in Romanian Language Technology*, Romanian Academy Publishing House.

Tufiş D., Barbu A.M., Pătraşcu V., Rotariu G., Popescu C. (1997). Corpora and Corpus-Based Morpho-Lexical Processing. În Tufiş D., P. Andersen (eds.) *Recent Advances in Romanian Language Technology*, Editura Academiei, Bucureşti.

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

Tufiş D., (2002). BALKANET - Tezaur lingvistic multilingv pentru limbile din Balcani, în acest volum.