Dan Tufiş, Florin Gh. Filip (coordonatori)

Limba Română
în
Societatea Informațională - Societatea Cunoaşterii

ACADEMIA ROMÂNĂ

Secția de Ştiința şi Tehnologia Informației

Institutul de Cercetări pentru Inteligență Artificială

Limba Română în Societatea Informațională - Societatea Cunoaşterii
Coordonatori: Dan TUFIŞ, Florin Gh. FILIP

Colecția
Societatea Informațională
Coordonator:
Prof. dr. ing. Doina BANCIU

ACADEMIA ROMÂNĂ

Bucureşti, decembrie 2002

Volumul de față a fost produs de
Institutul de Cercetări pentru Inteligență Artificială (RACAI) al Academiei Române
în cadrul Proiectului "Strategii şi soluții pentru Societatea Informațională -
Societatea Cunoaşterii în România"
din Programul Național de Cercetare-Dezvoltare "INFOSOC",
condus de Institutul Național pentru Cercetare-Dezvoltare în informatică (ICI),
Bucureşti

ACADEMIA ROMÂNĂ
Secția de Ştiința şi Tehnologia Informației
Institutul de Cercetări pentru Inteligență Artificială

Limba Romana

111

Societatea Informațională
Societatea Cunoaşterii

Coordonatori:
Dan TUFIŞ, Florin Gh. FILIP

^EDITURAf

Coediție

EDITURA Expert

Bucureşti, România

Editor şi coordonare editorială: **Valeriu IOAN-FRANC**
Redactori: **Mircea FAȚĂ, Paula NEACŞU, Irina STĂNESCU**
Concepția grafică, machetare şi tehnoredactare: **Luminița LOGIN**
Coperta: **Nicolae LOGIN**

ISBN 973-8177-83-9 Apărut 2002

DEDICAȚIE

*Acest volum este dedicat Academicianului Mihai Drăgănescu, Profesor şi mentorul unei întregi generații de specialişti în ştiința şi tehnologia informați în general şi al problemelor societății informaționale şi a cunoaşterii în specia Marea majoritate a contribuțiilor din acest volum aparțin unor experți ce fac par din Comisia de Informatizare a Limbii Române, comisie a Academiei Române la cărei naştere un rol esențial l-a avut Profesorul Drăgănescu, preşedintele Secției Ştiința şi Tehnologia Informației. Savantul Mihai Drăgănescu are numeroa contribuții în ştiința contemporană, binecunoscute atât în țară cât şi în străinăta Pentru cine îl cunoaşte pare incredibilă puterea sa de muncă, debordan creativitate şi neostoita căutare a noului. Profesorul Drăgănescu este indiscutab port-drapelul conceptului de societate informațională-societate a cunoaşterii România. în lucrările sale din urmă cu peste 25-30 de ani se regăsesc cu clarita multe concepte foarte actuale în zilele noastre, previziuni curajoase atunci, acu realități cotidiene. în lucrările domniei sale din ultima vreme, apare un no concept ce avem convingerea că se va impune: Societatea Conştiinței, o treap superioară a societății cunoaşterii. Nu este de mirare deci că în contextul societă informaționale şi a cunoaşterii profesorul Drăgănescu a susținut cu consecvență a afirmat cu claritate rolul Inteligenței Artificiale în devenirea noilor societăți d cunoaşterii. între domeniile Inteligenței Artificiale un loc de frunte în promovar principiilor societății cunoaşterii îi revine Tehnologiei Limbajului Natur Profesorul Drăgănescu a fost unul dintre puținii oameni de ştiință români care a înțeles şi au sprijin total aceste direcții. Cu aproape douăzeci de ani în urm (1983), Profesorul Drăgănescu edita (împreună cu Adrian Davidoviciu şi lo Georgescu) volumul "Inteligența Artificială şi Robotica' pentru ca trei ani m târziu (împreună cu Corneliu Burileanu) să editeze un alt volum de referin "Analiza şi sinteza semnalului vocal". Astăzi, cercetările mondiale în domeni tehnologilor lingvistice au atins un nivel de maturitate ce permit sinergizar eforturilor lingviştilor, informaticienilor, matematicienilor şi a altor specialişti d sectorul academic sau industrial, să abordeze proiecte mari, interdisciplinare avâ ca obiectiv prelucrarea automată, în mediile de comunicare electronică, a din ce ce mai multe*limbi naturale. Printre acestea, limba română îşi face loc încet d sigur. Volumul de față este o mărturie în acest sens. în același timp, volumul constituie într-o nouă confirmare a realităților pe care Profesorul Mih Drăgănescu le prefigura cu mulți ani în urmă.*

Dr. Dan Tufiş, m.c.A.R, Acad. Florin Gh. Fi

CUPRINS

INTRODUCERE 9

SECȚIUNEA I:
LINGVISTICĂ TEORETICĂ ŞI FORMALĂ; TERMINOLOGIE

Resurse lingvistice pentru limba română elaborate la Institutul de Lingvistică "Iorgu Iordan" - *Ioana Vintilă-Rădulescu* 19

Contribuția lingvisticii la studiul terminologiilor ştiințifice - *Angela Bidu-Vrănceanu* 33

Gramaticile generative nontransformaționale - *Emil lonescu* 39

Către o teorie X-bar funcțională - *Neculai Curteanu* 51

Teoria HPSG. Studiu de caz: acordul încrucişat - *Ana-Maria Barbu* 87

După 10 ani de experiență terminografică: noul model de date terminologice al TermRom - *Dan Matei* 109

Probleme de reprezentare a datelor terminografice într-o bază de date relațională - *Sorin Ghețaru* 121

SECȚIUNEA II:
TEHNOLOGII ALE LIMBAJULUI SCRIS

RO-BALKANET-ontologie lexicalizată, în context multilingv, pentru limba română - *Dan Tufiş, Dan Cristea* 137

Algoritmi de segmentare a textului în unități de tip clauzal - *Dan Gălea, Neculai Curteanu, Cristian Linteş* 165

O metodă automată pentru inserarea diacriticelor în texte în limba română- *Rada F. Mihalcea, Vivi A. Năstase* 191

Contribuții privind structura statistică de cuvinte în limba română scrisă - *Adriana Vlad, Adrian Mitrea* 207

Dezambiguizarea automată a cuvintelor din corpusuri paralele folosind echivalenții de traducere - *Dan Tufiş* 235

Referențialitate şi cursivitate în relație cu structura de discurs - *DanCristea* 269

DLIR - un sistem de căutare documentară multilingv - *Amalia Todiraşcu* 303

Mediu hermenofor pentru asistarea învățării unor concepte dintr-o limbă străină - *Ştefan Trăuşan-Matu* 317

SECȚIUNEA III:
TEHNOLOGII ALE LIMBAJULUI VORBIT

Experimente în vederea recunoaşterii vorbitorului - *Corneliu Burileanu, Luigi Bojan* 335

Prelucrarea inițială a textului de intrare în cadrul unui sistem de sinteză a vorbirii pornind de la text în limba română - *Dragoş Burileanu* 359

Utilizarea tehnicilor nuanțate (fuzzy) şi de dinamică neliniară pentru sinteza adaptivă a vorbirii - *Horia-Nicolai L. Teodorescu* 381

Dicționarele multimedia ale limbii române. Secvențe de implementări şi experimentări - *Dumitru Todoroi, Diana Micusa, Zinaida Todoroi, Ion Lingă, Ion Covalenco, Nicolae Objeleanu, Ştefan Spătaru, Stela Lungu, Virginia Țurcanu, Elena Cozlov, Nadejda Ambrozii, Victor Slobodeanu, Igor Coşeru, Cătălina Suruceanu* 401

Mediu pentru editarea transcrierilor fonetice în limba română. Realizarea atlasului lingvistic român pe regiuni - *Silviu Bejinariu, Vasile Apopei, Mariana Roman* 423

SECȚIUNEA IV:
DEZBATERI ŞI DISCUȚII

Asupra a doi vectori funcționali ai societății cunoaşterii: managementul cunoaşterii şi învățarea electronică. Cultura şi societatea cunoaşterii - *Mihai Drăgănescu* 441

între lingvistica matematică şi cea computațională - *Solomon Marcus* 471

între lingvistica matematică şi cea computațională: o altă perspectivă - *Dan Tufiş* 481

INTRODUCERE

Programul de cercetare aplicativă "Strategii şi soluții pentru Societa Informațională - Societatea Cunoaşterii în România (SI-SC), din subprogramul strategic, al Programului Național INFOSOC a avut ca principale obiective stabili unui program de veghe conceptuală pentru menținerea pe linia tendințelor mondi ale avansului SI-SC, sensibilizarea factorilor de decizie şi a publicului larg, crea unui cadru de reflecție prospectivă pe temele prioritare ale SI-SC: economi sociale, culturale, tehnologice, ambientale, precum şi operaționalizarea unor sol de interes prioritar pe plan național. în cadrul acestui proiect a fost elaborat volu *"Societatea Informațională - Societatea Cunoaşterii. Concepte, soluții şi strat pentru România"* (publicat la Ed. Expert in anul 2000), realizat sub coordona Academicianului Florin Gheorghe Filip. Acest volum avea ca scop construirea u viziuni si conținea o serie de studii şi cercetări care au aprofundat rezultat programului prioritar al Academiei Române privind *Societatea Informațională Societatea Cunoaşterii* şi au identificat o serie de orientări strategice cerute susținerea unei dezvoltări de tip "salt" a SI-SC în România. Prin prisma obiective proiectului, au fost analizate principalele aspecte conceptuale ale SI-SC, proble legate de infrastructurile informatice şi de comunicații ale SI-SC, forma profesională şi pregătirea generală a populației în şi pentru SI-SC, rolul ştiin cercetării şi inovării, aspecte sociale şi juridice, instituțiile statului şi relația lor cetățeanul, dezvoltarea economiei şi afacerilor, dimensiunea culturală a SI-S actorii sociali ai creării şi difuzării tehnologiei informației şi comunicațiilor în contex SI-SC. Studiile tematice, ancheta Delphi pentru consultarea opiniei experților priv tendințele globale şi opțiunile posibile de raportare la ele, scenariile de evolu elaborate au susținut funcția prospectivă a proiectului.

Funcția operativă a acestui proiect, respectiv identificarea de soluții tehn privind rezolvarea principalelor priorități identificate în faza analizei prospect urma sa se manifeste în perioada imediat următoare, printr-o dintr-o serie cercetări/dezvoltări tehnologice ce vor trata pe larg problematica specifică fiecăruia dintre direcțiile amintite anterior. Această serie este deschisă p prezentul volum ce înglobează contribuții ale unor specialişti români reprezenta în domeniul prelucrării automate a limbajului natural şi a resurselor lingvist necesare utilizării limbii române în mediile de comunicare electronică.

în [1] este definit conceptul de "Societate Informațională - Societate Cunoaşterii" (SI-SC) precum şi principalii săi vectori tehnologici şi funcționali. acest context "internetul dezvoltat" (ca vector tehnologic) şi "managemen

utilizării morale a cunoaşterii la nivel global" (ca vector funcțional) sunt prezentați ca factori motrici esențiali ai Societății Cunoaşterii, şi în perspectivă, a Societății Conştiinței. "Din momentul în care intervine Internetul cu marile avantaje pe care acesta le aduce (e-mail, comerț electronic şi tranzacții electronice, piața Internet, distribuția de 'conținut') prin cuprinderea în sfera informației electronice a unui număr cât mai mare de cetățeni se trece la societatea informațională. Cunoaşterea este informație cu înțeles şi informație care acționează. De aceea societatea cunoaşterii nu este posibilă decât grefată pe societatea informațională şi nu poate fi separată de aceasta. în acelaşi timp, ea este mai mult decât societatea informațională prin rolul major care revine informației-cunoaştere în societate." [1]

în 1984, William Gibson, un dizident cognitiv - după cum se auto-caracterizează, publică volumul SF *"Neuromancef* (Ace Book, July 1984, ISBN: 0-441-56959-5), carte care pe lângă o mulțime de premii literare i-a adus notorietatea şi pentru crearea termenului "cyberspace": "the total interconnectedness of human beings through computers and telecommunication without regard to physical geography... A consensual hallucination experienced daily by billions of legitimate operators, in every nation, by children learning mathematical concepts...a graphical representation of data abstracted from the banks of every computer in the human system. Unthinkable complexity. Lines of light ranged in the nonspace of the mind. Clusters and constellations of data. Like city lights receding..." (op. cit).

Termenul a făcut carieră, actualmente fiind o noțiune care din punct de vedere tehnic subsuma conceptul "Internet"(scris cu majusculă): "cyberspace: The impression of space and community formed by computers, computer networks, and their users; the virtual "world" that Internet users inhabit when they are online

The term internet (spelled with a lower case T) is distinguished from the Internet (spelled with the "I" capitalized). The Internet refers to a specific, historic, ubiquitous worldwide digital communication network." (cf. Glossary of Telecommunications, American National Standard T1.523-2001, www.atis.org/ ta2k/ cvberspace.html, 05.08.2002).

Dimensiunea tehnică (evocată mai sus) a noțiunii de *"ciberspațiu"* este complementată de dimensiunea socio-culturală şi din această perspectivă de problemele "satului global" previzionat de Societatea Informațională - Societatea Cunoaşterii. Idea atenuării schismei dintre specialiştii din domeniul tehnic şi cei din zona ştiințelor umaniste în contextul SI-SC este susținută puternic şi de M. Derouzos [5] , cel care a propus conceptul de "piața informațională", pe care îl consideră mai realist decât cel de "ciberspațiu". De altfel, dimensiunile socio-culturale ale SI-SC au fost evocate în capitolele 2, 3, 4 şi 6 ale volumului *"Societatea Informațională - Societatea Cunoaşterii. Concepte, soluții şi strategii pentru România"*

Printre componentele socio-culturale ale SI-SC, utilizarea limbii materne în mediile de lucru şi comunicare electronice a accesului universal la ciberspațiu [2, 3, 4] constitue priorități absolute.

în contextul actual, al comunicării mediate de tehnologia informației şi telecomunicații, limba devine obiect al investigației tehnice. Tehnologia limbaju impune metodologii specifice de cercetare/dezvoltare, dezvoltarea sau adaptar resurselor lingvistice fundamentale cum ar fi dicționarele, tezaurele, corpusurile gramaticile computerizate, în conformitate cu standardele sau recomandă existente. în funcție de resursele lingvistice disponibile, de volumul şi calitatea de compatibilitatea codificării lor în raport cu recomandările şi standard internaționale etc, se poate vorbi de *nivelul de tehnologizare* al unei limbi natura Nivelul de tehnologizare al unei limbi naturale este în corespondență directă statutul de limbă de *circulație electronică.* Această sintagmă, o parafrază expresia *limbă de circulație internațională,* încearcă să elimine antinomia, pe de cunoscută pe atât de goală în conținut spiritual şi cultural, "limbi mari/limbi mi Conceptul de "limbă de circulație electronică", pe lângă semnificația lui directă, profunde implicații culturale, sociale şi nu în ultimul rând economice implică dreptul fiecărui cetățean de a avea acces în propria limbă la cunoştințe informațiile şi serviciile ciberspațiului.

Promovarea limbii române în SI-SC presupune informatizarea limbii româ ca factor infrastructural fundamental (vector funcțional) precum şi stimularea utiliz curente (prin vectori tehnologici) a limbii române în utilizarea tehnologiilor şi serviciilor informatice. Acest obiectiv presupune eforturi umane şi materi substanțiale şi de dimensionarea lor se leagă orizontul de timp al realizării sale.

Volumul de față reuneşte lucrări ce tratează aspecte specifice prelucr limbajului natural, în marea lor majoritate cu aplecare directă asupra limbii româ Inerent, volumul de față nu poate acoperi întreaga arie problematică a domeniu după cum nici reprezentarea specialiştilor români în domeniul tehnologiei limbaju nu este completă, dar cititorul va găsi un larg evantai de direcții de cercetare, care specialiştii români au obținut rezultate importante.

Volumul este structurat în patru părți (aspecte teoretice şi probleme terminologie, prelucrarea limbajului scris, prelucrarea limbajului vorbit, dezbateri discuții) care pot fi citite în mod independent, în funcție de interesul specific cititorului.

Prima parte "Lingvistică teoretică şi formală; terminologie" cuprinde lucrări din domeniul lexicografiei, sintaxei şi terminologiei.

în lucrarea "Resurse lingvistice elaborate la Institutul de Lingvistică *«lo lordan»"* Ioana Vintilă Rădulescu face o trecere în revistă a celor mai importa resurse lingvistice realizate în cei peste 50 de ani de activitate la Institutul Lingvistică *«lorgu lordan».*

Angela Bidu-Vrânceanu prezintă în lucrarea "Contribuția lingvisticii studiul terminologiilor ştiințifice" concluziile a trei contracte de cercetare ştiințif

având ca obiect studiul terminologic al limbajului folosit în diverse domenii (matematică, filozofie, mineralogie, arte plastice).

Articolul "Gramaticile nontransformaționale" al lui Emil Ionescu face o prezentare generală a gramaticilor bazate pe unificare şi constrângeri precum şi a principalelor realizări, în contextul acestei paradigme, în cercetarea lingvistică din România.

Neculai Curteanu propune în lucrarea "Către o teorie X-bar funcțională" o reconsiderare a teoriei clasice X-bar prin perspectiva modelului propriu SCD (Segmentare-Coeziune-Dependență).

Ana-Maria Barbu prezintă în lucrarea sa "Teoria HPSG: studiu de caz: acordul încrucişat" principalei caracteristice ale teoriei HPSG şi discută în acest context un caz de dependență încrucişată specific limbii române, respectiv clauzele relative în care pronumele relativ este precedat de articolul genitival.

O serie de probleme legate de terminologia computațională sunt prezentate în ultimele două lucrări ale primei secțiuni. în articolul" După 10 ani de experiență terminologică: noul model de date terminologice al TERMROM" Dan Matei prezintă modelul dezvoltat în conformitate cu noile tendințe şi standarde în domeniu şi adoptat de Asociația Română de Terminologie - TERMROM.

Lucrarea lui Sorin Gețaru "Probleme de reprezentare a datelor terminografice într-o bază de date relațională" aduce în discuție aspecte specifice reprezentărilor standardizate necesare realizării dezideratului de interschimb şi interoperabilitate între diverse tezaure terminologice şi discută elementele distinctive ale standardului ISO-12200 MARTIF (Machine-Readable Terminology Interchange Format).

Secțiunea a doua a volumului ("Tehnologii ale limbajului scris") este deschisă de lucrarea lui Dan Tufiş şi Dan Cristea "RO-BALKANET - ontologie lexicalizată în context multilingv pentru limba română" care descrie stadiul dezvoltării unui dicționar, pentru limba română, structurat ca o rețea semantică, de tip EuroWordNet, rezultat al unui program european ce-şi propune extensia EuroWordnet (în prezent implementat pentru 10 limbi europene) cu încă 5 limbi.

Articolul lui Dan Gâlea, Neculai Curteanu şi Cristian Linteş "Algoritmi de segmentare a textului în unități de tip clauzal" tratează o problemă delicată a prelucrării limbajului natural, respectiv cea a identificării, în raport cu un anumit criteriu funcțional, a structurilor "clauzale" şi prezintă contrastiv doi algoritmi diferiți (unul dintre ei aparținând autorilor), atât prin prisma modelării lingvistice cât şi al performanței computaționale.

Rada Mihalcea şi Vivi Năstase prezintă în articolul lor o metodă de inserare automată a caracterelor diacritice în texte scrise (cu studiu de caz pentru limba română) fără diacritice şi comentează rezultatele proprii în comparație cele ale altor metode dezvoltate pentru rezolvarea aceleiaşi probleme.

Adriana Vlad şi Adrian Mitrea prezintă în lucrarea lor "Contribuții pri structura statistică de cuvinte în limba română scrisă" rezultate recente caracterizarea statistică a limbii române scrise, prin aproximarea ei ca un Markov ergotic multiplu cu ordin de multiplicitate mai mare decât 30, rezul obținute prin analiza riguroasă a unui corpus foarte mare de texte.

Articolul "Dezambiguizarea semantică automată în corpusuri paralele lui Dan Tufiş prezintă o alternativă la spinoasa problemă a dezambiguiz cuvintelor polisemantice, bazâdu-se pe extragerea cunoştințele implicite existe într-un corpus multilingv (creat de traducători profesionişti) şi apelând la tehnic euristici ale lingvisticii corpusului.

Dan Cristea prezintă în articolul "Referențialitate şi cursivitate în struct discursului" elementele definitorii ale teoriei sale asupra structurii discursiv textelor (teoria nervurilor) şi îşi exemplifică argumentația prin analiza dihotom structură-referențialitate şi structură-coerență.

în lucrarea "DLIR - un sistem de căutare documentară multilingv" Am Todiraşcu prezintă o abordare bazată pe logici terminologice, ontologii şi tehnic prelucrare a corpusurilor în implementarea unui sistem de regăsire document bilingv (română şi franceză).

Partea a doua a volumului se încheie cu articolul lui Ştefan Trăuşan-M "Mediu hermenofor pentru asistarea învățării unor concepte într-o limbă străi care după o prezentare a noțiunilor cu care operează în lucrare, descrie un mo de prelucrare a metaforelor utilizate în limbaje specializate (studiu de caz: limb financiar) incorporat într-un sistem de instruire inteligentă în învățarea concepte într-o limbă străină, sistem distribuit dezvoltat în cadrul unui proiect european.

Secțiunea a treia a volumului este dedicată problemelor de prelucrar vorbirii. Corneliu Burileanu şi Luigi Bojan se opresc asupra tehnicilor recunoaştere a vorbitorului ca etapă distinctă şi strict necesară pen recunoaşterea automată a vorbirii şi prezintă o parte a rezultatelor obținute către autori.

Lucrarea lui Dragoş Burileanu "Prelucrarea inițială a textului de intrare cadrul unui sistem de sinteză a vorbirii pornind de la text în limba româ abordează problemele sintezei limbajului vorbit pornind de la un text în forn electronic şi detaliază etapa de preprocesare a textului ca etapă primară procesul transformării sale în semnal vocal inteligibil şi coerent.

Tot în domeniul sintezei vorbirii se plasează şi lucrarea lui Horia Nico Teodorescu "Utilizarea tehnicilor nuanțate (fuzzy) şi de dinamică neliniară pen

sinteza adaptivă a vorbirii" ce subliniază rolul esențial al prozodiei şi al modelării sale algoritmice în realizarea unor sinteze vocale de calitate, purtătoare de informație emoțională.

Un proiect de anvergură, este prezentat de Dumitru Todoroi, Diana Micusa, Zinaida Todoroi, Ion Lingă, Ion Covalenco, Nicolae Objeleanu, Ştefan Spătaru, Stela Lungu, Virginia Turcanu, Elana Cozlov, Nadejda Ambrozii, Victor Slobodeanu, Igor Coşeru şi Cătălina Suruceanu în lucrarea "Dicționarele multimedia ale limbii române. Secvențe de implementări şi experimentări".

Secțiunea a treia a volumului se încheie cu lucrarea elaborată de Silviu Bejinariu, Vasile Apopei şi Mariana Roman "Mediu pentru editarea transcrierilor fonetice în Limba Română. Realizarea Atlasului Lingvistic Român pe Regiuni" ce prezintă un instrument ce permite realizarea facilă a transcrierilor fonetice într-un limbaj standardizat (IPA), oferă extensii specifice de adnotare fonetică (realizate până acum manual) şi prefigurează realizarea variantei computerizate a atlaselor lingvistice româneşti.

Ultima secțiune a volumului (Dezbateri şi discuții) conține trei contribuții. Prima dintre ele, elaborată de Mihai Drăgănescu, "Asupra a doi vectori funcționali ai Societății Cunoaşterii: Managementul Cunoaşterii şi învățarea Electronică. Cultura şi Societatea Cunoaşterii" reprezintă liantul dintre volumul precedent *(Societatea Informațională - Societatea Cunoaşterii. Concepte, soluții şi strategii pentru România,* coordonator Fl. Gh. Filip) şi volumul de față, rafinând clasificarea din lucrarea anterioară şi adâncind o serie de probleme ridicate în [1].

Ultimele două contribuții reprezintă două puncte de vedere asupra problematicii prelucrării limbajului natural, prima poziție "între lingvistica matematică şi cea computațională" fiind susținută de Solomon Marcus, iar cea de a doua "între lingvistica matematică şi cea computațională: o altă perspectivă" fiind prezentată de Dan Tufiş.

Mulțumiri

Coordonatorii acestui volum, mulțumesc tuturor celor care au participat la realizarea proiectului "Strategii şi soluții pentru societatea informațională-societatea cunoaşterii în România" derulat cadrul programului național INFOSOC. Mulțumiri speciale se cuvin directorului programului INFOSOC, Profesor Doina Banciu, care a susținut şi a manifestat un interes deosebit față de desfăşurarea acestui proiect.

Referințe bibliografice

[1] M. Drăgănescu "Societatea informațională şi a cunoaşterii. Vectorii socie cunoaşterii" în *F.G. Filip (coord.) Societatea Informațională - Societ Cunoaşterii. Concepte, soluții şi strategii pentru România.* Academia Romă Editura Expert, ISBN 973-8177-42-1, 2001, pp. 43-112

[2] *** The Multilingual Information Society, Report of Commission of the Europ Communities, COM(95) 486/final, Brusseîs, November 1995.

[3] *** Multilingualism in an Information Society, International Sympos organized by EC/DGXIII, UNESCO and Ministry of Foreign Affairs of the Fre Government, Paris 4-6 December 1997.

[4] *** Promotion and Use of Multilingualism and Universal Access to Cyberspa UNESCO 31st session, November 2001.

[5] M. Dertouzos. "What It will Be". Harper Edge. New York,1997 (trad. în română "Ce va fi", Ed. Tehnică, Bucureşti, 2000).

Secțiunea I

LINGVISTICA TEORETICA SI FORMALĂ; TERMINOLOGIE

Resurse lingvistice pentru limba română elaborate la Institutul de Lingvistică "Iorgu Iordan"

Ioana VINTILĂ-RĂDULESCU
Institutul de Lingvistică "Iorgu Iordan - Al. Rosetti"
Bucureşti, Calea 13 Septembrie 13
e-mail: ioanar@fx.ro

1. Consideraţii generale

Numind *resursă* în general o "rezervă sau sursă de mijloace (materiale sau spirituale) susceptibile de a fi valorificate într-o împrejurare dată"[1], înţelegem prin *resurse lingvistice* pentru limba română izvoarele fundamentale de informaţii cu privire la aceasta, stocate convenabil (chiar dacă încă preponderent în manieră tradiţională) şi care, în calitate de componente ale *culturii* în sensul cel mai larg, sunt susceptibile de a fi valorificate pentru studierea limbii române, precum şi în diverse scopuri conexe, inclusiv aplicative, în cadrul *societăţii informatice* actuale.

Cât priveşte Institutul de Lingvistică "Iorgu Iordan"[2], acesta nu mai există formal ca atare, deoarece la începutul anului 2002, printr-o hotărâre de guvern adoptată la propunerea conducerii Academiei Române, s-a produs re-unirea sa şi a Institutului de Fonetică şi Dialectologie "Al. Rosetti". (Spunem reunire întrucât cercetările de fonetică şi de dialectologie formaseră iniţial obiectul unui sector, respectiv al unei secţii a Institutului de Lingvistică din Bucureşti al Academiei Române (înfiinţat în 1949), devenită din 1961 centru şi apoi institut independent.) întrucât în 1998 fusese oficializată, tot prin hotărâre de guvern, propunerea celor două institute, aprobată de Prezidiul Academiei, de a-şi adăuga fiecare în titulatură numele fostului său director, institutul în cadrul căruia cele două nuclee care au fuzionat acum îşi continuă de fapt activitatea poartă numele dublu de Institutul de Lingvistică "Iorgu Iordan - Al. Rosetti".

Fără îndoială, cele mai numeroase şi mai importante resurse lingvistice pentru limba română s-au realizat la acum fostul Institut de Lingvistică "Iorgu

[1] **** (1975).* Dicţionarul limbii române (DLR). Serie nouă. Tomul IX, Litera R, *Bucureşti,* s.v.
[2] *Pentru o imagine de ansamblu asupra activităţii acestui institut şi a istoriei sale v. Mioara Avram, Marius Sala, Ioana Vintilă-Rădulescu (coordonatori) (1999).* Institutul de Lingvistică "Iorgu Iordan". 50 de ani de existenţă (1949-1999), *Bucureşti.*

Iordan", înglobând, până în 1961 direct şi apoi numai indirect, şi contribuţia colegilor foneticieni şi dialectologi[3], precum şi, în unele cazuri, în colaborare cu alte institute de specialitate din ţară ale Academiei - Institutul de Lingvistică şi Istorie Literară "Sextil Puşcariu" din Cluj şi Institutul de Filologie Română "Alexandru Philippide" din Iaşi - şi cu cadre didactice de la facultăţile de profil mai ales ale Universităţii din Bucureşti. Această activitate este continuată şi în noul cadru organizatoric de sectoarele fostului institut, pe care în cele ce urmează îl vom numi, pe scurt, *Institutul.*

2. Resurse lexicografice

Dintre resursele lingvistice tradiţionale dezvoltate până în prezent de Institut, cele mai importante din punctul de vedere care interesează aici sunt cele **lexicografice - dicţionarele** (mono- şi bilingve) -, activitatea lexicografică din Institut, începută încă de la înfiinţarea sa, desfăşurându-se din 1959 în cadrul unui sector specializat cu acest profil, condus până în 1985 de Mircea Seche, iar de atunci încoace de Ion Dănăilă[4].

2.1. Dicţionare monolingve

2.1.1. Dintre dicţionarele româneşti monolingve se distinge, prin anumite trăsături ale sale, dicţionarul "explicativ general academic" intitulat pur şi simplu ***Dicţionarul limbii române -*** dar mai cunoscut ca "Dicţionarul Academiei" a cărui realizare se apropie de sfârşit şi care va cuprinde o mare parte a "tezaurului" lexical al limbii române - fără a putea şi nici a intenţiona să includă însă ansamblul cuvintelor româneşti folosite în toate epocile, în toate regiunile şi în toate domeniile[5]. în ciuda marilor sale calităţi, care sunt bine cunoscute şi asupra cărora nu credem deci că mai este nevoie să insistăm aici, acest dicţionar prezintă un dezavantaj major din punctul de vedere al utilizării sale ca resursă de bază (pe lângă faptul că nu se prezintă şi sub forma unei variante electronice, care nici nu putea fi imaginată până nu de mult) şi anume caracterul său fatalmente neunitar, datorat faptului că a fost elaborat pe parcursul a aproape un secol[6], de unde mai deosebiri dintre cele două părţi ale sale: cea publicată între 1907 şi 1949 s conducerea marelui lingvist Sextil Puşcariu şi cea care a început să apară 1965 şi a cărei publicare se apropie, în fine, de sfârşit. "Seria veche" a dicţionaru academic, desemnat de aceea prin sigla DA, cuprinde literele *A-C* (inclu puţinele neologisme scrise acum cu *k-,* iar în DA cu c/?-) şi *F-J* complet, iar liter D şi *L* parţial (până la cuvântul de, respectiv *lojniţă),* totalizând 3.142 de pagini tipar, format mare, dintre ele lipsind în întregime, după cum se observă, litera Această primă jumătate a dicţionarului se distinge prin lista de cuvinte, bogată r ales sub aspectul fondului tradiţional, prin tratarea amănunţită a semantismul bazată pe numeroase citate, prin dimensiunile şi valoarea comentariu etimologic, precum şi prin traducerea sensurilor în limba franceză[7]. Desigur, aveau cum figura în aceste prime volume numeroasele neologisme încetăţenite româneşte după elaborarea lor, ilustrarea sensurilor prin utilizarea lor de că autori mai noi şi în general toate aspectele care sunt rodul evoluţiei ulterioare limbii române, al cercetărilor dialectale, etimologice, filologice etc. mai recente şi dezvoltării lingvisticii şi metodelor ei, în general. Din 1965 dicţionarul şi-a reîncep apariţia, în format asemănător, ca *Serie nouă* (de data aceasta sub o siglă diferi menţionată în titlu, DLR), cu litera *M,* sub conducerea, la început, a lui Ion Iordan, Alexandru Graur şi Ion Coteanu, iar actualmente a lui Gh. Mihăilă şi Mar Sala. Noua serie păstrează, în mare, principiile lui Sextil Puşcariu, dar beneficia de toate avantajele elaborării sale mai aproape de zilele noastre: ea inclu modificări şi amplificări reflectând evoluţia limbii române, a lexicografiei române şi a studiului limbii române în ansamblu, precum şi a lingvisticii în general, dar mai cuprinde, în schimb, traducerea sensurilor (în anii '60 nefiind conside oportun acest lucru, deşi era util mai ales pentru cunoaşterea limbii române către străini, fără a fi, este drept, uzual într-un dicţionar monolingv explicativ), secţiunea etimologică a fost redusă, dicţionarul păstrându-şi însă caracterul isto (sensurile sunt date în ordinea atestării lor în texte şi în alte surse)[8]. Institu bucureştean a redactat literele *M, N, P, S* şi *Z*[9] şi este pe cale de a înch

[3] *Aceştia au produs mai ales "resurse" de un tip specializat, concretizate în principal în atlase lingvistice şi în arhiva fonogramică a limbii române, de care nu ne vom ocupa în mod direct aici, dar care, ca şi contribuţiile similare ale altor institute, au avut şi un aport indirect la resursele fundamentale despre care vorbim, printre izvoarele cărora s-au numărat*

[4] *Pentru detalii cu privire la lucrările acestuia v. Ion Dănăilă (1999).* Sectorul de lexicologie şi lexicografie, *în Mioara Avram, Marius Sala, Ioana Vintilă-Rădulescu, op. cit, p. 98-113.*

[5] *Ideea, relativ utopică şi controversată, a înregistrării şi chiar a descrierii semantice a întregului inventar lexical al limbii române (ILEX) din toate timpurile, incluzând atât numele comune, cât şi cele proprii (v. Ion Dănăilă (1993).* Pentru un inventar general al limbii române, *în "Limba română" XLII, nr. 2, p. 61-68), nici nu a început a fi pusă în practică.*

V., printre altele, Marius Sala (1999). Institutul de Lingvistică "Iorgu Iordan" la 50 de ani *Mioara Avram, Marius Sala, Ioana Vintilă-Rădulescu, op. cit, p. 35-37.*

[7] *Pentru o descriere amănunţită a DA v. Mircea Seche (1969).* Activitatea lexicografică a Sextil Puşcariu, *în* Schiţă de istorie a lexicografiei române, *voi. II,* De la 1880 până asta *Bucureşti, p. 42-72.*

[8] *V. şi Mircea Seche (1969).* Seria nouă a Dicţionarului academic general *în* Schiţă istorie a lexicografiei române, *voi. II,* De la 1880 până astăzi, *Bucureşti, p. 72-79.*
Iorgu Iordan, Al. Graur, I. Coteanu (red. resp.) et al. (1965-2000). *Dicţionarul limbii roma (DLR). Serie nouă,* Bucureşti: T. VI, *Litera M,* 1965-1968 (apărut iniţial în fascicule); Partea 1, *Litera N,* 1971; Partea a 2-a, *Litera* O, *1969]* VIII, *Litera P.'*Partea 1, *P-PĂZ* 1972; Partea a 2-a, PE-PÎNAR, 1974; Partea a 3-a, *PÎNĂ-POGRIBANIE,* 1977; Partea 4-a, *POGRIJENIE-PRESIMŢIRE,* 1980; Partea a 5-a, *PRESIN-PUZZOLANĂ,* 1984; *Litera R,* 1975; X. *Litera S.* Partea 1, *S-SCLĂBUC,* 1986; Partea a 2-a, *SCLAD SEMÎNŢĂRIE,* 1987; Partea a 3-a, *SEMN-SÎVEICĂ,* 1990; Partea a 4-a,[%] *SLA*

reluarea şi terminarea literei *D* absentă din prima parte (trei volume); numai primele patru litere elaborate la Bucureşti însumează 51.847 de cuvinte şi variante, totalizând 5.839 p. Institutului din Cluj i-au revenit literele O, *R, T, Ţ* (totalizând 2.044 de pagini de tipar), *U* (aflată sub tipar) şi, din prima parte, reluarea şi terminarea unei părţi din litera L, iar celui din Iaşi - literele *Ş, V* (Ş şi prima parte din cele trei ale literei *V* - singura dintre acestea apărută până acum - totalizând 599 de pagini de tipar), IV, *X,* Y, precum şi, din prima parte, elaborarea literei lipsă £ şi reluarea şi terminarea unei părţi din litera *L*\ pentru etimologii au fost consultaţi specialişti din mai multe centre universitare. Majoritatea literelor au apărut, unele pe sărite *(M* între 1965 şi 1968, *N* în 1971, O în 1969, P între 1972 şi 1984, *R* în 1975, S între 1986 şi 1994, Ş în 1978, *T* în 1983, Tîn 1994, prima parte din *V* (până la a *veni)* în 1997 şi Zîn 2000) - în total 20 de volume -, cu excepţia literelor *D, E, K, L, U,* a puţinelor cuvinte începând cu litera Q şi a ultimelor părţi ale literei *V* (începând cu *venin),* la care se adaugă literele *W, X* şi Y Deosebirea cea mai importantă constă în tipurile de cuvinte reprezentate în cele două serii: la majoritatea primelor litere ale alfabetului (cu excepţia celor care s-au redactat abia acum), neologismele sunt slab reprezentate, nu numai din cauza faptului că foarte multe nici nu se încetăţeniseră încă în limba română la vremea elaborării volumelor respective, dar şi din cauza reticenţei lui Puşcariu cu privire la acest sector al vocabularului; într-o situaţie asemănătoare se află termenii regionali, deoarece cercetările dialectale se aflau în acea vreme abia la început. Prima parte prezintă în schimb avantajul de a putea servi ca bază pentru o prelucrare bilingvă, întrucât includea şi traducerea sensurilor în limba franceză, la care a trebuit să se renunţe în perioada comunistă. Reluarea şi completarea acestui dicţionar, absolut necesară, nu ni se mai pare astăzi recomiandabil şi nici posibil de realizat prin mijloace tradiţionale (fişe etc), ci exclusiv pe baze informatizate. Ea ar trebui să valorifice, printre altele, şi banca de texte şi cea de inovaţii a limbii române, despre care va fi vorba mai departe. Ar fi necesar ca partea publicată înainte de 1949 să fie reluată şi adusă la zi, cu atât mai mult cu cât puţine persoane şi chiar biblioteci posedă dicţionarul în întregime (chiar în cazul seriei noi, tirajele diverselor litere au fost diferite şi în continuă scădere), iar îmbătrânirea hârtiei în cazul seriei vechi o face fragilă şi greu de consultat. Având în vedere că pentru noua serie a dicţionarului s-au adunat, manual, peste şase milioane de fişe cu extrase şi atestări (dintre acestea, în DLR au fost incluse cea 3.200.000 de citate[10], reprezentând aproximativ 88% din totalul textului), este de sperat că la reluarea, într-un viitor mai mult sau mai puţin apropiat, a primei serii se va putea uza de

SPONGHIOS, 1992; Partea a 5-a, *SPONGIAR-SWING,* 1994; XI Partea 1, *Litera* Ş, 1978; Partea a 2-a, *Litera T, T-TOCĂLIŢĂ,* 1982; *Partea a* 3-a, *TOCĂNA-TWIST,* 1983; XII, Partea 1, *Litera Ţ,* 1994; ***XIII,*** *Partea 1, Litera V, V-VENI,* 1997; XIV, *Litera* Z, 2000.

[10] *în legătură cu reflectarea noilor norme ortografice ale limbii române în volumele DLR elaborate după 1993, semnalăm faptul că forma* sânt, *reflectând un fonetism real, vechi şi popular, este păstrată în citatele în care nu era folosit* sunt.

avantajele elaborării computerizate, valorificându-se băncile de date în curs elaborare în institut.

Având în vedere diferenţele semnalate (dintre care unele se regăsesc între primele şi ultimele litere din seria nouă), este foarte binevenită ideea actuali responsabili ai DLR de a se publica, pentru operativitate, un *Supliment* - care poate realiza relativ mai lesne - "care să înregistreze neologismele adoptate limba literară de la începutul secolului" 20 "până în prezent, precum şi o serie cuvinte regionale incluse în atlasele lingvistice şi în culegeri de pe teren s termeni vechi extraşi din documente ale secolelor al XVI-lea - al XVIII-lea, edita în ultimele decenii"[11].

2.1.2. Din motivele expuse mai sus, la care se adaugă şi faptul că DA/DL este accesibil mai ales specialiştilor şi mai puţin publicului larg, institu bucureştean pregăteşte între timp, la sugestia conducerii Academiei Române, sinteză a marelui dicţionar academic, fără citate şi izvoare şi cu un sistem foa economic de prezentare a informaţiilor lexicografice. Acest ***Mic dicţion academic*** (MDA)[12] (care va avea totuşi patru volume), inclus, alături de DL printre lucrările fundamentale ale Academiei Române, va avea cea 175 000 intrări (cea 125.000 de cuvinte şi cea 50.000 de variante); primul volum (A-C) a fo publicat în anul 2001 de editura Univers Enciclopedic. Proiectul *Micului dicţion academic,* numit astfel în opoziţie cu "marele" dicţionar academic, şi-a propus reducă decalajul dintre cele două serii ale acestuia, îmbogăţind primele litere baza unor surse lexicografice mai noi. La rândul său, acest nou dicţionar prezir însă dezavantajul de a fi fost obligat, prin dimensiuni, să renunţe la citate ilustrative, ceea ce limitează posibilitatea utilizării lui ca sursă de informa morfologice, gramaticale şi stilistice; numărul neobişnuit de mare de abrevi netransparente, utilizate din acelaşi motiv de economie, constituie un argume suplimentar în favoarea realizării unei versiuni electronice a MDA care să perm regăsirea automată a informaţiilor.

2.1.3. Spre deosebire de DA/DLR, o reflectare în general unitară vocabularului limbii române oferă ***Dicţionarul explicativ al limbii român*** despre a cărui siglă, DEX, se afirmă, pe drept cuvânt, că a devenit un apelat denumirea, care ar fi trebuit protejată prin înregistrare, a fost preluată abuziv *Noul dicţionar explicativ al limbii române* publicat pe CD-Rom de firmele Litera sigla NODEX, sugerând că ar fi "un nou DEX". Prima ediţie, un volum de 1.049 pagini, cuprinzând 56.569 de cuvinte şi variante, a fost urmată de un ***Suplimen***

[11] *Marius Sala, G. Mihăilă (2000).* Cuvânt înainte, *în* Dicţionarul limbii române (DLR). Se ^nouă. Tomul XIV. Litera Z, *Bucureşti, p. VI.*

V. /. Dănăilă (1994). De ce este nevoie de un MDA?, *în "Limba română" XLIII, 9-10, 397-406 şi Marius Sala (2001).* Prefaţa, *în* Micul dicţionar academic (MDA), *voi. I,* A- ^*Bucureşti.*

I. Coteanu, Luiza Seche, M. Seche (conducătorii lucrării) et al. (1975). Dicţiona explicativ al limbii române *(DEX), Bucureşti.*

Dicționarul explicativ al limbii române *(DEX-S)*[14]. Ediția a doua a DEX[15] totalizează 1.204 pagini; această ediție, care se publică în continuare în tiraje succesive, totalizase numai în primii patru ani de la apariție 65.000 de exemplare vândute, după un calcul sumar rezultând că la 42 de locuitori ai României revenea un *DEX*. Actualmente, se poate într-adevăr afirma că, prin **DEX,** *best-sellerul* lingvisticii româneşti, Institutul a intrat în marea majoritate a caselor din România. Se preconizează ca DEX să fie realizat, în fine, într-un viitor relativ apropiat, şi în format electronic. El a fost deja supus, de către Centrul de Cercetări Avansate în învățarea Automată, Prelucrarea Limbajului Natural şi Modelarea Conceptuală al Academiei Române, codificării conform TEI[16]. Se estimează că ediția a III-a a DEX, concepută sub conducerea lui Ion Dănăilă, va avea în plus față de precedenta cea 30.000 de cuvinte. Sub conducerea lui Ion Coteanu şi Ion Dănăilă, la sectorul de specialitate al Institutului a fost conceput şi un ***Nou dicționar explicativ al limbii române*** (NEX), cu caracteristici diferite de cele ale **DEX:** inventar de cea 100.000 de cuvinte şi variante (deci aproape de două ori mai multe decât prima ediție a DEX), definiții mai concise, prin eliminarea sinonimelor şi - din păcate!-, neincluderea etimologiei cuvintelor; revizuit de cei doi responsabili, el aşteaptă introducerea în calculator, în vederea efectuării corelațiilor semantice definiționale şi sinonimice.

2.1.4. DEX a scos practic din circulație dicționarele explicative mai vechi, limitate la limba română literară, DLRLC şi DM[17]. Prima siglă reprezintă ***Dicționarul limbii române literare contemporane***[18]**,** elaborat de institutele din Bucureşti şi Cluj pornind de la "baza manuscrisă" a DA şi apărut între 1955 şi 1957 în patru volume. El se mai foloseşte şi astăzi - deşi din el lipsesc cuvintele, sensurile şi citatele neconforme cu ideologia vremii - pentru citatele cu care, spre deosebire de dicționarele de dimensiuni comparabile mai noi, sunt ilustrate sensurile cuvintelor (chiar dacă, pentru unele neologisme, citatele provin, aşa cum era obligatoriu în epocă, din traducerile "operelor clasicilor" marxism-leninismului!). El mai merită deci atenție în virtutea faptului că, spre deosebire de DEX şi de MDA,

Ion Coteanu, Ion Dănăilă, Nicoleta Tiugan (conducătorii lucrării) et al. (1988). Supliment la Dicționarul explicativ al limbii române (DEX-S) *Bucureşti.*

[J] *Ion Coteanu, Lucreția Mareş (sub conducerea) et al. (1996),* Dicționarul explicativ al limbii române *(DEX), ediția a ll-a, Bucureşti.*

' Dan Tufiş (2000). Cercetare şi colaborare internațională în ingineria lingvistică la RACAI, *în "Terminologia în România şi în Republica Moldova", Cluj-Napoca, p. 34-36 şi* Recherche et collaboration internationale en industries de la langue â l'Academie Roumaine, *în "Terminometro Hors-serie n° 4. La terminologie en Roumanie et en Republique de Moldova", p. 38-40.*

[7] *Pentru detalii cu privire la aceste două dicționare v. Mircea Seche (1969).* Dicționarele explicative ale limbii române literare, *în* Schiță de istorie a lexicografiei române, *voi. II,* De la 1880 până astăzi, *Bucureşti, p. 135-147.*

' D. Macrea, E. Petrovici (sub direcția) et al. (1955-1957). Dicționarul limbii române literare contemporane *(DLRLC), Editura Academiei, Bucureşti, voi. I,* A-C; *//,* D-L *1956' III* M-R *1957; IV,* S-Z, *1957.*

include citate ilustrative, care din păcate au fost eliminate din dicționarele următoare.

2.1.5. O versiune prescurtată a acestui dicționar, cu un inventar puțin măr şi cu adăugarea etimologiei cuvintelor, dar cu eliminarea citatelor, a fost publicată de Institutul din Bucureşti în 1958 sub titlul ***Dicționarul limbii române moderne*** (abreviat DM).

2.1.6. Un dicționar de un tip special, cu o utilitate mult mai largă decâ aceea care i se recunoaşte de obicei, elaborat de data aceasta de colectivul d gramatică al Institutului (condus până de curând de Mioara Avram[20]}, est ***Dicționarul ortografic, ortoepic şi morfologic al limbii române*** (DOOM) \ Est singurul dicționar al limbii române (mai bogat decât DEX) care conține ampl informații cu privire la formele flexionare ale cuvintelor variabile incluse, putân servi astfel (chiar dacă aceste informații nu sunt exhaustive) ca sursă pentru stud şi aplicații de morfologie. Institutul ar.e în prezent în lucru, sub conducere subsemnatei, o a doua ediție, parțial revăzută şi adăugită, a DOOM (care v cuprinde şi cuvinte neînregistrate în nici un dicționar românesc până în prezent Aceasta va apărea în anul 2003, inclusiv pe CD-Rom, şi va trebui să serveasc drept bază unui nou corector ortografic şi morfologic, care să țină seamă d modificarea unor recomandări oficiale în raport cu cele încă în vigoare.

2.1.7. în fine, un dicționar mai puțin obişnuit, ***Dicționarul invers***[22]**,** în car cuvintele sunt ordonate alfabetic pornind dinspre sfârşitul lor, este deosebit de ut specialiştilor pentru studierea terminațiilor, a desinențelor şi a sufixelor, dar poeților,' fiind utilizabil şi ca dicționar de rime. Această lucrare - care, spun "legenda", a valorificat experiența din copilărie a uneia dintre autoare, care folosis în joacă o *păsărească* de acest fel - ar merita şi ea o nouă elaborare, pe baza un inventar mai bogat şi actualizat de cuvinte şi a unui program care să permit "răsturnarea" lor automată.

2.1.8. Institutul a publicat, încă din 1968, un dicționar al lexicului un autor, primul ales neputând fi altul decât Eminescu - *Dicționarul limbii poetice a l Eminescu*[23]*,* care însă, la acea vreme, nu se putea baza, evident, pe stabilire concordanțelor aşa cum se realizează ea în zilele noastre.

2.1.9. Institutul a elaborat de asemenea o serie de dicționare ale limb române pe epoci sau pe probleme, cum sunt *Dicționarul limbii române litera*

[19] *D. Macrea (sub direcția) (1958).* Dicționarul limbii române moderne, *Bucureşti.*

[20] *Pentru activitatea acestuia v. Mioara Avram (1999).* Colectivul de gramatică, *în Mioa Avram, Marius Sala, Ioana Vintilă-Rădulescu, op. cit., p. 113-125.*

[21] *Mioara Avram (red. resp.) et al. (1982).* Dicționarul ortografic, ortoepic şi morfologic limbii române *(DOGM), Bucureşti, 1982.*

22 **** (fgsjy* Dicționar invers, *Bucureşti. V. şi Mircea Seche (1969).* Schiță de istorie lexicografiei române, *voi. II,* De la 1880 până astăzi, *Bucureşti, p. 254-255.*

[23] *Tudor Vianu (sub redacția) et al. (1968).* Dicționarul limbii poetice a lui Eminesc *Bucureşti.*

vechi[24] şi *Dicţionarul împrumuturilor latino-romanice în limba română veche*[25], publicate de sectorul de limbă literară, filologie şi poetică[26], condus de Ion Gheţie,' iar în prezent de Alexandru Mareş - şi *Dicţionarul elementelor româneşti din documentele slavo-române*[27], elaborat la sectorul de slavistică[28] - dicţionare destinate în primul rând specialiştilor.

2.1. 10. Un cercetător din institut, Constant Mâneca, a publicat, împreună cu Florin Marcu, un extrem de util, cu toate criticile care i s-au adus, *Dicţionar de neologisme*[29], reluat şi dezvoltat, după moartea celui dintâi, de Florin Marcu, în numeroase variante, de diverse dimensiuni, la diferite edituri, inclusiv pe CD-Rom.

2.1.11. Se află în lucru şi ***Dicţionarul etimologic al limbii române*** (DELR) - coordonator: Marius Sala -, altă lucrare fundamentală a Academiei Române, la care colaborează cercetători din toate sectoarele Institutului, cercetători din Cluj şi Timişoara şi cadre didactice de la universităţile din Bucureşti, Cluj şi Timişoara.

2.1.12. Pe lângă resursele privitoare la numele comune, Institutul a elaborat şi importante lucrări consacrate numelor proprii[30].

Astfel, în domeniul toponimiei, după clasica lucrare a lui Iorgu Iordan[31], s-a realizat în Institut ***Dicţionarul toponimic al României,*** partea I, ***Oltenia***[32]***,*** elaborat sub conducerea lui Gh. Bolocan în colaborare cu cadre didactice de la Universitatea din Craiova, din care au apărut în perioada 1993-2001 primele trei volume, precum şi al doilea dicţionar din serie, consacrat ***Munteniei*** şi aflat în curs de definitivare.

[24] Mariana Costinescu, Magdalena Georgescu, Florentina Zgraon (1987). *Dicţionarul limbii române literare vechi (1640-1780). Termeni regionali* Bucureşti.

[25] *Gh. Chivu, Emanuela Buză, Alexandra Roman Moraru (1992).* Dicţionarul împrumuturilor latino-romanice în limba română veche (1421-1760) *Bucureşti.*

[26] *V. Ion Gheţie (1999).* Colectivul de limbă literară şi filologie, *în Mioara Avram, Marius Sala, Ioana Vintilă-Rădulescu, op. cit,* p. *132-143.*

[27] *Gheorghe Bolocan (redactor responsabil) et al. (1981).* Dicţionarul elementelor româneşti din documentele slavo-române. 1374-1600, *Bucureşti.*

[28] *Cu privire la care v. Virgil Nestorescu (1999).* Sectorul de lexicografie bilingvă. Fostul sector de slavistică, *în Mioara Avram, Marius Sala, Ioana Vintilă-Rădulescu, op cit p 165-174.*

[29] F. Marcu, C. Mâneca (1961-1978). *Dicţionar de neologisme,* Bucureşti, 1961; ed. II revăzută şi adăugită, 1966;[3]1978. V. şi Mircea Seche (1969). *Schită'de istorie a Jexicografiei române,* voi. Il, *De la 1880 până astăzi,* Bucureşti, p. 154-159.

Pentru activitatea în acest domeniu v. Gheorghe Bolocan, Ecaterina Mihăilă (1999). Colectivul de onomastică *şi Domniţa Tomescu (1999).* Grupul de lucru PatRom, *în Mioara Avram, Marius Sala, Ioana Vintilă-Rădulescu, op. cit, p. 125-132.*

[31] Iorgu Iordan (1952-1963). *Nume de locuri româneşti în Republica Populară Română* Bucureşti, 1952; *Toponimia românească,* Bucureşti, 1963.

[2] Gh. Bolocan (sub redacţia) et al. (1993-2001). *Dicţionarul toponimic al României. Oltenia* (DTRO), voi. **I–III,** Craiova, Editura Universitaria.

în domeniul onomasticii, de asemenea urmând altei lucrări clasice a lui Ior Iordan[33], Institutul colaborează şi la proiectul internaţional PatRom, care realizea un dicţionar istoric de antroponimie romanică, în care este reprezentată şi lim română, şi din care până acum a fost publicat un prim volum de prezentare[34].

2.2. Dicţionare bilingve şi multilingve

2.2.1. Pe lângă dicţionarele monolingve ale limbii române, Institutul realizat şi unele din cele mai importante dicţionare bilingve[35] (englez-român german-român[37], rus-român[38], ceh-român[39] şi sârb-român[40] - perechea s dicţionarul român-sârb, fiind în curs de redactare; un dicţionar francez-român rămas nepublicat) şi frazeologice (spaniol-român, sub tipar, şi român-spaniol, curs de elaborare), cărora li se adaugă dicţionare bilingve[41] - care au început a transpuse şi pe CD-Rom - şi dicţionare frazeologice româneşti[42] şi bilingv elaborate de unii membri ai Institutului; *Dicţionarul elen-român,* lucrare colecti se apropie şi el de sfârşit.

2.2.2. Institutul a colaborat şi la mai multe dicţionare multilingve[44], din care se distinge în mod deosebit un lexicon multilingv de un tip special - adevărată premieră internaţională - ***Dicţionarul elementelor latineşti sava din limbile romanice,*** elaborat la sectorul de romanistică (condus iniţial de mar

[33] *Iorgu Iordan (1983).* Dicţionar al numelor de familie româneşti, *Bucureşti.*

*34 *** (fQQj)* pictionnairehistorique d'anthroponymie romane (PatRom). Presentation d projet, *Tubingen.*

[35] *V. şi Ilinca Constantinescu. (1999).* Fostul sector de germanistică, *în Mioara Avra Marius Sala, Ioana Vintilă-Rădulescu, op. cit, p. 174-179.*

[36] *L. Levi̯tchi (red. resp.) etal. (1974).* Dicţionar englez-român, *Bucureşti. Suplimentul la acest dicţionar, care nu a mai apărut, coordonat de Ilinca Constantinescu, va fi inclus în nouă ediţie, mult mărită, a dicţionarului, aflată sub tipar şi care va reprezenta cel mai bo dicţionar englez-român.*

[37] *M. Isbăşescu, Măria Iliescu (coord. şi revizie) et al. (1966, 1988).* Dicţionar germ român, *Bucureşti, 1966; ediţia a ll-a revăzută şi îmbogăţită, Bucureşti, 1988.*

[38] *Gheorghe Bolocan (redactor responsabil) (1964).* Dicţionar rus-român, *Bucureşti.*

[39] *S. Staţi (red. resp.) et al. (1967).* Dicţionar ceh-român, *Bucureşti.*

[40] *M. Tomici (1998-2000).* Dicţionar sârb-român, *3 voi, Timişoara.*

[41] *Gh. Bolocan (1972).* Dicţionar bulgar-român, *Bucureşti - Sofia; Gh. Bolocan et al. (198* Dicţionar român-rus, *Bucureşti - Moscova; Al. Calciu, C. Duhăneanu, D. Munteanu (1979).* Dicţionar român-spaniol, *Bucureşti; Ana Canarache (coord.) (1967, 1978).* Dicţionar român-francez, *Bucureşti,*[2]*1978; M. Isbăşescu (red. resp.) (1963),* Dicţionar român-german, *Bucureşti; Valeria Neagu (2001).* Dicţionar român-spaniol *(cu transpune pe CD-Rom), Bucureşti.*

[42] *V. Breban et al. (1969).* Dicţionar de expresii şi locuţiuni româneşti *Bucureşti.*

[43] *Gh. Bolocan et al. (1968).* Dicţionar frazeologic rus-român, *Bucureşti; H. Mantsch et (1979).* Dicţionar frazeologic român-german, *Bucureşti.*

*44 *** flQQiy* Dictionnaire de la presse ecrite et audiovisuelle. Espagnol-frangais-itali portugais-roumain, *Paris; *** (2001).* Usage Dictionary of Anglicisms in Selected Europe Languages (UDASEL) *Oxford ş.a.*

romanist Iorgu Iordan, apoi de Marius Sala şi în prezent de subsemnata)[45], în colaborare cu cadre didactice de la Facultatea de Limbi şi Literaturi Străine a Universității din Bucureşti şi în coordonarea prof. dr. Sanda Reinheimer Rîpeanu, decanul Facultății. Negăsindu-şi un editor "clasic" din cauza costurilor prea ridicate, acest dicționar va fi publicat direct pe Internet, sub auspiciile Universității din Bucureşti.

3. Bănci de date

3.1. Institutul a avut în proiect încă din anii 1978-80 realizarea primei bănci computerizate de date lingvistice din România ***(Banca de date fono-morfo-semantice a limbii române - BANDASEM)",*** cel dintâi modul fiind cel de semantică, proiectat pentru un *Dicționar confruntativ de sinonime, de analogii şi de asociații al limbii române (DCSAAs).* Redactarea acestuia, care a ajuns la litera S, s-a făcut însă cu mijloace tradiționale, deşi prin colaborarea cu Centrul de Calcul al Universității din Bucureşti se elaborase un modul de program în sistemul Socrate pentru recunoaşterea şi selectarea, ca probă, a analogiilor şi a asociațiilor cuvântului *blitz.* Elaborarea DCSAAs a fost întreruptă pentru un timp în favoarea lucrărilor prioritare al Academiei, iar reluarea lui se va putea face, sperăm, cu mijloacele informatice disponibile actualmente[47].

3.2. O minibancă inițiată în cadrul sectorului de gramatică al Institutului, a cărei alimentare a fost din păcate întreruptă în favoarea concentrării forțelor pentru realizarea ediției a doua a "Gramaticii Academiei", este ***Banca de inovații a limbii române,*** bazată pe monitorizarea presei scrise şi audiovizuale actuale.

3.3. Având în vedere că în DOOM informația este atomizată, în folosul cititorului neprofesionist, în cadrul fiecărui cuvânt-titlu în parte, dar este greu de sistematizat de către specialist, Institutul are în proiect, începând din 2003, realizarea unui baze de date care să permită nu numai elaborarea unui ***Nou dicționar ortografic, ortoepic şi morfologic al limbii române*** şi a unor dicționare specializate de un tip asemănător, precum şi aducerea lor permanentă la zi, ci şi gruparea cuvintelor în clase în funcție de caracteristicile lor fonetice, grafice şi morfologice[48].

Cu privire la activitatea acestuia v. Marius Sala (1999). Sectorul de limbi romanice şi clasice, *în Mioara Avram, Marius Sala, Ioana Vintilă-Rădulescu, op. cit, p. 147-164.*

[46] *Ion Dănăilă (2000).* Proiecte de prelucrare electronică a vocabularului limbii române, *în "Terminologia în România şi în Republica Moldova", Cluj-Napoca, p. 36-37.*

[47] *Partea de fonetică/grafematică şi de morfologie a* BANDASEM *a fost cedată institutului omolog din Cluj, pentru care v. Felicia Şerban et al. (2000).* Baza de date a limbii române, *în "Terminologia în România şi în Republica Moldova", Cluj-Napoca, p. 37-38 şi La* base de donnees de la langue roumaine, *în "Terminometro Hors-serie n° 4. La terminologie en Roumanie et en Republique de Moldova", p. 40-42.*

[48] *Clasificarea cuvintelor româneşti conform modului lor de flexiune, realizată de Alf Lombard, Constantin Gâdei (1981).* Dictionnaire morphologique de la langue roumaine,

3.4. Institutul are în proiect şi elaborarea sau definitivarea terminologice[49] (dicționare terminologice bi- şi multilingve, valo elaborate în cadrul proiectului PRACTEAST din cadru programului C al Comisiei Europene[50] şi un dicționar al termenilor oficiali); de al membri ai Institutului au colaborat la realizarea ***Băncii de date*** (BDT) multilingve a Asociației Române TermRom[51], care, cu sprijin terminologie şi inginerie lingvistică a Uniunii Latine, este accesib TermRom găzduit de CIMEC (http://www.cimec.ro/tr) şi, de curând, şi Reprezentarea României (prin subsemnata) în Rețeaua Panlatină d (Realiter)[52] şi în Rețeaua Francofonă de Amenajare Lingvistică constitui desigur un sprijin în dezvoltarea resurselor terminologice română în conformitate cu normele şi recomandările internaționale.

Lund - Bucureşti, bazată pe inventarul DEX„ prezintă unele inexact insuficientei cunoaşteri de către autori a limbii române actuale; ea co bazele realizării, în Republica Moldova, a unui pachet de programe destin nivel morfologic, pentru care v. Elena Boian et al. (2000). Instrumentar lingvistice, *în "Terminologia în România şi în Republica Moldova", Cluj-Nap* Instruments pour applications linguistiques, *în "Terminometro Hors-s terminologie en Roumanie et en Republique de Moldova", p. 42-44; o gru unui număr limitat de cuvinte ale limbii române a fost realizată de Flora Ş Şoşa (1999) în* îndreptar ortografic şi morfologic; *Bucureşti.*

[] *V. Ioana Vintilă-Rădulescu (1999).* Institutul de Lingvistică "Iorgu Iordan" d *"Terminologia în România şi în Republica Moldova", Cluj-Napoca, p. 13-1* Linguistique Iorgu Iordan de Bucarest, *în "Terminometro Hors-serie n° 4. en Roumanie et en Republique de Moldova", p. 22-23.*

[] *Nicoleta Petuhov. (2000).* Colaborarea românească la proiectul PR *"Terminologia în România şi în Republica Moldova", Cluj-Napoca,* collaboration roumaine au projet Practeast, *în "Terminometro Hors-s terminologie en Roumanie et en Republique de Moldova", p. 64-66.*

[] *Dan Matei (2000).* Banca de date terminologice a TermRom, *în "Terminol şi în Republica Moldova", Cluj-Napoca, p. 29-30 şi La* banque de donnees de TermRom, *în "Terminometro Hors-serie n° 4, La terminologie en ^Republique de Moldova", p. 32-33.*

[] *Dan Matei (2000).* Prezența românească în rețeaua panlatină de termino *în "Terminologia în România şi în Republica Moldova", Cluj-Napoca,* presence roumaine dans le Reseau panlatin de terminologie Realiter, *în Hors-serie n° 4. La terminologie en Roumanie et en Republique de Moldova*

[] *Ioana Vintilă-Rădulescu (2000).* Colaborarea în cadrul ACCT/Agenți mentale a Francofoniei şi al Rifal, *în "Terminologia în România şi în Rep Cluj-Napoca, p. 51-52 şi* La cooperation dans le cadre de l'A Intergouvernementale de la Francophonie), *în "Terminometro Hors-s terminologie en Roumanie et en Republique de Moldova", p. 57-58.*

4. Corpusuri

O altă categorie importantă de resurse lingvistice o constituie corpusurile, la Institut fiind în curs de realizare o ***Bancă de texte româneşti,*** care cuprinde texte din secolele al XVI-lea - al XVIIMea, introduse integral în calculator, şi în care se prevede introducerea câtorva sute de texte din toate epocile. Inițiată de directorul institutului, acad. Marius Sala, *Banca* a fost deja valorificată în elaborarea unor teze de doctorat, printre altele la aceea a Janei Balacciu-Matei. Pentru exploatarea ei deplină în vederea identificării primelor atestări ale cuvintelor limbii române din fondul vechi, necesare MDA şi *Dicționarului etimologic al limbii române* (DELR), a îmbogățirii dicționarelor limbii române în general şi a dezvoltării studiilor privind istoria limbii române literare şi a limbii noastre în ansamblu este necesară achiziționarea unor programe de ultimă oră, precum şi specializarea unor persoane pentru utilizarea lor eficientă. Sperăm de asemenea că într-un viitor nu prea îndepărtat se va realiza şi dorita joncțiune cu Banca de texte din faza modernă şi contemporană a limbii române, proiectată a se realiza la Centrul de Studii Româneşti de pe lângă Universitatea din Anvers, inaugurat în primăvara anului 2000 sub conducerea cunoscutei romaniste şi romaniste Liliane Tasmowski.

5. Resurse bibliografice

Amintim pe scurt şi principalele resurse bibliografice privitoare la limba română elaborate de Institut sau de membri ai acestuia[54]. *Bibliografia limbii române,* inițiată de Al. Rosetti şi definitivată de Aurel Nicolescu, a rămas nepublicată. *Bibliografia românească de lingvistică (BRL)* referitoare la lucrările de lingvistică apărute în țară începând din 1944 apare anual în revista "Limba română"; în 1999, ea totalizase deja 64.340 de titluri, în peste 3.300 de pagini de tipar; se preconizează introducerea în calculator a tuturor numerelor din BRL în vederea publicării unui volum cu itemurile ordonate pe autori şi pe domenii (descrise şi separate mai amănunțit decât în forma apărută, cronologic, cu indice de domenii, materii, cuvinte, autori etc).

Pentru domeniul terminologiei s-au realizat bibliografii ale dicționarelor terminologice, respectiv ale studiilor de terminologie[55] şi ale standardelor româneşti de/cu terminologie[56], precum şi un repertoriu bio-bibliografic al terminologilor

'1. Coteanu, I. Dănăilă (1970). Introducere în lingvistica si filologia românească Probleme. Bibliografie, *Bucureşti; T. Vianu (red. resp.) et al. (1972).* Bibliografia analitică a limbii române literare. 1780-1866, *Bucureşti; Gh. Chivu, Mariana Costinescu (1974).* Bibliografia filologică românească. Secolul al XVI-lea, *Bucureşti.*

' Anca Fezi et al. (2000). Bibliografia lucrărilor de terminologie (1990-1999). România, *în "Terminologia în România şi în Republica Moldova", Cluj-Napoca, p. 103-113 şi pe discheta anexată revistei Terminometro Hors-serie n° 4. La terminologie en Roumanie et en Republique de Moldova".*

'Aurora Pețan, EdySăvescu (2000). Standarde româneşti de/cu terminologie (1990-1999). România, *în "Terminologia în România şi în Republica Moldova", Cluj-Napoca, 2000,* p.

din România[57], inclus în repertoriul internațional al terminologilor din domeniul neolatin pregătit de Uniunea Latină şi accesibil pe Internet.

6. Concluzii

Nu ne vom referi aici la alte tipuri de lucrări (gramatici[58], tratate[59], enciclopedii[60] etc.) elaborate de Institut sau de cercetători ai acestuia ori la alte tipuri de resurse care ar merita să fie elaborate de noul institut, pentru a înlocui lucrări mai vechi şi a valorifica posibilitățile oferite culturii de societatea informațională, de exemplu un nou dicționar de frecvență al limbii române ş.a.

Deşi dicționarele pe CD-Rom şi cele pe Internet sunt solicitatete de tot mai mulți utilizatori din țară şi din străinătate, care cer tot mai des informații cu privire la eventuale dicționare româneşti on-line, până în prezent a existat la noi o anumită reticență a editurilor proprietare ale drepturilor asupra edițiilor pe suportul tradițional de hârtie față de acest nou mod de difuzare. Nu trebuie însă să existe temerea că folosirea şi a noilor suporturi ar diminua vânzarea cărților, în condițiile în care, în ciuda tuturor eforturilor, un procent încă infim din populația României are acces la PC-uri. De altfel, practica altor țări a arătat că, în mod neaşteptat, difuzarea şi în format electronic chiar a sporit desfacerea cărților, cărora le-a făcut în felul acesta reclamă şi care prezintă, la rândul lor, alte avantaje în utilizare în raport cu CD-Romurile, cele două tipuri specializându-se şi în funcție de necesități. Astfel, având în vedere culegerea lor computerizată, atât DEX, cât şi MDA şi DOOM ar putea fi primele dicționare ale Institutului difuzate în viitor şi pe CD-Rom.

Credem că şi diverse lucrări valoroase ale Institutului, care, exclusiv din motive financiare, nu-şi găsesc editori de ani de zile, nici în țară, nici în străinătate (ca *Bibliografia limbii române, Dicționarul spaniolei americane* ş.a.), ar putea

117-126 şi pe discheta anexată revistei Terminometro Hors-serie n° 4. La terminologie en Roumanie et en Republique de Moldova".

[57] *Adriana Marinescu (2000).* Repertoriul bibliografic al terminologilor. România, *în "Terminologia în România şi în Republica Moldova", Cluj-Napoca, 2000, p. 128-139 şi pe discheta anexată revistei Terminometro Hors-serie n° 4. La terminologie en Roumanie et en Republique de Moldova".*

[8] **** (1954, 1963).* Gramatica limbii române, *Bucureşti, ed. I, 1954; ed. a ll-a, revăzută adăugită, 1963; Mioara Avram (1986, 1997, 2001).* Gramatica pentru toți, *Bucureşti, 198*

[s] [2]*1997;*[3]*2001.*

[59] *Al. Rosetti (redactor responsabil) et al. (1965, 1969).* Istoria limbii române. *Bucureşti, vo*

I. Limba latină, *voi. al ll-lea; Al. Graur, Mioara Avram (1970-1989).* Formarea cuvintelor limba română, *Bucureşti: I. Fulvia Ciobanu, Finuța Hasan (1970).* Compunerea; //. *Mioa Avram et al. (1978).* Prefixele, *1978; III. Laura Vasiliu (1989).* Sufixele, 7. Derivare

[60] verbală *etc.*

Marius Sala, Ioana Vintilă-Rădulescu (1981). Limbile lumii. Mică enciclopedie, *Bucureş (1984).* Les langues du monde. Petite encyclopedie, *Bucureşti - Paris; Marius Sa (coord.) etal. (1989).* Enciclopedia limbilor romanice, *Bucureşti; (2001),* Enciclopedia lim române, *Bucureşti.*

valorificare prin aducerea lor la cunoştinţa celor interesaţi pe această cale, tot mai utilizată în societatea informaţională actuală. O condiţie pentru viitor este realizarea din capul locului a lucrărilor institutului pe calculator, care a devenit posibilă prin tot mai buna dotare tehnică a Institutului, realizată prin eforturile directorului său, precum şi prin însuşirea, de către un număr tot mai mare de cercetători din Institut, în special din generaţiile tânără şi mijlocie, a cunoştinţelor de operare pe calculator, inclusiv, în unele cazuri, a lucrului cu baze de date.

Prin realizarea proiectelor de editare pe CD-Rom şi pe Internet vom recupera relativa întârziere în acest domeniu faţă de difuzarea în România, de către Grupului Editorial Litera din Republica Moldova şi firma Litera Internaţional, cu sediul în Bucureşti, a unor CD-Romuri cuprinzând, în diverse combinaţii, mai multe titluri[6]. Sperăm că CD-Romurile consacrate unor dicţionare ale Institutului vor fi, deşi tot protejate, mai uşor de instalat decât cele de la Litera şi că vor oferi mai multe facilităţi în utilizare decât acestea, care nu sunt foarte practice, mai ales pentru cercetători, în ciuda structurii lor modulare şi a interfeţei lor comune, despre care în reclamă se spune că permit activarea simultană a tuturor dicţionarelor.

Pentru progresul cercetărilor şi dezvoltarea şi prelucrarea resurselor la nivelul exigenţelor pe plan mondial, credem că în viitor se impune o mai bună colaborare, în interes reciproc, între lingvişti şi informaticienii preocupaţi de probleme asemănătoare.

Corectorul electronic ORTO 2001 ROM SP, Dicţionarul ortografic al limbii române, Gramatica uzuală a limbii române, Noul dicţionar explicativ al limbii române^ Marele dicţionar de neologisme *de Florin Marcu,* Dicţionarul de dublete etimologice ale limbii române *de Marcu Gabinschi şi un* Dicţionar de termeni de afaceri englez-român.

Contribuţia lingvisticii la studiul terminologiilor ştiinţifice

Angela BIDU-VRÂNCEANU
Universitatea din Bucureşti, Edgar Quinet nr. 5-7
vrancean@gpsnet.ro

1. Se admite "laicizarea" ştiinţelor [1] sau importanţa lor socio-culturală economică şi pedagogică tot mai mare în societăţile moderne. Aceasta înseamn că *limbajele specializate* şi *terminologiile* lor nu mai reprezintă coduri tota inaccesibile vorbitorilor obişnuiţi, nespecializaţi sau de altă specialitate. în direcţi deschiderii, chiar şi parţiale a codurilor ştiinţifice, *dicţionarele generale [2],* car includ un număr destul de mare de termeni ştiinţifici joacă un rol deosebit pentru asigura accesul la sensul specializat oricărui vorbitor insuficient informat, pentru a ajuta să rezolve ambiguităţile de diferite tipuri şi chiar să utilizeze adecvat terminologie. Permanenta raportare la dicţionarele generale ca forme instituţio nalizate de reglare a uzului nu numai al cuvintelor din limba comună, ci şi termenilor specializaţi constituie premisa de la care pornim pentru a susţin importanţa lingvisticii în descrierea terminologiilor ştiinţifice, în receptarea s utilizarea lor adecvată chiar şi de către nespecialişti.

Pe aceste poziţii s-a situat activitatea în cadrul a trei contracte de cercetar ştiinţifică pe anii 1997, 1999 şi 2000, finanţate de CNCSIS (Consiliul Naţional d Cercetare Ştiinţifică). Au fost studiate limbajul **filozofic,** terminologiile **matematică mineralogică** şi din **artele plastice** şi, dintr-o perspectivă mai limitată **medicina lingvistica** şi **ştiinţele politice.** Rezultatele cercetărilor au fost publicate în dou volume: *Lexic comun, lexic specializat* [3], care conţine studii cu caracte monografic şi *Lexic ştiinţific interdisciplinar* [4], reprezentând o sinteză lexicografiei generale şi specializate pentru termenii din fiecare dintre domeniil studiate care apar mai mult decât într-o terminologie ştiinţifică.

în toate cercetările întreprinse s-a urmărit adoptarea *unei grile metodo logice comune* atât pentru clase de cuvinte din limba comună *(abstractele),* cât pentru termenii specializaţi din orice domeniu. S-a obţinut atât caracterizare fiecărei terminologii studiate în parte, cât şi desprinderea unor trăsături general ale terminologiilor ştiinţifice, relevante din punct de vedere lingvistic. S-au avut î vedere aspecte *paradigmatice* privind diferitele modalităţi de definire a sensulu relaţiile semantice *(monosemie/polisemie, hiponimie, sinonimie)* din perspectiv necesităţii ca termenii ştiinţifici să fie monoreferenţiali, univoci din punct de vede

semantic şi să nu aibă sinonime. Analiza **sintagmatică** a gradului de non-determinare contextuală ca o condiție de exprimare a sensului specializat a individualizat terminologiile științifice studiate, de la o *libertate contextuală* mai mare (terminologia **matematică, mineralogică)** sau relativă (terminologia **filozofică)** până la o *strictă determinare contextuală* (terminologia **politică** şi din **artele plastice).** Acolo unde independența contextuală e mai mare, determinările contextuale exprimă în mod similar în diferite terminologii **(matematică, filozofică, lingvistică)** subcategorii științifice care dezambiguizează lexicul științific intedisciplinar. Caracterizarea termenilor științifici prin mărci diastratice în dicționarele generale şi enciclopedice ca tipuri de informații sintagmatice reprezintă un aspect foarte important pentru uzajul adecvat de către specialişti, aspect deficitar, inegal rezolvat.

De pe poziția receptorului nespecializat care decodează sensul total sau parțial, un rol important îl are *definiția lexicografică* care, spre deosebire de cea *terminologică* trebuie să fie mai mult sau mai puțin *naturală* şi prin aceasta accesibilă. Existența celor două tipuri de definiții ale termenilor specializați este în general admisă şi compararea lor este favorizată de prezentarea sintetică, sinoptică propusă de noi [4]. Chiar şi în cazul definițiilor strict terminologice, Em. Vasiliu [5] a susținut şi demonstrat prin diferite exemple relevanța diferită a unor componente de sens pentru vorbitorul specialist sau non-specialist. Pornind de la aceste constatări de principiu, ar fi justificat ca termenii științifici să aibă *definiții alternative,* ştiințifice şi pre-ştiințifice [6], condiționate atât de o interpretare semantică, cât şi de una pragmatică. Din această perspectivă, definițiile termenilor ştiințifici în dicționarele generale ar trebui să difere de cele din dicționarele specializate pentru a facilita deschiderea codurilor ştiințifice şi pentru a dezambiguiza lexicul ştiințific interdisciplinar (din principiu, de interes mai larg) sau tangențele cu limba comună. Din păcate, cu mici excepții **(matematica)** selecția termenilor ştiințifici şi definirea lor nu diferă aproape deloc în dicționarele generale şi în cele specializate.

2. Din perspectivă lingvistică, terminologiile investigate prezintă o serie de particularități:

Matematica se caracterizează prin cel mai mare grad de abstractizare şi de ermetism la nivelul sensurilor şi definițiilor lor. Compararea definițiilor specializate cu cele din dicționarele generale arată că acestea din urmă definesc diferit şi mai accesibil termenii, fără a afecta precizia lor semantică. Sensurile univoce, fără sinonime nu sunt condiționate contextual; sintagmele mai mult sau mai puțin fixe diferențiază subcategorii conceptuale (de ex. *sistem de ecuații, ~ de curbe, ~ de numerație, ~ de referință)* şi nu afectează independența semantică a acestora. Această terminologie dispune de cea mai bună marcare diastratică în DEX, chiar dacă există numeroase situații în care apartenența la matematică rezultă numai din definiție (manieră de caracterizare practicată sistematic şi nu întotdeauna convenabil de DEX în cazul altor terminologii). **Matematica** are cel

mai bogat lexic ştiințific interdisciplinar, cei mai numeroşi termeni comuni fiind cu **fizica, filozofia, logici,** dar şi cu **lingvistica, biologia, arhitectura** ş.a.; termenii interdisciplinari îşi păstrează aproape neschimbat sensul, indiferent de domeniul în care se utilizează. Dacă în unele cazuri (relația cu **fizica, logica, filozofia)** punctul de plecare pentru lexicul interdisciplinar nu se poate stabili cu certitudine, în destule alte situații, **matematica** este sursa "împrumutului" făcut de alte ştiințe **(arte plastice, arhitectură, lingvistică** ş.a.)

Mineralogia reprezintă şi ea un grad mare de ermetism sau închidere a codului, majoritatea termenilor fiind univoci semantic, monoreferențiali şi implicit, independenți contextual. Determinările contextuale reprezintă subtipuri, ca şi în alte terminologii **(matematică, filozofie** de ex.; *acvamarin brazilian, ~ sintetic, ~ siamez,* etc.) Are un număr mai limitat de termeni comuni cu alte ştiințe **(chimia, artele plastice, simbolistica)** şi, cel puțin pentru ultimele două, **mineralogia** este punctul de origine al termenilor interdisciplinari. în ciuda caracterului strict specializat al acestei terminologii, marcarea diastratică din dicționarele generale este deficitară.

Terminologia **filozofică** se caracterizează printr-un grad oarecare de ambiguitate, determinat de variații de interpretare în funcție de curente şi tipuri de texte, dar şi de contactele cu alte ştiințe sau cu limba comună. De aceeadefinițiile termenilor **filozofici** nu se pot limita la dicționare, fiind necesară analiza strategiilor argumentative şi a figurilor textuale; Invers proporțional cu această necesitate de dezambiguizare, DEX-ul prezintă o marcare diastratică deficitară atât pentru termenii filozofici, cât şi pentru celelalte terminologii cu care se stabilesc interdisciplinarități, cum ar fi **matematica, lingvistica** şi alte domenii **umaniste.** O bună parte a lexicului ştiințific interdisciplinar are ca punct de plecare **filozofia,** al cărei sens se păstrează ca o medie semantică în majoritatea disciplinelor. Ca şi în alte ştiințe, determinarea contextuală exprimă în general subtipuri (de ex. *sistem al ştiințelor, ~ axiomatic, ~ filozofic).*

Terminologia **artelor plastice** prezintă aspecte paradoxale. Maniera de înregistrare şi de definire echivocă, imprecisă a acestor termeni în dicționarele generale dă impresia unui nespecialist de falsă accesibilitate, interpretare contrazisă categoric de definițiile precise, riguroase din dicționarele şi textele specializate. Dependența contextuală strictă a numeroşi termeni din **artele plastice,** al căror sens specializat e condiționat de sintagmele fixe în care apare (de ex. *acord cromatic, compoziție de gen, semn plastic)* reprezintă o altă caracteristică a acestei terminologii. **Artele plastice** au un lexic ştiințific interdisciplinar bogat, în care se remarcă faptul că sunt preluați cu unele modificări semantice (privind interesul pentru acest domeniu) termeni din alte ştiințe, cum ar fi **chimia, mineralogia, matematica, fizica.** DEX-ul nu utilizează decât mărcile diastratice (pictură), (sculptură) dispuse nesistematic şi rar, ceea ce contribuie la o tratare deficitară a acestei terminologii.

Lexicul ştiinţelor politice prezintă, din prespectiva analizei întreprinse de noi, o serie de particularităţi (unele asemănătoare cu artele plastice). Se remarcă dependenţa contextuală strictă a acestei terminologii, nici unul dintre termeni nefiind total liber contextual. Sensul specializat în ştiinţele politice se exprimă, deci, aproape exclusiv pe cale sintagmatică, în contexte mai mult *(celulă de criză, agregare de interese,* de ex.) sau mai puţin fixe (diverse combinaţii cu adjectivul politic în sintagme nominale: *capital politic, cartel ~ algoritm ~, contract ~, dialog ~ alternanţă politică).* Preia (fără să fie niciodată punct de plecare termeni din numeroase şi variate ştiinţe: economia, filozofia, dreptul, dar şi lingvistica, biologia, medicina, geografia, fizica, psihologia, sportul. în majoritatea acestor cazuri nu există o motivare de conţinut strictă (dincolo de întrebuinţarea metaforică), ceea ce determină, în mare parte, mai curând un lexic ştiinţific interferent decât unul interdisciplinar. Poate şi din cauza modificărilor continue şi rapide din domeniul politicii, DEX-ul înregistrează în mică măsură termeni şi sensuri din acest domeniu diastratic, ceea ce constituţie un dezavantaj în impunerea acestei terminologii.

3. Analiza lingvistică a limbajelor ştiinţifice (care ar putea fi extinsă) permite caracterizarea unor terminologii ca *"puternice"* (matematica, mineralogia de ex.), iar a altora mai *"slabe"* în diferite forme şi grade (de ex. ştiinţele politice, artele plastice), cu dificultăţi mai mari de deschidere a codurilor în cazul primei categorii.

Delimitarea componentelor de sens relevante diferit în funcţie de vorbitori specializaţi şi nespecializaţi ar putea constitui o bază obiectivă pentru rezolvarea mai eficientă a definiţiilor alternative în dicţionarele generale, foarte importante în "laicizarea" ştiinţelor necesară în grade diferite în epoca actuală. Exprimarea sensului specializat condiţionat de dependenţele contextuale mai mici (pentru terminologiile "puternice") sau mai mari (pentru terminologiile "slabe") constituie o caracterizare lingvistică relevantă. în schimb, în unele cazuri (ca pentru terminologia politică), determinările contextuale sunt mai favorabile, "transparenţei" semantice sau deschiderii codurilor specializate.

Analiza lexicului ştiinţific interdisciplinar (LSI) poate contribui şi ea la determinarea specificului unor terminologii. Ştiinţele care constituie sursa, punctul de plecare pentru o parte a LSI îşi susţin, şi pe această cale, statutul de terminologie "puternică" (de ex. matematica, fizica şi, din acest punct de vedere filozofia). Dimpotrivă, atunci când punctul de plecare nu se poate stabili aproape niciodată la nivelul unor terminologii (ştiinţele politice, artele plastice), aceasta constituie o modalitate de determinare specifică. Diferenţierea interdisciplinarităţilor (cu o motivare de conţinut determinată de considerarea referentului din diferite puncte de vedere sau de un transfer conceptual) de simplele interferenţe (mai puţin sau deloc motivate, cu modificări de sens ale termenilor, multe metaforice) se bazează pe aprecierea distanţei semantice, verificată obiectiv.

Dat fiind rolul dicţionarelor generale în impunerea şi extinderea terminologiilor ştiinţifice, de interes pentru diferite categorii de vorbitori, carenţele constatate în tratarea sensului şi în marcarea lor diastratică riguroasă conduc la concluzia necesităţii unei reconsiderării şi remedieri a manierei de tratare din perspectiva "laicizării" ştiinţelor.

Referinţe bibliografice

[1] F. Rastier (1995) Le terme; entre ontologie et linguistique. *Banque des mots* 1995/7, p. 35-65.

[2] DEX - Dicţionar explicativ al limbii române, (1996) ed.a 2-a sub coord. acad- I. Coteanu şi Dr. Lucreţia Mareş, Ed. Univers Encilopedic, Bucureşti 1996.

[3] A. Bidu-Vrânceanu - coordonator (2000). *Lexic comun, lexic specializat,* Editura Universităţii din Bucureşti, 2000, cu colaboratorii: Alice Toma (matematică), Silvia Săvulescu (mineralogie), Claudia Ene (filozofie), Alexandra Vrânceanu (arte plastice).

[4] A. Bidu-Vrânceanu - coordonator (2001). *Lexic ştiinţific interdisciplinar,* Editura Universităţii din Bucureşti, 2001, cu colaboratorii: Silvia Săvulescu (ştiinţe politice şi mineralogie), Alice Toma (matematică),Claudia Ene (filozofie), Alexandra Vrânceanu (arte plastice).

[5] Em. Vasiliu (1980). Sens şi definiţie lexicografică "Studii şi cercetări lingvistice", an XXXI, 465, 1980.

6] Em. Vasiliu (1982/1983). Adevăr analitic şi definiţie lexicografică "Analele ştiinţifice ale Universităţii "Al. I Cuza" din Iaşi", secţiunea III, tom XXVIII/XXIX, 1982/1983.

Gramaticile generative nontransformationale

Emil IONESCU
Universitatea Bucureşti, Facultatea de Litere
Str. Edgar Quinet nr. 5-7,
Email: eionescu@racai.ro

Acest articol este o prezentare generală a gramatic
nontransformationale (GNT) şi a prezenței lor în cercetarea
România. în prima secțiune a articolului este descrisă geneza aces
secțiunea a doua, sunt prezentate pe scurt caracteristicile lor, în ti
treia şi a patra se menționează principalele realizări ştiințifice
existență instituțională ale curentului. Partea a cincea este consac
şi paşilor care au dus la pătrunderea acestor gramatici în mediile
noi. Concluziile articolul se vor a fi o pledoarie în sprijinul eforturilor
acestei direcții în cultura ştiințifică românească.

1. Gramaticile generative nontransformationale: apariț

Gramaticile generative nontransformationale reprezintă
lingvisticii formale contemporane, o direcție extrem de influentă şi d
dinamism. Istoria acestei direcții este, desigur, mai recentă
generativismului din care face parte. Este însă o istorie deja bo
Printre altele, diversitatea se exprimă şi prin faptul că suntem ob
despre *gramatici* şi nu despre o gramatică nontransformațională, pu

Putem plasa începuturile acestei istorii la cumpăna dintre
Sunt anii când programul gramaticii universale al lui Noam Ch
punctul să depăşească starea de impas atinsă prin faza denu
mişcării "teoria standard". Privită din perspectiva prezentului, lucra
lui Chomsky ("Lectures on Government and Binding") tocmai ace
niază: depăşirea crizei prin propunerea unui model nou de gramatic

Punctele în care gramatica universală este reformulată în c
"Government and Binding" (GB) nu sunt puține şi nici neînsemnat
importantă modificare a fost operată într-una din componentele car
cele mai mari speranțe: componenta transformărilor. Formulată suc

conceptului de transformare în cadrul modelului GB înseamnă două lucruri: simplificare şi îngrădire. Simplificare, deoarece marea varietate de transformări se reduce acum la o singură operație: deplasarea unui constituent oarecare a. Şi îngrădire, pentru că deplasarea nu se poate produce oricum, ci numai în condițiile în care anumite reguli foarte generale, numite *principii,* sunt respectate.

Nu toți adepții generativismului au fost însă mulțumiți cu noua propunere. Ceea ce s-a reproşat a fost că transformările rămâneau mai departe mecanisme prea puternice - în ciuda îngrădirilor şi a simplificărilor - deoarece ele operau pe un domeniu prea larg: cel al structurilor sintactice. O altă obiecție viza temeiurile mentale ale operației de deplasare: în ciuda plauzibilității aparente a acestei ipoteze, nu există dovezi - susțineau criticii - că mintea implicată în utilizarea limbajului ar face uz de o astfel de operație. în sfârşit, existau cercetători care considerau că noul model de gramatică universală era greoi din punct de vedere computațional, tocmai din cauza operației de deplasare: anume, pentru fiecare deplasare de constituenți, este necesară o verificare a compatibilității dintre principii şi deplasarea constituentului.

în ansamblu, divergențele legate de conceptul de transformare au pregătit cea mai mare ruptură pe care a cunoscut-o în istoria sa curentul gramaticii universale. Criticii radicali ai conceptului de transformare au propus renunțarea la acest mecanism, propunere pe care Chomsky şi cei ce l-au urmat nu au acceptat-o niciodată. începând cu anul 1981, ruptura se oficializează. Apar pe rând Gramatica Lexico-Funcțională (LFG - Bresnan şi Kaplan), Gramatica Sintagmatică Generalizată (GPSG - Gazdar, Klein Pullum şi Sag), Gramatica Arborilor Adăugați (TAG - Joshi), Gramatica Centrilor de Sintagmă (HPSG - Pollard şi Sag), Gramaticile Categoriale de Unificare (CUG- Uzkoreit)

2. Caracteristicile G N T

Dincolo de varietatea lor, gramaticile nontransformationale au un set de trăsături comune:

- Exploatează în mod generalizat reprezentările în termeni de trăsături
- Fac recurs la mecanismul unificării
- Se bazează pe constrângeri
- Sunt gramatici lexicaliste
- Au adecvare computațională

2.1. Reprezentări: structurile de trăsături

Reprezentările în termeni de trăsături sunt bine cunoscute în lingvistica modernă, datorită fonologiei şi semanticii structurale. GNT au meritul de a fi generalizat această notație la scara întregii teorii lingvistice. Prin perechea trăsătură (atribut)-valoare, orice fel de informație lingvistică morfologică, sintactică semantică, pragmatică - îşi găseşte adecvată. Câteva exemple: notația [P(arte de)V(orbire): nume anumită entitate lingvistică este un nume. Reprezentarea [F(gerunziu] precizează că avem a face cu un verb la gerunziu; [RAM(ură): v(aloare)n(on)v(idă)] spune că obiectul lingvistic în structură internă şi este prin urmare o sintagmă. Este uşor de rem atribut-valoare aplică principiul general al funcțiilor: unui ar corespunde o anumită valoare, întocmai cum unui argument dat î anumită valoare, datorită unei legi specifice de corespondență. R care se face uz în GNT sunt denumite *structuri de trăsături.*

2.2 Unificarea

GNT se mai numesc şi gramatici de unificare. Unificarea structurile de trăsături. Unificarea a două structuri de trăsături A şi este structura minimală de trăsături care cuprinde în acelaşi timp Dacă o astfel de structură nu există, unificarea "eşuează" (ceea c Unificarea verifică aşadar compatibilitatea dintre două structuri produce o structură rezultantă care conține toată informația din st unificării, lată câteva exemple:

(1) [CAT: det] u [CATmume] = ± (eşec)

(2) [CAT:det]u[ACORD:[NUM:singfl= CAT: det ACORD:[NU

(3) [CAT:nume]u | ACORD: | GEN: mase / NUM :sing | | CAT:num / ACORD:

Operația de unificare din primul exemplu eşuează pen rezultantă ar trebui să conțină atributul CAT cu două valori diferite nume). Unificarea se realizează normal în (2) şi (3), şi produce complexă.

Se poate remarca faptul că rolul unificării este acela de a ce este corect în variate compartimente de limbă. Dacă are lo informații fonologice, aceasta explică un aspect al corectitudinii for limbă dată. O unificare de informații morfologice dă seama d corectitudinii morfologice, ş.am.d. Nu e însă exclusă nici unifica diferite, de exemplu, semantice şi morfologice, semantice şi sintac

2.3. Constrângeri

în exemplul (1) din paragraful precedent, unificarea eşuează deoarece nici o structură de trăsături nu poate avea valori diferite pentru acelaşi atribut. Aceasta este o "lege" inerentă unificării, tot astfel cum în logica bivalentă o "lege inerentă" este terțiul exclus. Se poate spune că (1) defineşte o limită a unificării şi implicit o constrângere asupra acestei operații. Constrângerea este de natură formală, pentru că derivă din natura însăşi a unificării. Dar pentru scopurile unei teorii lingvistice, astfel de constrângeri nu pot fi suficiente. Polona, de pildă, face la verbele de persoana I deosebirea între verbele folosite de un bărbat şi cele folosite de o femeie. Verbul are aşadar gen în polonă, dar nu şi în română. Pentru a face această diferență între cele două limbi trebuie să' se admită că unificarea informației de gen cu cea de verb se poate face în polonă dar nu se poate face şi în română. Numai că de această dată constrângerea privind unificările nu mai are temei formal. Nu se poate spune că în mod necesar verbul are sau nu gen. Unificările acestor informații sunt prin urmare "contingente", sau cu un alt termen, "empirice", tocmai pentru că ele nu derivă din natura însăşi a operației. Gramatica unei limbi se descrie mai ales în termenii unificărilor "contingente".

2.4 Lexicalism

în teoriile contemporane ale gramaticii, lexicalismul este o opțiune privitoare la modul în care este concepută structura cuvintelor în relația lor cu sintaxa. Există teorii, precum GB, care consideră că procesul de constituire morfologică a cuvintelor are loc în sintaxă. în acest sens, GB este o morfosintaxă deoarece generalizează operația de deplasare la nivelul morfologiei înseşi, prin mecanismul numit "deplasare centru-centru" (engl, "Head to Head Movement"). Gramaticile de unificare adoptă o strategie distinctă: ele consideră că procesele de constituire morfologică a cuvintelor sunt independente de sintaxă. în această perspectivă, rezultatul proceselor morfologice furnizează sintaxei inputul necesar: cuvintele gata formate. Modularizarea celor două componente ale gramaticii se dovedeşte preferabilă mai ales în cazul limbilor cu morfologie bogată.

Un alt aspect al lexicalismului asumat de GNT este ilustrat de modul în care sunt construite explicațiile de gramaticalitate. Explicațiile în GNT se sprijină în măsura posibilului (dar într-o măsură mult mai mare decât în alte teorii) pe proprietățile cuvintelor. In istoria generativismului, pasivul, de pildă, a fost considerat multă vreme o structură explicabilă *sintactic,* adică o construcție rezultată din transformări ale unei alte structuri sintactice. GNT afirmă însă că nu e nevoie să se recurgă la structuri sintactice anumite, deoarece toate elementele de care e nevoie pentru a explica o construcție pasivă pot fi codificate la nivelul cuvintelor[1]. Un tratament asemănător

Preferința aceasta pentru un compartiment de limbă în defavoarea altui compartiment, atunci când se pune problema mecanismelor care justifică o anumită construcție nu e înțeleasă încă nici azi de unii lingvişti. Este vorba de aceia care cred că a avansa o explicație lexicalistă atunci când există deja una sintactică pentru un fenomen oarecare

poate fi observat în cazul dependențelor la distanță, sau în cel al c ridicare (engl. "raising"), unde rolul unităților lexicale în determinare acestor construcții este de asemenea semnificativ.

2.5 Adecvare computațională

în lingvistică, o teorie este considerată adecvată, dacă domeniul de fapte pentru care este construită ca o explicație. O mo limbi, de pildă, este adecvată dacă prin regulile propuse dă seamă morfologic corecte ale limbii supuse analizei.

Acest principiu foarte general a fost nuanțat de către Chom este deja celebră: pornind de la ideea că utilizarea limbajului este minții omeneşti, Chomsky a susținut că o teorie trebuie socotită ad pentru că produce explicații ale cazurilor de corectitudine, ci mecanismele utilizate sunt dovedite (sau cel puțin presupuse) a fi în mintea omenească. Quine afirmase că dacă avem două gramatici c diferite explică aceeaşi realitate lingvistică, nu există criterii s alegere a uneia dintre ele. Chomsky a replicat că un astfel de criter el fiind măsura în care fiecare dintre aceste gramatici se folose cunoscute ca aparținând minții în procesele ei cognitive.

Criteriul suplimentar formulat de Chomsky în evaluarea teo a apropiat comunitatea generativiştilor de cea a psihologilor şi cercetările de psiholingvistică. S-au obținut rezultate interesante ş ipoteze neaşteptate. De pildă, regulile de constituenți sînt socot operațiuni cu mare probabilitate de a fi folosite de intel Recursivitatea este şi ea considerată a fi o proprietate de care int face uz în utilizarea limbajului.

Criteriul lui Chomsky a condus însă şi la cercetări cu re judecat. De pildă, despre realitatea psihologică a *urmelor,* conc teoriei GB, s-a argumentat şi pro şi contra, şi este foarte dificil ch poată lua o poziție.

Un lucru este cert totuşi în evoluția raporturilor dintre teo realitatea ei psihologică: comparativ cu faza de început, interesul p psiholingviştilor față de ipotezele venite din comunitatea "chomsky semnificativ. A crescut însă interesul psiholingviştilor pentru ipote lumea inteligenței artificiale. Este celebră în acest sens ipoteza c cunoştințelor lexicale a lui Quillian, care a atras atenția în mod spe

înseamnă doar a propune variațiuni pe aceeaşi temă. Diferențele sunt în şi privesc mecanismele cognitive angajate în utilizarea limbajului. Este procesarea unităților lexicale este mai uşor de efectuat decât unele structurilor sintactice. Acest fapt oferă un criteriu valoros de judecare a gramatici privite din unghi cognitiv.

de psihologi şi de psiholingvşti. Un al treilea factor intra astfel în joc, rezultatul fiind că unele teorii lingvistice au devenit atente la operațiile şi mecanismele utilizate de inteligența artificală. Erau exact teoriile generative netransformaționale. Consecința principală a acestei deplasări de interes a fost că teoriile în cauză au devenit accesibile utilizării automate. Cu alte cuvinte - şi spre deosebire de gramaticile lui Chomsky - ele pot fi implementate computațional.

Vom numi adecvarea unei teorii la domeniul de fapte pe care îl abordează *adecvare lingvistică.* Măsura în care o teorie lingvistică aparține (sau poate fi presupusă a aparține) minții omeneşti defineşte *adecvarea* ei *psihologică.* lar gradul în care ea este livrabilă inteligenței artificiale indică *adecvarea* ei *computațională.* Direcția actuală a curentului de idei pare să fie următoarea: legăturile şi dialogul dintre psihologia cognitivă şi inteligența artificială sunt într-o continuă creştere, astfel încât adecvarea computațională a unei teorii lingvistice are şanse mari să-i confere şi adecvare psihologică. Pe această direcție sunt plasate gramaticile generative netransformaționale.

3. Realizări

Una dintre cele mai importante realizări ale gramaticilor nontransformationale îl reprezintă numărul mare de aplicații. O enumerare a limbilor supuse analizelor nu este posibilă aici, dar se poate preciza că aproximativ doua treimi din familiile de limbi (considerate in eşantioanele lor reprezentative) au fost analizate din perspectiva netransformațională. Este caracteristic acestor analize faptul că refuză deosebirea chomskyană centru-periferie ("core-periphery"). Ele se concentrează asupra varietății de date oferite de corpusuri.

Ceea ce este însă cel mai important sub aspectul realizărilor este faptul că GNT au reuşit să producă replici viabile la analizele paradigmei dominante, cea chomskyană. O serie de fenomene gramaticale - privite de obicei ca fiind de la sine caracterizabile prin mecanismul deplasării constituenților - au primit in cadrul GNT analize alternative. Aşa s-a întâmplat cu construcțiile pasive, cu fenomenul de ridicare (şi mai general cu fenomenele de depedență limitată), cu construcțiile nonlocale (precum topicalizările, structurile relative si interogative). In această privință, GNT au continuat tradiția firească, inaugurată de structuralism, tradiție constând in regândirea fenomenelor de limbă odată cu fiecare nouă şcoala lingvistică.

4. Forme instituționale de susținere

GNT sunt bine reprezentate instituțional. Ele şi-au făcut loc în primul rând în programele curiculare ale unor universități de prestigiu, precum Universitatea Stanford, Universitatea Statului Ohio (Columbus), Universitatea Tuebingen, Universitatea Saarbruecken, Universitatea Groningen, King's College din Londra Universitatea Edinburgh, Universitatea Paris 7. Extensiile acestor programe curiculare sunt şcolile de vară. O prestigioasă şcoală de acest fel ("European Summer School in Logic Language and Information" - ESSLLI) este organizată anual din 1989, cu rolul de diseminare a evoluțiilor si curentelor formate în interiorul gramaticilor netransformaționale. Este apoi de semnalat, în aceeaşi linie a "didacticii" gramaticilor nontransformationale, nou înființata scoală de vară de la Konstaz (Germania).

în planul congreselor ştiințifice, HPSG si LFG au de multă vreme propriile lor conferințe anuale. lar un congres ținut o dată la doi ani - cel de gramatici formale - urmăreşte să adune sub acelaşi acoperiş toate şcolile aceleiaşi familii.

Până de curând, gramaticile nontransformationale nu au avut o revistă proprie. Lucrările însă au fost şi sunt publicate in reviste de prestigiu, precum "Computațional Linguistics" "Natural Language and Linguistic Theory", "Journal of Linguistics", "Language" sau "Langages". O revistă orientată explicit spre aceste gramatici este editată de puțină vreme la cunoscuta editură olandeză Kluwer. Este vorba despre revista "Grammars". De asemenea, pe lângă Centrul de Studi asupra Limbajului si Informației de la Universitatea Stanford există de mai multă vreme o deja celebră editură care publică lucrările esențiale ale domeniului.

5. Gramaticile nontransformationale în România

Prezența GNT în România poate fi discutată având în vedere două coordonate: cea a contribuțiilor ştiințifice şi cea a programelor curiculare.

Din primul punct de vedere, întâia contribuție (după cunoştința noastră, ce puțin) a venit din partea Adrianei Costăchescu ([14]). Adriana Costachescu este autorul unui studiu, din perspectiva GPSG (teorie care a precedat si inspira HPSG), asupra relației dintre coordonarea adversativă si subordonarea concesivă. Studiul a fost elaborat in 1993 şi publicat in 1996.

Lucrări de prezentare generală a diferitelor forme de GNT sau, dimpotrivă, de prezentare a trunchiului comun - unificarea - au fost publicate în ultimii şase ani de Adrian Atanasiu, Verginica Barbu, Ana-Maria Barbu, Florentina Hristea, Emi Ionescu şi Rodica Tătar.

Printre "pionierii" aplicațiilor acestor gramatici la limba română trebuie menționați Liviu Ciortuz şi cercetătoarea italiană Paola Monachesi. Amândoi au folosit teoria HPSG. Rolul lui Monachesi în stimularea aplicațiilor de acest tip la limba română trebuie în mod special subliniat. Studiile sale asupra criticelo pronominale din română au determinat o "mobilizare" a energiilor câtorva

cercetători români. Este vorba despre Ana-Maria Barbu, Emil Ionescu şi Amalia Todiraşcu.

Ana-Maria Barbu a aplicat HPSG în analiza elementelor gravitând în jurul verbului - adverbul de negaţie, semiadverbele, auxiliarele - şi a ajuns la concluzia că acestea sunt mai apropiate de afixe decât de cuvinte. Concluzia analizei se întâlneşte cu concluzia exprimată în lucrarea Valeriei Guţu Romalo, "Morfologie structurală a limbii române", în care formele compuse ale verbelor sunt considerate forme cu afix mobiL

O alta contribuţie a Anei-Maria Barbu priveşte ordinea constituenţilor in grupul nominal. Valorificând sugestiile de analiză ale lui Valerio Allegranza', Ana-Maria Barbu a propus o clasificare a constituenţilor grupului nominal, care este relevantă pentru problema ordinii acestora. Analiza produce astfel soluţii clare şi eficiente într-o problemă complicată de gramatică a limbii române.

Semnalând unele neajunsuri în analiza GB a fenomenului de anticipare clitică a complementului direct nominal în română, Verginica Barbu şi Emil Ionescu propun o abordare alternativă HPSG. Analiza poate fi extinsă şi la alte limbi care prezintă fenomenul în cauză. Analiza susţine că pronumele neaccentuate nu au un comportament uniform, proprietăţile lor depinzând de faptul dacă participă sau nu la structuri de dublare. Noutatea abordării vine din faptul că fenomenul anticipării obiectului direct este în mod ultim justificat prin proprietăţile lexicale ale verbului tranzitiv.

Un fenomen care, în aparenţă cel puţin, implică recursul la mecanismul deplasării - este vorba de prezenţa pronumelor neaccentuate în acuzativ în contexte în care ele nu sunt subordonate faţă de vreun element din acel context - este tratat într-un alt studiu asupra cliticelor pronominale româneşti[2] (). Studiul arată că ipoteza deplasării constituenţilor nu este necesară în analiza fenomenului. Este propusă în alternativă o analiză fără deplasări care captează toate proprietăţile fenomenului.

O analiză HPSG este propusă de asemenea pentru fenomenul negaţiei duble şi multiple în română [23]. în sfârşit, Amalia Todiraşcu abordează într-unui din studiile sale asupra limbii române, o categorie de dependenţe limitate (aşa-numitele *tough-constructions),* din aceeaşi perspectivă HPSG.

în aceeaşi linie a contribuţiilor ştiinţifice, merită amintită o intiativă instituţională: acreditarea de către CNCSIS,' în anul 2001, a Centrului de Lingvistică Computaţională de pe lângă Facultatea de Litere. Centrul este perechea universitară a Centrului de Studii Avansate în Inteligenţă Artificială. Apariţia sa a fost semnalată în buletinul european ELSNEWS.' Unul dintre programele de cercetare pe anul 2002 ale centrului are în vedere dezvoltarea aplicaţiilor de gramatici netransformaţionale la limba română.

în engleză, fenomenul este cunoscut sub numele de tlitic climbing", si este ilustrat în română de structuri de tipul Nu-l pot suferi pe Ion.

în planul programelor curiculare, GNT şi-au făcut loc mai g
întâmpinate uneori nu doar cu neîncredere, ci şi cu ostilitate. A e
fericire un sprijin substanţial şi constant al factorilor de decizie.
decanul Facultăţii de Litere, acad. prof. Dan Horia Mazilu, la recto
Bucureşti, prof. dr. Ioan Mihăilescu, la prorectorul aceleiaşi instituţi
Pânzaru, şi la acad. Dan Ioan Tufiş, directorul Centrului de Stud
Inteligenţă' Artificală al Academiei Române, cărora autorul aces
exprimă via şi profunda sa gratitudine, pentru susţinerea pe care a
în iniţiativele sale. Mulţumită acestui sprijin, au devenit realitate
care pot fi considerate succese:

- în programa cursurilor opţionale de limbă pentru an
  Facultăţii de Litere a fost introdus în 1996 un curs
  GPSG, iar din 1997 pînă în 2001 s-a ţinut un curs
  gramatici de unificare cu referire specială la HPSG.
- Din 1999, se predă la Facultatea de Matematică a
  Bucureşti un curs opţional de un an de prelucrare au
  naturale, în care un loc important îl ocupă gramaticile de
- Din 1997 pînă în prezent masteratul de lingvistică teore
  de Litere din cadrul aceleiaşi universităţi găzduieşte
  semestru de teorie HPSG aplicată la limba română.
- Din 1999, acelaşi masterat oferă un seminar de
  implementare computaţională.
- în anul 2000, un proiect de dezvoltare a componente
  computaţională în cadrul masteratului de lingvistică te
  sprijin de finanţare din partea Băncii Mondiale şi
  României, sprijin care a făcut posibile printre altele or
  cicluri de conferinţe pe teme de GNT (în special HPSG
  de Litere a Universităţii Bucureşti. Au conferenţiat Ivan
  tatea Stanford), Anne Abeille şi Daniele Godard (Uni
  7), Ştefan Muller (Universitatea din Jena), Robert Malou
  Groningen), Howard Gregory (King's College, Londra),
  (Universitatea Tubingen), toţi fiind personalităţi re
  domeniului. Mulţumită aceluiaşi program, cercetătorii
  petrece stagii de specializare la universităţile din Lille s
  au putut participa la manifestări reprezentative, cum
  UNESCO asupra spaţiilor virtuale şi multilingvismul
  (aprilie 2001), colocviul de gramatici bazate p
  Trondheim (august 2001), sau congresul de prelucra
  limbilor naturale de la Tokyo,' (noiembrie, 2001). Cea
  realizare legată de acest program, a constat însă în p
  mobilităţi studenţeşti, concretizate în vizitele de studiu
  masteratului de lingvistică teoretică, la universităţile
  Tubingen, Paris 7 şi Sienna.

6. Concluzii

Deşi GNT au pătruns în mediile ştiințifice din România mai târziu decât în alte țări, faptul că ele sunt prezente la noi este un lucru încurajator. Există tentația de a privi aceste eforturi de sincronizare cu mişcarea de idei din domeniul lingvisticii formale drept tentative mimetice şi superficiale. Este o greşeală gravă. Diversele comunități de lingvişti pot desigur ignora un curent, precum cel prezentat mai sus, dar aceasta este o atitudine, pentru a spune aşa, pe proprie răspundere. GNT şi teoria lingvistică pe care ele au inspirat-o şi-au făcut deja loc în lingvistica zilelor noastre şi au devenit una din paradigmele majore. în plus, dubla deschidere a acestor gramatici către psihologia cognitivă, pe de-o parte, şi către inteligența artificială, pe de altă parte, recomandă această paradigmă drept cadrul *privilegiat* de dialog interdisciplinar din ştiințele umaniste ale contemporaneității. Din acest triunghi, sunt aşteptate să apară noi aplicații - unele au şi apărut deja - care vor extinde într-un mod neaşteptat conceptul de lingvistică aplicată. Pentru toate aceste motive, tentativele de a păstra un contact viu şi de perspectivă cu comunitatea ştiințifică internațională a GNT reprezintă o investiție sigură pe temen lung.

Bibliografie

[1] Abeille, A. *Les nouvelles syntaxes. Grammaires d'unification et analyse du frangais,* Armând Colin, Paris, 1993

[2] Atanasiu, A. *Curs de lingvistică matematică,* Editura Universității Bucureşti, 1998

[3] Barbu, A.M. *Gramatici categoriale. Studiu comparativ cu gramaticile de constituenți,* "Limba Română", XVLI, 4-6, p 239-252, Ed. Academiei, 1997

[4] Idem, *Complexul verbal,* "Studii şi Cercetări Lingvistice", Ed. Academiei, sub tipar.

[5] Idem, *Romanian Determiners: Order and Classification,* "Revue Roumaine de Linguistique", Ed. Academiei, sub tipar

[6] Idem, *Funcțiile sintactice în Teoria X-Bară,* "Studii şi Cercetări Lingvistice", Ed. Academiei, sub tipar Barbu, A.M. şi E. Ionescu *Teorii gramaticale contemporane: Gramatica Centrilor de Sintagmă,* "Limba Română" 1 1996 31-55

[7] Idem, *Accusative Clitic Doubling in Romanian,* Liviu Ciortuz, Paola Monachesi, Hans Uszkoreit (editori) "Informai Proceedings of the GE&GL Workshop: Grammar Engineering and Grammar Learning", Tuşnad, România, 1997

[8] Barbu, V. *Despre gramaticile de unificare,* Analele Universității Bucureşti, seria limbă şi literatură română, 2001, p. 45-52

[9] Barbu, V. şi E. Ionescu *Anticiparea complementului direct în limba română în perspectiva HPSG,* Lucrările colocviului "Perspective moderne asupra limbii române", Bucureşti, Editura Universității din Bucureşti, (sub tipar)

[10] Borsley, R. *Syntactic Theory: A Unified Approach,* Edward Arnold, London, 1991

[11] Bresnan, J (editor) *The Mental Representation of Grammatical Relations,* MIT, Press, Ca. Mass, 1982

[12] Ciortuz, L. *An HPSG Kernel for Romanian,* manuscris, 1996

[13] Ciortuz, L, P. Monachesi, şi H. Uszkoreit (editori; *Informai Proceedings of the GE&GL Workshop: Grammar Engineering and Grammar Learning,* Tuşnad, România, 1997

[14] Costăchescu, A. *"Coordination" adversative et "subordination" concessive,* Iliescu, M. şi S. Sora, (editori), Rumänisch: Typologie, Klassification, Sprachcharakteristik, Mtinchen, 1996, p. 121-134

[15] Gazdar, G, E. Klein, G. Pullum şi I. Sag, *Generalized Phrase Structure Grammar,* Cambridge, Harvard University Press, 1985

[16] Gerlach, B. şi J. Grijzenhout (editori) *Clitics in Phonology, Morphology and Syntax,* John Benjamins Publishing Company, Amsterdam / Philadelphia, 2000

[17] Hristea, F. *Introducere în procesarea limbajului natural cu aplicații în PROLOG,* Editura Universității Bucureşti, Bucureşti, 2000

[18] Iliescu, M. şi S. Sora, (editori), *Rumänisch: Typologie, Klassification, Sprachcharakteristik,* Mtinchen, 1996, p. 121-134

[19] Ionescu, E. *A Type of SOV Construction in Romanian,* "Cahiers de Linguistique Theorique et Appliquee", tomes XXXII-XXXIII, 1995-1996, 19-39

[20] Idem, *Accusative Weak Pronouns in Romanian,* "Cahiers de Linguistique Theorique et Appliquee", tomes XXXII-XXXIII, 1995-1996,19-39

[21] Idem, *Accusative Clitic Doubling in Romanian,* "Cahiers de Linguistique Theorique et Appliquee" tomes XXXII-XXXIII, 1995-1996, 53-73

[22] Idem,, *Accusative Clitic Climbing in Romanian,* "Cahiers de Linguistique Theorique et Appliquee", tomes XXXII-XXXIII, 1995-1996, 74-87

[23] Idem, *A Quantification-based Approach to Negative Concord in Romanian* in Geert-Jan M. Kruijff and Richard T. Oehrle (editori), Proceedings of Formal Grammar Conference Utrecht,1999, p. 25-36

[24] Idem, *pro-Drop: An HPSG Account without Lexical Rules,* "Bucharest Working Papers in Linguistics", voi. I, nr.1, 1999, 117-124

[25] Idem, *On the Status of PE in the Direct Object Construction in Romanian,* Romanian Journal of Information Science and Technology, volume 4, numbers 3-4, 2001, p. 293-310

[26] Joshi, A. Introduction to Tree Adjoining Grammar, *Manaster Ramer, A. (editor)* The Mathematics of Language, *John Benjamins, Amsterdam,1987, p. 87-114*

[27] Kruijff, G-J. *M. and R. T. Oehrle (editori),* Proceedings of Formal Grammar Conference, *Utrecht, 1999*

[28] Manaster Ramer, A. (ed.) *The Mathematics of Language,* John Benjamins Publishing Company, Amsterdam, 1987

[29] Monachesi, P. *Clitic Placementin the Romanian Verbal Complex,* Gerlach and Grijzenhout (2000), p. 255-294.

[30] Pollard, C. şi I. A. Sag, *Information-based Syntax and Semantics,* CSLI, University of Chicago Press 1987

[31] Idem, *Head-driven Phrase Structure Grammar,* The University of Chicago Press, Chicago, 1994

[32] Shieber, St. *An Introduction to Unification-based Theories of Grammar,* CSLI, University of Chicago Press, 1986

[33] Tătar, D. *Inteligență artificială,* Editura Albastră, Cluj, 2001

[34] Todiraşcu, A. *Romanian **Tough**-Constructions,* Ciortuz, L, P. Monachesi, şi H. Uszkoreit (editori; *Informai Proceedings of the GE&GL Workshop: Grammar Engineering and Grammar Learning,* Tuşnad, România, 1997

[35] Wood, M. McGee, *Categorial Grammars,* Routledge London and New York, 1993

Către o teorie X-bar funcțională

Neculai CURTEANU
Institutul de Informatică Teoretică, Academia Română, Filiala Iaşi
curteanu@iit.tuiasi.ro

1. Teorii X-bar mai vechi şi mai noi

Scopul prezentei lucrări este dublu: **(a)** de a propune o nouă X-bar schemă, numită X-bar schemă *funcțională* şi *recursivă* (pe scurt, FX-bar schemă), mai generală şi mai adecvată decât cele existente, care să satisfacă cerințele unei abordări funcționale a limbajului natural (LN), în particular, ale strategiei lingvistice SCD (Segmentare-Coeziune-Dependență) [1], [2], şi *(b)* de a pune în evidență faptul că teoria FX-bar propusă poate reprezenta o posibilă (şi necesară) soluție la următoarea problemă ridicată de Noam Chomsky în teoria Minimalist Program [3]: în două capitole diferite, Chomsky afirmă (în două abordări diferite, aparent contradictorii, asupra structurii sintactice a LN) atât importanța crescândă a teoriei X-bar cât şi posibilitatea ca teoria X-bar standard să fie *"largely eliminated in favor of bare essentials"* (vezi secțiunea 5).

1.1. Teoria X-bar clasică

Printre (sub)teoriile care reprezintă substanța majoră pentru câteva teorii formale importante asupra sintaxei (LN), un rol fundamental este jucat de către aşa-numita teorie X-bar. X-bar schemele propuse sunt de obicei însoțite de definiții, ipoteze, restricții, principii şi alte (sub)teorii gramaticale care specifică într-o cât mai mare măsură modul concret în care X-bar schemele sunt utilizate pentru a construi structurile sintactice de bază ale LN. în general, teoria X-bar stabileşte categoriile gramaticale principale, proiecțiile lor lingvistice (minimale şi maximale), relațiile de dominare dintre categorii în cadrul acestor proiecții, sub-, co-, sau supra-ordonarea lor. Toate aceste aspecte asigură numai coloana vertebrală (infrastructura) consistentă a structurii sintactice în reprezentarea LN. Un capitol de o importanță deosebită este relația dintre teoria X-bar şi alte sub(teorii) sintactice şi semantice care formează întregul corpus al unei anumite teorii lingvistice.

Prima formă a X-bar teoriei este propusă de către Noam Chomsky în 'ucrarea *Remarks on Nominalizations* (1970) [4]. Chomsky scoate în evidență diferențele reale existente în următoarele sintagme nominale:

(1.1) *John 's criticism of the book*

(1.2) *John's criticizing the book;*

în special datorită şablonului verbal (similar cu al verbului *"criticize")* rezultat din gerunziul nominal (pentru engleză) *"criticizing",* în comparație cu forma nominală derivată *"criticism".*

Teoria X-bar originală propusă de Chomsky identifică trei categorii lexicale *primitive,* N [Eng: *noun],* V [Eng: *verb]* şi A [Eng: *adjective],* fiecare dintre ele cu câte două categorii sintagmatice corespunzătoare. Mai exact, utilizând notația X = N, V, A, categoria gramaticală X se întâlneşte ca *nucleu* [Eng: *head]* într-o categorie intermediară X' (sau X1, sau X^1), tradițional numită X-bar, precum şi într-o categorie maximală X" (sau X2, sau X^2), tradițional numită XP, reprezentând *proiecția maximală* a categoriei gramaticale X (lexicală sau nelexicală). Categoria X este numită *nucleul sintagmelor* X' (sau X1) şi X" (sau X2) care o conțin. Să mai notăm că prescurtarea pentru categoria *prepozițională* este P.

Ulterior au fost considerate *patru* categorii lexicale, bazate pe următoarele combinații ale celor două trăsături N şi V (considerate ca fiind generice pentru categoriile lexicale):

N	este o categorie X cu trăsăturile [+N, -V];
V	este o categorie X cu trăsăturile [-N,+V];
A	este o categorie X cu trăsăturile [+N, +V];
P	este o categorie X cu trăsăturile [-N, -V].

Teoria X-bar poate fi înțeleasă şi ca o specificare a modalității în care unele categorii gramaticale sunt *dominate* de către altele, deci ca o teorie a dominanței gramaticale (sau, aşa cum spune Chomsky, a *"guvernării),* care arată cum un nucleu (sau o categorie lingvistică) X se proiectează (se extinde) către categoriile mai complexe (structurile sintagmatice) X' (sau X1) şi X" (sau X2, sau XP). Structurile sintactice X1 şu X2 devin categorii gramaticale esențiale ale organizării şi reprezentării textului în LN.

Deci, X-bar teoria clasică consideră că X, împreună cu o secvență de *complemente* (sau *argumente,* notate Argj) este imediat dominată de X1, în timp ce Xl împreună cu o secvență de *specificatori* (notată Speq) este imediat dominată de către X2 (sau. XP). Utilizând binecunoscutele notații din domeniul teoriilor lingvistice formale, (X' = X1, X" = X2 = XP), categoriile lexicale şi gramaticale ale teoriei X-bar clasice a lui Chomsky sunt următoarele:

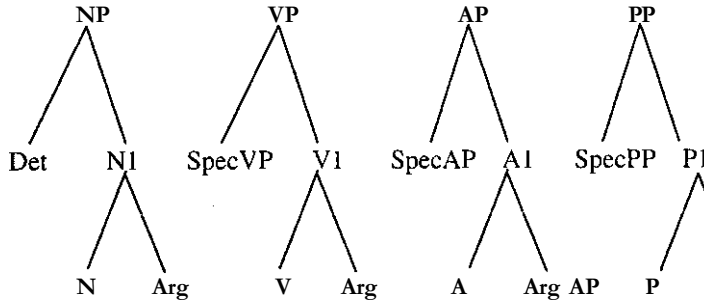


Figura 1.1. Proiecțiile categoriilor lexicale din teoria X-bar clasică

1.2. Extinderea teoriei X-bar la categorii non-lex

Stowell [5] propune ca teoria X-bar clasică să fie extin nelexicale sau *funcționale.* în particular, categoria gramaticală S Rom: *frază],* care corespunde uneia sau mai multor propoziții grama este văzută ca 12 sau IP, deci ca proiecția maximală a categoriei ne INFL [Eng: *Inflectionaf].* Nucleul nelexical I (INFL) reprezintă mulțim de flexionare atribuite nucleului lexical al clauzei-matrice (propoziția chiar una regentă) dintr-o frază, aşa cum sunt timpul, aspectul etc. fraze. Remarcăm *categoria* S, care introduce un anumit grad de analiza gramaticală, atât în engleză cât şi în română. Termenul realitatea lingvistică codificată de categoria S ar trebui să fie ac *gramaticală"* pentru engleză [Eng: *(grammatical) clause],* şi *gramaticală"* pentru limba română, cu două sorturi principale prescurtată CLF sau mai simplu CL, şi *clauză infinită,* prescurtată Cl

Astfel în extensia nelexicală a teoriei X-bar, S este proie maximală a categoriei virtuale (nelexicale) I, în timp ce S1 este văz sau CP, unde nucleul C este un *complementizator,* o categorie corespunde unei expresii (unui delimitator) sau unei sintagme ca clauză subordonată, e.g. pronume relativ, conjuncție, locuțiune con Teoria *X-bar extinsă* acreditează următoarele structuri:

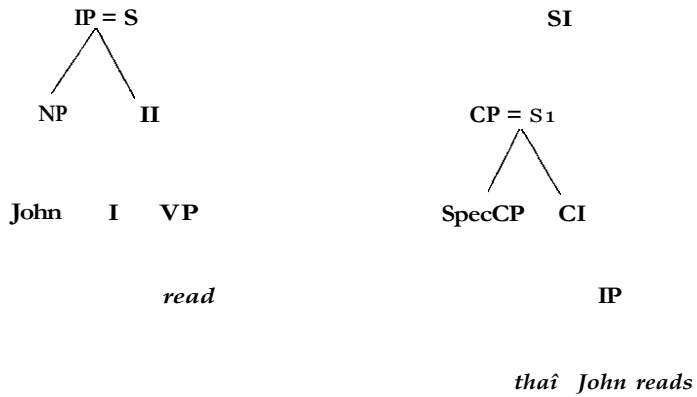


Figura 1.2. Teoria X-bar extinsă la categorii nelexicale

Sunt necesare câteva remarci:

(a) Teoria X-bar extinsă utilizează terminologia de "categorii nelexicale (sau funcționale)", prin care Stowell, Chomsky şi alți lingvişti definesc *noile nuclee* ale structurilor sintactice considerate. Categoria virtuală "I" este, desigur, una nelexicală, şi susține o anumită funcționalitate depinzând de categoria lexicală căreia îi este atribuită. Categoria C nu este, de obicei, nelexicală (exceptând situația, posibilă, când ea lipseşte) deoarece C corespunde unor categorii gramaticale lexical nevide. în ceea ce priveşte funcționalitatea lui C, suntem de acord că C corespunde într-adevăr unor funcții şi relații sintactice şi semantice importante pe care le numim *marcheri de propoziție* [1], [2], [6], uneori incluşi în clase mai largi cum sunt cea a *marcherilor de discurs* [7], reprezentând în acelaşi timp şi un element (deci o relație) de co-referință în cadrul fenomenului de *legare,* şi/sau o *"barieră"* [8] în cadrul *teoriei limitării* [9]. Aceste aspecte multi-funcționale ale categoriei C nu sunt contradictorii ci doar complementare, întregind un tablou complex al funcționalității lexical-semantice pentru o categorie lingvistică atât de specială cum este C.

(b) A doua observație este dedicată rolului unor categorii nelexicale în cadrul X-bar schemelor extinse. Din Fig. 1.2. reiese că subiectul NP are rolul (nesigur) al unui specificator pentru S = IP, în timp ce VP reprezintă complementul categoriei virtuale I. De asemenea, S1 = CP se consideră a fi proiecția maximală a categoriei C, în timp ce complementul sintagmei CP este IP. Admițând că în engleză, din punct de vedere sintactic, această supoziție are sens deoarece categoria C reprezintă nucleul acestor sintagme, în alte limbaje, inclusiv româna, acest lucru este nedecis, în special din perspective semantice şi funcționale. Unele abordări funcționale ale acestor probleme sunt discutate în mai multe lucrări, dar

ne vom restrânge să menționăm aici soluțiile oferite de către *teoria gramaticii funcționale* [10] şi *strategia lingvistică* SCD [1], [2], [6]. Un interes special prezintă abordarea *lexicală* (inclusiv *funcțională)* a teoriei X-bar ca subteorie de bază în cadrul teoriei sintactice HPSG [Eng: *Head-driven Phrase Structure Grammar]* [11]. O analiză comparativă cu *FX-bar schema* propusă în această lucrare va fi făcută într-o lucrare viitoare.

1.3. X-bar schemele din teoria GB

X-bar schemele propuse de teoria *Government and Binding* (GB) a lui Chomsky [5] sunt următoarele:

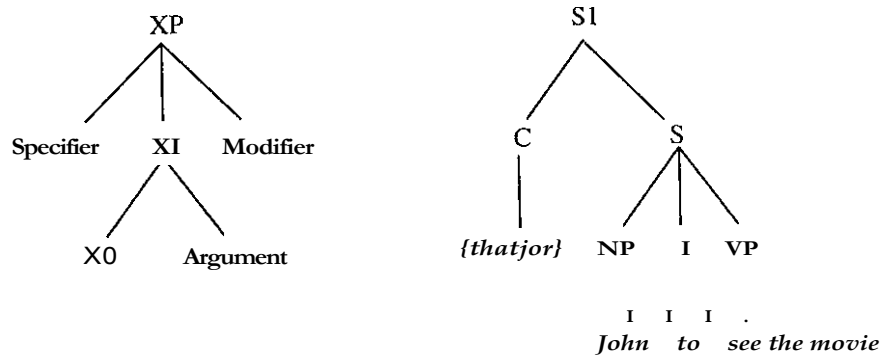


Figura 1.3. X-bar schema generală din GB, X = N, V, A, P, S

în teoria GB există următoarele X-bar *echivalențe* pentru proiecțiile categoriilor gramaticale (lexicale şi nelexicale).

Tabelul 1.3

X	XI	X2
N	NI	NP
V	VI	VP
A	A1	AP
P	PI	PP
I	S	SI

în lucrările GB [5] si cele care urmează, Chomsky consideră categoria I ca fiind nucleul lui S, iar complementizatorul C ca fiind nucleul iui S1. în subsecțiunea următoare teoria sintactică GPSG a lui G. Gazdar [12] face un important pas

înainte către lexicalitate şi către utilizarea explicită a trăsăturilor lingvistice atribuite categoriilor gramaticale.

1.4. Teoria X-bar în GPSG

în teoria lingvistică GPSG [Eng: *Generalized Phrase Structure Grammar]* [12], [13] etc, (sub)teoria X-bar joacă de asemenea un rol central, o sintagmă a LN fiind definită ca proiecția *trăsăturilor lingvistice* atribuite *nucleului* [Eng: *head]* acelei sintagme. Informația cuprinsă în trăsăturile nucleului determină caracteristicile principale ale comportamentului sintactic al sintagmelor LN. Reamintim că o categorie sintactică în GPSG se reprezintă ca o mulțime de perechi *<trăsătură, valoare>.* De exemplu, eticheta NP [Eng: *noun phrase]* (sau N2), prin care se notează o sintagmă nominală, reprezintă o abreviere pentru mulțimea {<N, +>, <V, ->, <BAR, 2>}, unde BAR este *numele trăsăturii ce* codifică *nivelul de proiecție* a categoriei sintactice N = {<N, +>, <V, ->}. Trăsătura BAR poate lua valorile 0, 1, 2. Teoria GPSG consideră N, V, A şi P ca fiind *categorii sintactice majore.* Toate celelalte sunt considerate de GPSG ca fiind *categorii minore:* determinatori, complementizatori, marcheri, cuantificatori, alte particule etc. Categoriile majore sunt considerate de către teoria GPSG ca având întotdeauna o *valoare* pentru *trăsătura* BAR. Valoarea BAR pentru categoriile minore nu este definită niciodată în GPSG.

Teoria sintactică a GPSG aduce câteva elemente noi şi interesante comparativ cu teoria GB: **(a)** X-bar schemele au, ca şi în GB, trei nivele de proiecție (valorile trăsăturii BAR); **(b)** Pentru economia reprezentării, GPSG propune ca în X-bar schemele de bază, nivelul proiecției lingvistice să fie conservat când se trece de la nucleu către expresiile subcategorizate, mai puțin în cazul în care acest lucru se face prin (alte) reguli explicite; **(c)** Printr-un mecanism de *moştenire implicită,* nivelele BAR de proiecție a nodului-rădăcină şi ale nodurilor-fiice rămân aceleaşi, mai puțin în cazul în care există o indicație contrară expresă.

O altă caracteristică este aceea că în GPSG nu se întâlnesc categorii abstracte, non-lexicale, cum ar fi "I" (INFL) din GB. Acest lucru este posibil deoarece în GPSG, pentru aceste categorii nelexicale, nu există un nivel de proiecție pe care ele să fie reprezentate (sub nivelul lexical BAR = 0). Consecința este aceea că, în GPSG, S este proiecția unei categorii V. Mai exact, proiecțiile maximale ale lui V sunt VP, S, şi S1, depinzând de următoarele valori luate de către trăsăturile SUBJ şi COMP (= complementizator = C):

V[BAR 2][SUBJ-][COMP NIL] = VP;

V[BAR 2][SUBJ +][COMP NIL] = S;

V[BAR 2][SUBJ +][COMP a] = S1; unde a *e {that, for, whether, if).*

în sfârşit, trebuie să remarcăm că GPSG trebuie să rezolve problemele întâlnite în mod obişnuit în formalismele gramaticale bazate pe unificarea lingvistică (şi/sau logică), de exemplu PATR-II [14], HPSG [15], [16] etc. O astfel

de problemă este, în particular, transmiterea informației despre *timpul verbului* între forma flexionară codificată de verb şi nodul S. Pentru teoriile lingvistice care permit inserarea în arborele de derivare a cuvintelor flexionate, aşa cum este cazul cu GPSG, HPSG etc, informația despre forma flexionară trebuie să poată fi mutată în ambele direcții pe nivelele X-bar schemei. Din aceasta derivă, în GPSG, condiția ca V să fie *nucleul structurii clauzale* care corespunde categoriei S. Pe de altă parte, în GB, informația asupra timpului unui verb poate fi transmisă dinspre nodul I către proiecția sa în S înainte ca I să fie combinat cu forma flexionată a verbului din S. Această situație poate produce potențiale dificultăți procedurale şi de reprezentare.

Este important de menționat că proiecțiile categoriilor din Tabelul 1.3 rămân aceleaşi pentru GPSG şi LFG [Eng: *Lexical Funcțional Grammar] (vezi* de exemplu [13]), cu diferența notabilă că prima celulă din ultima linie a Tabelului 1.3 este goală, deoarece în aceste două teorii lingvistice (ca şi în altele), categoria virtuală I lipseşte.

1.5. O formulare recursivă a X-bar schemelor din teoria Tbarr

Vom propune în această subsecțiune o *formulare recursivă* a teoriei X-bar avându-şi originea în *teoria barierelor* (TBarr) [8], [17] şi fiind compatibilă cu teoria sintactică a *Programului Minimalist* (MinP) [3] şi cu modelul său gramatical din *Principii şi Parametri* (P&P) [3]. în conformitate cu MinP şi P&P, gramaticile concrete ale limbajelor naturale (LN) reale pot fi modelate de mulțimi de parametri şi valorile lor, care specifică principii şi teorii lingvistice universal valabile. Pentru o asemenea setare (asignare) a valorilor parametrilor, relațiile de precedență (de ordonare liniară) dintre categoriile gramatice sunt obținute din proprietăți ca marcarea cazuală, atribuiri de roluri tematice *((0* - roluri şi *8* - marcheri), împreună cu alte relații şi marcheri ce se aplică la nivelul sintagmelor, clauzelor, şi unităților de discurs. Din acest motiv, relațiile de precedență pentru X-bar schemele propuse pot fi utilizate independent pe arborii sintactici considerați, informația de ordonare (liniară) a categoriilor fiind dată de următorii *parametri de precedență.*

(OrdPar) *Un anumit parametru (depinzând de limbaj) precizează dacă secvența de specificatori precede sau succede nucleul, iar un alt parametru (depinzând de limbaj) precizează când secvența complementelor precede sau succede nucleul din X-bar schemă.*

De exemplu, în *engleză,* specificatorii preced de obicei nucleele lor nominale, în timp ce în *română,* în mod normal, ei succed nucleelor. în general complementele (argumentele) succed nucleele lor şi în engleză şi în română. Un caz special al argumentului este *subiectul* (sintactic). Această exprimare a (OrdPar) poate fi încă particularizată în funcție de categoriile lexicale concrete, din LN concrete. De exemplu, atât în română cât şi în engleză, când o sintagmă

adjectivală (adverbială) este *predicațional activă,* fiind urmată de anumite argumente (complemente sau adjuncți), atunci este obligatoriu ca ea să succeadă propriul nucleu şi nu să îl preceadă.

Consecința principală a parametrizării dependentă de limbaj a precedenței categoriilor lingvistice este că în exprimarea teoriilor lingvistice se pot utiliza arbori neordonați, iar principiile propuse de teoria X-bar primesc un puternic caracter de independență relativ la regulile de dominare ale structurilor sintagmatice. Este important faptul ca X-bar schemele obținute în cadrul teoriei X-bar considerate să asigure proiecții adecvate ale categoriilor lexicale, permițând inserarea *adjuncților,* obținerea categoriilor de proiecție maximală, şi acceptarea faptului că unele proiecții minimale sau maximale din *structura de adâncime* pot fi vide (deci noduri care să domine *categorii vide),* conform [9], [8], [17].

Fiind stabilit *principiul* (OrdPar), teoriile GB şi Tbarr consideră următoarele trei nivele ale proiecției din teoria X-bar, sintetizate de următoarele *reguli (principii)* şi de *X-bar schemele* corespunzătoare:

(PXO) Fiecare nod XO dintr-o schemă X-bar este fie vid, neavând nici o trăsătură, fie este nodul-mamă al unui element lexical a cărei categorie gramaticală şi trăsături sunt specificate la nivelul lexiconului.

XO X0[F]

NIL *lex-item[F]*

Figura 1.5.1. Nodul XO în TBarr

(PX1) Fiecare nod X1 (X' sau X¹) având trăsăturile lexicale F este fie nodul-rădăcină al exact unui nod X (care este *nucleu)* cu trăsăturile F şi al unei secvențe de noduri XP (care sunt *complemente,* sau *argumente),* fie este rădăcina unui nod identic X1 împreună cu exact un nod XP (care este *adjunct).*

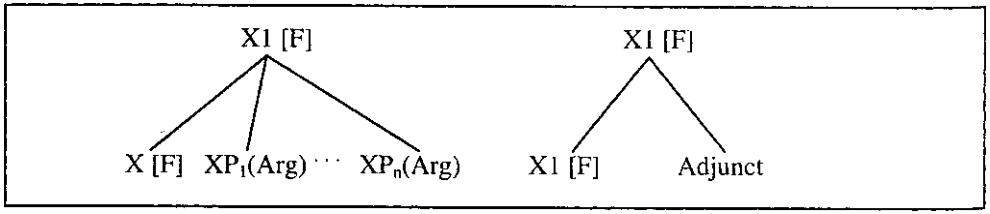


Figura 1.5.2. Nodul X1 în TBarr

(PX2) Fiecare nod XP care are trăsăturile lexicale F trebuie să satisfacă una şi numai una din următoarele condiții: **(i)** XP este un nod-frunză (nu mai are nici un nod-fiică) şi mulțimea F este vidă; **(ii)** XP este rădăcina unei secvențe de XPs *(specificatori)* şi a exact unui nod X1 moştenind trăsăturile F; **(iii)** XP este rădăcina unei secvențe de XPs *(complemente,* sau *argumente)* şi a exact unui nod X cu trăsăturile F; **(iv)** XP este rădăcina unui alt nod XP moştenind trăsăturile F şi a exact unui nod XP.

O *observație importantă* este aceea că unele dintre *secvențele* XP specificate în regulile (PX1) şi (PX2) pot fi *vide.*

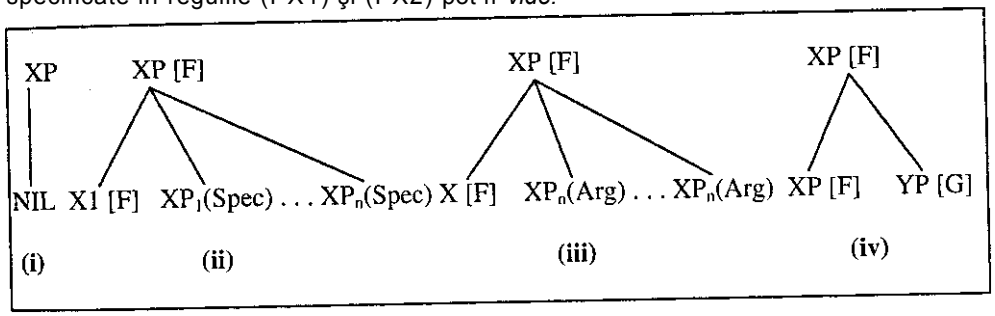


Figura 1.5.3. Nodul X2 în teoria TBarr

Combinând recursiv X-bar schemele rezultate din *regulile* (X_no)-(X_{n-1})-(Xp2) se pot obține toate structurile sintactice întâlnite în *X-bar teoria clasică* şi *extinsă*

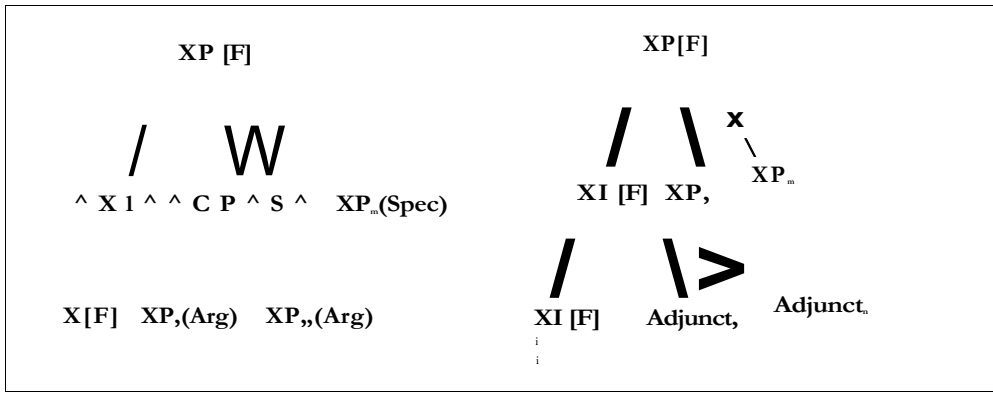


Figura 1.5.4. Formele generale (şi recursive) ale X-bar schemelor din TBarr

2. X-bar teoria din modelul P&P al teoriei MinP

2.1 Sistemul Chomskyan al gramaticii universale

Această subsecțiune conturează câteva aspecte implicate de către teoria X-bar în cadrul teoriilor MinP *(Minimalist Program)* şi P&P *(Principles and Parameters)* [3]. Pentru a înțelege contextul, este necesar să schițăm *teoria* lui *Chomsky a* gramaticii universale UG [Eng: *Universal Grammar]* şi a relațiilor sale cu abordarea MinP bazată pe P&P. Sunt introduse următoarele concepte de UG.

Capacitatea utilizării şi înțelegerii LN se bazează în esență pe proceduri care pot genera obiecte numite *descrieri structurale* (SDs). SDs sunt *expresii* de limbaj. Teoria unui LN particular constituie gramatica acestuia, în timp ce teoria tuturor limbajelor şi a expresiilor pe care le generează ele reprezintă *Gramatica Universală* (UG).

Se consideră că UG specifică anumite *nivele lingvistice,* sau sisteme de reprezentare a informației lingvistice. UG a lui Chomsky [3] presupune că fiecare SD este o secvență (8, a, *n, X)* de *patru* reprezentări pe următoarele nivele, respectiv: *structură de adâncime* (D-structură), *structură de suprafață* (S-structură), *formă fonetică* (PF) şi *formă logică* (LF). O ipoteză constructivă pentru UG este aceea că limbajul este scufundat în *sisteme de performanță* care permit ca exprimări în LN să fie folosite pentru articulare, interpretare, referire, interogare, reflecție şi alte acțiuni, în timp ce SDs devin un complex de instrucțiuni pentru aceste sisteme de performanță.

O altă ipoteză standard pentru construcția UG este aceea că un LN este format din două componente: un *lexicon* şi un *sistem computațional.* Această construcție este o inovație esențială comparativ cu teoria GB, care pretinde independența sa față de orice aspecte computaționale sau de implementare. Lexiconul specifică elementele de intrare pentru sistemul computațional, în timp ce acesta foloseşte intrările de lexicon pentru a genera derivări şi SDs. Derivarea unei exprimări lingvistice particulare implică alegerea elementelor din lexicon şi evaluarea, construind perechea pe două nivele de performanță, numite şi *reprezentări de interfață.* Una din ipotezele de bază ale teoriei lui Chomsky *Minimalist Program* este aceea că în construcția SD, utilizând lexiconul şi sistemul de evaluare, sunt luate în considerare *numai două* nivele de interfață, corespunzând lui PF (formă fonetică) şi lui LF (formă logică), împreună cu mulțimile de perechi *(n, X)* rezultate din cele două forme.

în abordarea P&P a teoriei lingvistice MinP, UG asigură un *sistem de principii* fixat, asociat cu un tablou finit de *parametri evaluați* (pe un număr finit de valori). Regulile pentru un LN particular se reduc la alegerea valorilor pentru aceşti parametri. Noțiunea de construcție gramaticală este eliminată, împreună cu regulile particulare de construcție, specifice gramaticilor generative. Construcții ca

VP, clauză relativă, pasivul etc. devin doar elemente ale unei taxonomii generale, sau colecții de fenomene explicate prin interacțiunea principiilor de UG, legate (setate) cu anumite valori fixate ale parametrilor.

în sistemul computațional al UG există un set de *principii invariante,* fiecare cu un domeniu de *opțiuni* restrânse la elementele funcționale şi proprietățile generale ale lexiconului. O *selecție* Z printre aceste opțiuni determină LN concret. în schimb, un limbaj determină o mulțime infinită de SDs lingvistice, fiecare pereche *(n, X)* fiind obținută din nivelele de interfață (PF, LF), respectiv. *Achiziția de limbaj* implică fixarea mulțimii 2, în timp ce *gramatica* limbajului se reduce la specificarea lui 2. în fine, un *sistem de parsare* care este invariant şi neantrenat (cum adesea se presupune) poate fi văzut ca o transformare a perechii (I, TI) într-o schemă structurată similară cu o SD. Condițiile asupra reprezentărilor LN impuse pentru diferite principii şi (sub)teorii, cum ar fi teoria *legării,* teoria *cazurilor,* ^-teoria etc, sunt satisfăcute pe nivelele de interfață ale sistemelor de performanță. Toate aceste ipoteze fac parte din teoria MinP a lui Chomsky şi din construcția sa pentru UG.

2.2 (Sub)teoria X-bar în contextul teoriei MinP

Sistemul computațional al unui LN concret preia reprezentările unei forme date şi le modifică, în timp ce UG trebuie să furnizeze mijloacele de a reprezenta o mulțime de elemente din *lexicon* într-o formă care să poată fi accesată şi procesată de către sistemul computațional. Forma sub care este accesat lexiconul de către sistemul computațional poate fi considerată ca fiind o anumită *versiune* a teoriei X-bar. Schemele X-bar pot fi asociate în mod natural cu *structuri de trăsături lingvistice* [18], ca un *tip de date lingvistice* standard şi invariant pentru a reprezenta şi a procesa LN eficient. în strategia SCD, *schemele X-bar augmentate* [19] considerate până acum nu sunt doar tipuri de reprezentare a datelor la nivelul lexiconului ci ele pot asigura structurile invariante fundamentale pentru a reprezenta şi a procesa textul în LN la nivel sintactic [1], [2], [6].

în teoria *Minimalist Program* şi modelarea P&P a UG, proprietățile şi relațiile esențiale sunt formulate în termenii simpli şi elementari ai *teoriei X-bar.* Astfel, o *structură* X-bar este compusă din *proiecțiile* lingvistice ale *nucleelor* selectate din lexicon. în *schema X-bar* a teoriei MinP reprezentată în Fig. 2.2.1. sunt prezente două relații locale: relația *Specificator-Nucleu* de la ZP la X, şi relația *Nucleu-Complement* de la X şi YP (ordinea categoriilor nu este esențială, fiind stabilită de către parametri P&P adecvați de ordonare). Relația Nucleu-Complement (Nucleu-Argument) nu este numai "locală" ci şi fundamentală deoarece este asociată *(8) relațiilor tematice.*

Dacă, pentru moment, nu este luată în considerare *relația de adjuncție* sau adjuncții se consideră a se afla printre argumentele-complemente, X-bar structurile pot fi reduse la X-bar schema din Fig. 2.2.1, cu următoarele specificări: (a) Sunt considerate numai relațiile locale (deci nici o relație de proiecție între X şi vreo sintagmă inclusă în proiecțiile maximale YP sau ZP); **(b)** Relația *Nucleu-Complement* reprezintă *relația locală de nucleu* [Eng: *core relation];* **(c)** O relație locală *admisibilă* a schemei X-bar din MinP este cea *Nucleu-Nucleu.* De exemplu, relația unui verb predicativ cu nucleul predicațional (deverbal) al unei sintagme nominale pe care o subcategorizează; **(d)** O altă relație în X-bar schema din MinP este *legătura de lanț* [Eng: *chain* //n/c], corespunzând unui *lanț de dominare* sau de *guvernare.*

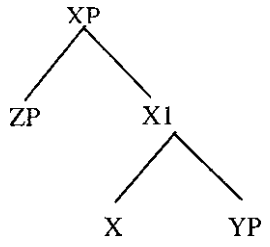


Figura 2.2.1. Schema X-bar din teoria MinP

Guvernarea realizată de nucleu joacă un rol central în toate componentele teoriei MinP asupra UG. Una dintre problemele-cheie este asignarea corectă a trăsăturilor nucleului. în HPSG şi SCD, de exemplu, acest lucru este realizat la nivel lexical (BAR = 0), după aplicarea flexionării, cât şi la nivel de lexicon (nivel de proiecție notat convențional cu BAR = -1) pentru clasa categoriilor lingvistice cu *proprietăți funcționale (predicaționale, relaționale),* fie ele verbe, substantive, adjective, marcheri de sintagmă, marcheri de discurs etc. care antrenează un comportament sintactic funcțional [2], [6]. în particular, pentru teoria MinP, subteorii ca *S-guvernarea* şi *guvernarea de caz,* corespunzând *6-marcării* şi *Caz-marcării,* sunt cele mai importante forme de dominare. Un studiu comparativ al guvernării categoriilor (dependență, dominare), relație prezentă firesc în cele mai importante teorii sintactice formale existente în acest moment, este inclus în [20].

Structurile propuse de teoria X-bar trebuie "animate" de către (sub)teoriile (de asemenea complementare) conținute în MinP şi P&P, şi care explicitează fenomenele de *guvernare, legare, limitare* etc. ce s-au dovedit a fi importante pentru orice teorie lingvistică deoarece ele asigură reguli pentru organizarea *lexiconului* şi a *sistemului computațional* care generează şi recunoaşte SDs.

De exemplu, în funcționarea *teoriei cazurilor in* contextul schemelor X-bar din MinP, ipoteza standard din MinP este aceea că, într-o frază (propoziție), relația *Specificator-Nucleu* atrage după sine *cazul structural* pentru *poziția de subiect,* în timp ce *poziția de obiect* primeşte cazul sub guvernarea nucleului V, incluzând construcții în care obiectul marcat cazual de către un verb nu este complementul său ci doar un adjunct (aşa-numita *marcare de caz excepțională).*

în continuare este prezentată structura X-bar de bază a *clauzei* în teoria MinP, cu următoarele notații uzuale: C = COMP = Complementizator, T = Timpul, Agr_s = acordul subiectului; Agr_o = acordul obiectului etc.

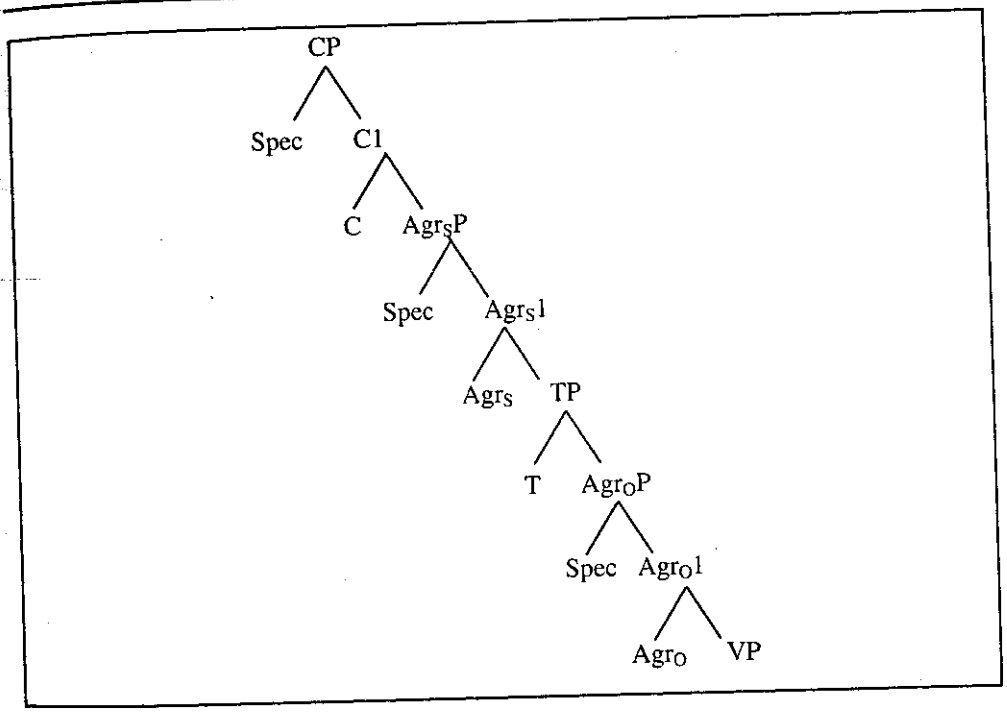


Figura 2.2.2. X-bar structura clauzei în teoria MinP

Schemele X-bar clauzale clasice din Fig. 1.2. şi Fig. 1.3. sunt expandate în Fig. 2.2.2., cu următoarea posibilă interpretare funcțională: X-bar schema MinP are ca nucleu VP, care îşi selectează sintagma-Obiect (sau argument, mai general) prin acord şi marcare, afectată apoi de Specificator. Un timp finit T aplicat sintagmei Verb-Obiect generează sintagma TP [Eng: *tensed phrase],* căreia i se aplică apoi aceleaşi funcții de selecție a subiectului (acord, marcare, specificare), generând sintagma Verb-Obiect-Subiect, care este de fapt clauza finită simplă (notată S). în fine, prin aplicarea asupra lui S (văzută ca sintagmă Agr_sP) a unui complementizator C (sau marcher clauzal, marcher de discurs etc.) se obține o clauză "completă" ce poate, prin recursie, să genereze orice frază [Eng: *sentence].*

Alte exemple de X-bar scheme bazate pe MinP şi P&P, ce pot fi discutate în contextul mai general al fenomenelor de guvernare sunt date de Fig. 2.2.3. care urmează.

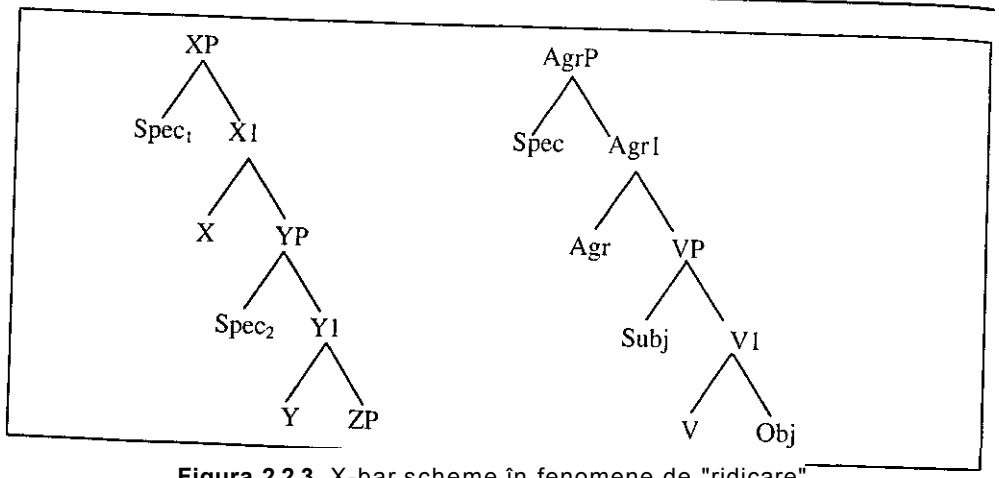


Figura 2.2.3. X-bar scheme în fenomene de "ridicare" la nivel de Spec în MinP

Concluzia este aceea că teoria X-bar din MinP sintetizează relațiile fundamentale de dependență, descrise de X-bar schemele propuse, şi implicate în procesele de organizare a lexiconului şi a sistemului computațional din UG. X-bar teoria în abordarea MinP reflectă în principal *aspectele statice* întâlnite în fenomenele de *guvernare* (c-comandă, m-comandă, bariere, categorii de blocare etc), în *teoria legării* şi în procesele de *referință-coreferință,* în stabilirea dependențelor la mare distanță (extra-clauzale) etc Nu vrem să intrăm în detalii şi să explicităm mecanismele de lucru ale X-bar schemelor considerate, ci mai curând să atragem atenția asupra *teoriei X-bar* ca o *componentă fundamentală* a unei teorii lingvistice noi şi elaborate cum este MinP şi modelul său P&P [3].

Teoriile MinP şi P&P nu reprezintă un punct-terminus pentru evoluția teoriei X-bar. Dimpotrivă, asigură o bază de pornire pentru o strategie radical diferită în care Chomsky examinează cele mai serioase argumente pentru a abandona (!) teoria X-bar [3; Cap. *Categorii şi transformări].* Această alternativă şi consecințele sale sunt discutate în secțiunea 5, şi ar trebui să reprezinte una dintre cele mai importante provocări prezente pentru domeniul analizei şi proiectării teoriilor lingvistice [21].

Unul dintre principalele scopuri ale secțiunii care urmează este de a introduce propunerea noastră de *scheme X-bar funcționale (scheme FX-bar)* în cadrul strategiei lingvistice SCD. Propunerea noastră o considerăm a fi o poziție pragmatică şi echilibrată în *direcția* teoriei X-bar, atrăgând atenția asupra adevăratului său rol şi oportunităților computaționale din lingvistica reală, înțelegerea corectă a aspectelor *statice* şi *dinamice* ale acestei versiuni a teoriei X-bar ar trebui să fie de asemenea o consecință a unei priviri cuprinzătoare a întregului context al teoriilor lingvistice care stabilesc principiile de dependență,

pasele de marcheri, categoriile şi ierarhiile, regulile de referire şi structurare, în Strânsă relație cu formele şi regulile de construcție ale (sub) schemelor FX-bar.

3. Scheme X-bar funcționale şi strategia lingvistică SCD

în [19], în contextul *strategiei lingvistice SCD (Segmentare-Coeziune-Dependență)* [22], [19], [1], [2], [6], este definită o clasă de *scheme X-bar augmentate (scheme AX-bar),* scheme destinate a reprezenta *invarianți sintactici generali de* reprezentare şi operare cu structurile gramaticale ale LN, în particular pentru limba română, ca soluție la problemele de analiză şi generare automată a LN. *Schemele FX-bar (funcționale)* propuse aici completează şi extind *schemele AX-bar* [19], şi pot fi interpretate în mai multe moduri: (1) din punct de vedere *static,* schemele FX-bar pot furniza câteva de tipuri fundamentale de date pentru reprezentarea informației lingvistice în structuri de trăsături lingvistice, standardizate şi tipizate; (2) din punct de vedere *dinamic,* schemele FX-bar pot codifica informația lingvistică în formă procedurală ca *funcții şi relații standard* ce sunt (recursiv) apelate în cadrul proceselor de analiză şi generare a LN; (3) schema FX-bar generală poate fi de asemenea interpretată şi utilizată ca un automat pe baza căruia să se realizeze o analiză *on-line* a textului unei fraze, cuvânt cu cuvânt.

3.1. Câteva preliminarii asupra SCD

Sunt necesare unele precizări asupra noțiunilor şi notațiilor cu care lucrează *strategia lingvistică* SCD. Unul dintre elementele importante este că nivelul 2 (BAR = 1) în X-bar schema clasică joacă un *rol-cheie* în SCD pentru construcția structurilor sintactice, şi este utilizat sub numele de *grup nominal* (NG), *grup verbal* (VG), *grup adjectival-adverbial* (AG), în general XG, pentru X = N, V, A. *Grupul XG* corespunde *proiecției lexicale* X1, cu X = N, V, A, şi *clauzei minimale* CLO, în X-bar schema fundamentală propusă în Fig. 3.2.1.

Să menționăm că orice XG (X1) este un XP (X2), dar nu şi invers, deoarece proiecția categoriei X în cazul XG lucrează numai pentru nivelul BAR ^1. SCD face de asemenea distincție între câteva tipuri de NGs (NGs elementare, Predicaționale, non-predicaționale, etc), VGs (VGs la un timp finit şi la un timp non-finit) etc

O altă trăsătură esențială şi specifică a SCD este un tratament adecvat al *Proprietăților funcționale* ale categoriilor lingvistice, ca şi al tuturor *categoriilor Naționale* şi sintagmelor (expresiilor) de *discurs.* Mecanismul utilizat pentru a obține acest lucru se bazează pe *clase de marcheri lingvistici* şi *ierarhiile* lor [1], 12]i [6], [7], [45]. Câteva observații se impun:

(a) *Marcherii* din SCD, numiți *marcheri de structuri sintagmatice* (PS-Ms) [Eng: *phrase-structure markers],* sunt cu totul diferiți de ceea ce teoria lui Chomsky numeşte formal *"marcheri de sintagmă"* [Eng: *phrase-markers]* în [17], sau *T(ree)-marcheri* în [3]. *Marcherii Chomsky* sunt definiți ca *"tăieturi orizontale"* (sau "factorizări") în cadrul unui arbore de derivare, sau ca fiind arborele însuşi. Mult mai apropiați de ceea ce sunt PS-Ms în HPSG [16], *marcherii de structuri sintagmatice* (PS-Ms) din SCD sunt acele categorii lexicale şi nelexicale care se aplică cuvintelor şi structurilor sintagmatice (PSs) cu scopul de *evidenția,* de a *marca,* anumite funcții şi relații sintactice şi semantice pe care PSs respective le joacă în cadrul unei exprimări. Punerea în evidență a anumitor funcții care se aplică PSs se referă la (cel puțin) câteva elemente: *tipul funcției* (sintactic, semantic, relațional, logic, pragmatic, discursiv etc), *locul,* în text, *unde începe aplicarea* funcției sau relației, şi *domeniul (domeniile, conexe sau nu)* de aplicare a funcției sau relației (limitele textuale între care se aplică).

Exemple tipice de PS-Ms din SCD sunt: (a) trăsăturile *predicative* generate de către *categoriile predicaționale* (de fapt, verbe, substantive, adjective şi adverbe predicaționale); *(b)* acele *mijloace gramaticale* prin care sunt introduse *noi* NGs (grupuri nominale în limbajul SCD), VGs, AGs (Caz-marcarea, acordul, gradele de comparație, etc); (c) acele *categorii* şi *expresii* (numite şi *marcheri de discurs)* care introduc *noi* clauze; (of) PS-Ms care introduc *proprietăți relaționale* asupra PSs şi clauzale (de exemplu de *marcheri* de tip logic cum sunt structurile *dacă-atunci-altfel, deoarece,* etc, dar şi marcheri de tip sintactico-semantic cum sunt aceia care introduc *categorii* şi *clauze subordonate* etc)

(b) SCD se aseamănă din unele puncte de vedere cu abordarea [16] a HPSG şi, parțial, cu [15], care exploatează, pentru prima o$ră în clasa teoriilor lingvistice bazate pe *gramatici* de PSs (PS-Gs), într-o mult mai mare măsură, categoria lingvistică a marcherilor PS-Ms. în [16], Pollard & Sag "postulează o nouă parte a *marcherilor* de discurs,... ce se remarcă ... printr-un nou atribut al categoriilor (în plus față de NUCLEU şi SUBCAT) numită MARKING, cu valori din sortul *marking".* Teoria HPSG enunță PRINCIPIUL MARCĂRII [16, p. 400] după cum urmează:

"într-o sintagmă cu nucleu, valoarea trăsăturii MARKING este lexical-identică cu cea a trăsăturii MARKER-DAUGHTER dacă aceasta există, şi cu cea a trăsăturii HEAD-DAUGHTER în caz contrar.

Modul în care HPSG [16] pune la lucru PS-Ms reprezintă un bun şi esențial pas înainte, deşi credem că nu exploatează îndeajuns potențialul funcțional şi relațional al diferitelor clase de marcheri şi ierarhiile acestora (aşa cum face strategia SCD, vezi şi [7], [45]).

(c) Continuând şi extinzând construcția limbajului, ca o expresie de convergență între gramatica categorială şi *Minimalist Program,* Chomsky [3] consideră *transformările generalizate* (GTs) şi concepe un demers de înlocuire a

X-bar teoriei, ce explică în Programul Minimalist structura constituenților (sintagmatici) complecşi, prin GT *Merge* care construieşte obiecte sintactice pornind de la obiecte sintactice simple (de exemplu, *"speaks"* şi *"French"* sunt "reunite" într-un nou obiect sintactic *"speaks French"* etc). Mai multe formalizări ale acestui nou curent al ideilor lui Chomsky pot fi găsite în cadrul gramaticilor logice multi-modale şi de tipuri categoriale, *e.g.* [21], [23], [24] etc (vezi şi secțiunea 5).

(d) Dintr-o perspectivă diferită dar oarecum similară, *gramatica funcțională* (FG) [25] a lui Simon Dik, orientată funcțional şi semantic, încearcă să facă aceleaşi lucruri. Ca şi în SCD, FG găseşte *patru tipuri ierarhice* de bază ale categoriilor relaționale, aceste tipuri corespunzând într-o bună măsură cu *clasele de marcheri* PS-Ms şi *ierarhiile* lor stabilite în SCD [7], [2], [6]. PS-Ms reprezintă acele mijloace lingvistice de "suprafață" pe care le utilizează un limbaj natural pentru a organiza sintactic şi semantic structurile codificate în construcții gramaticale. Se impune în viitor o analiză comparativă între cele *patru nivele* sau *"straturi'* din organizarea formală şi semantică furnizată de FG [25], şi cele *patru nivele* de proiecție lingvistică, împreună cu clasele de marcheri corespunzătoare, din SCD: (1) *cuvântul* (lexical); (2) *sintagma* XG (X = N, V, A) subclauzală; (3) *clauza* (finită şi infinită); (4) *unitatea de discurs* (una sau mai multe fraze, care să formeze un *segment* de discurs).

(e) în fine, privitor la utilizarea intensivă a *caracterului predicațional* pe care categoriile lexicale majore (N, V, A) îl poartă (proprietate moştenită sau dobândită apoi de alte categorii gramaticale), *strategia lingvistică* SCD este comparabilă în special cu FG, cu accentul particular pe ierarhiile de delimitare şi marcare aplicate structurilor sintactico-semantice. SCD porneşte de la *lexicon* şi stabileşte la acest nivel o *taxonomie predicațională inițială* pentru categoriile lexicale majore. Un exemplu simplu al acestei taxonomii predicative este dat de către cele două categorii importante de substantive comune: *substantive existențiale* sau *obiectuale,* a căror *predicaționalitate (trăsătură* PRED) este EXIST *(e.g.* [Eng: *student, table;* Rom: *elev-student, masă])* şi a căror reprezentare funcțională reflectă categorii individuale sau personale, de exemplu predicatul uni-variabil *student(X), masă(X)* etc, şi substantive de tip-predicațional, a căror predicaționalitate *(trăsătură* PRED) are valoarea ACT, *e.g.* [Rom: *întâlnire, invidie, marcare* etc], şi ale căror reprezentări funcționale depind de mai multe variabile, de exemplu *întâlnire(X,* Y,...), *invidie(X,* Y,...), *marcare(X,* Y) etc Substantivele proprii şi/sau personificările sunt codificate prin constante ale variabilelor din predicatele de mai sus. Câteva din remarcile anterioare vor fi aprofundate în concluziile finale ale lucrării.

Schemele FX-bar, ca şi precursoarele lor *schemele AX-bar* [19], reflectă Pentru SCD faptul că un XPG (grupul sintagmatic de nucleu X), sau mai simplu XG, conține un *nucleu,* reprezentat printr-o categorie lexicală (nevidă) sau printr-o categorie virtuală (vidă), înconjurat (prin relații de *coeziune)* de specificatori şi/sau Codificatori de tipul A (adjectival-adverbial). Este esențial să facem următoarea specificare: un XG din SCD nu include nici un complement (argument obligatoriu)

sau adjunct. Complementele şi adjuncţii, împreună cu *nucleele* de nivel BAR = 1 formează nivelul BAR = 2 în FX-bar schema propusă în Fig. 3.2.1. Pentru un anumit nivel de specificare semantică, FX-bar schemele nu fac o distincţie clară între complemente (argumente obligatorii) şi adjuncţi, considerând toate structurile subcategorizate ca fiind argumente *sintactice]* clasificări ulterioare (suplimentare) sunt făcute pe baza şabloanelor verbale şi *restricţiilor* sintactice, semantice, şi pragmatice asupra componentelor şablonului, la nivel de lexicon.

O problemă a cărei soluţie poate influenţa în mod special şi teoria X-bar este aceea a asignării corecte a complementelor şi adjuncţilor, în particular, a stabilirii corecte a dependenţelor dintre grupurile nominale (NGs). Soluţia acestei probleme nu se poate obţine la nivel sintactic, iar o soluţie completă nu se poate obţine uneori nici chiar în contextul unui nivel semantic minimal (vezi [26], [27]). Chomsky remarcă realitatea că "... *the distinction between modifiers and arguments is notoriously difficult in certain cases"* [9, p. 44]. Exemple simple ilustrează această problemă: în TBarr [8], sintagmele *"the students ofphysics"* este văzută ca un NP cu un *argument* PP, în timp ce sintagma *"the students in. the yard"* este considerată a fi un NP cu un *adjunct modificator* PP. De fapt, în numeroase LNs, inclusiv engleză, se pot aduce multiple argumente serioase pentru ca cele două sintagme să poată fi la fel de bine interpretate fie într-un fel, fie în celălalt.

Soluţia SCD pentru acest exemplu foarte particular este următoarea (schiţând şi soluţia problemei generale): substantivul *"students"* este obiectual, adică *nu are* o natură *predicaţională* prin el însuşi, astfel că ambele sintagme nominale care îl succed sunt considerate de către SCD ca fiind *modificatori* pentru NG *"students".* Natura acestor modificatori poate fi diferită deoarece *"physics"* este introdus de marcherul de *caz* (genitiv) "of, în timp ce *"the yard"* este introdus de marcherul *prepoziţional "in".* în general, când nucleul lui NG posedă o trăsătură *predicaţională,* atunci NG care urmează nucleului predicaţional asigură o distribuţie sintactică ce satisface un anumit şablon (verbal) al predicatului (verbului) corespunzător.

Clasele din PS-Ms şi ierarhiile lor din SCD [7], [45] sunt responsabile pentru delimitarea structurilor sintagmatice propuse de schemele FX-bar, şi pentru stabilirea dependenţelor sintactico-semantice. Diferitele tipuri de marcheri sunt adesea aplicate simultan (deci multiplu) asupra aceloraşi categorii gramaticale, în cadrul anumitor nivele de structurare (proiecţii pe BAR-nivel). Similar cu unele teorii lingvistice (LFG, FG, şi parţial HPSG) dar contrar altora (GB, GPSCB etc), SCD nu consideră *prepoziţia* (X = P) ca fiind o categorie lexicală majoră. în SCD, P primeşte rolul unui marcher (funcţional), având atât proprietăţi de marcher de caz cât şi de complementizator. Categoriile HPSG PP[+PRD] sau PP[-PRD] (vezi [16]) sunt irelevante pentru SCD deoarece trăsătura +PRD în HPSG este atribuită numai lui PP subcategorizat de un V, în timp ce trăsătura *(predicaţională)* PRED din SCD poate fi în mod egal atribuită lui V, N, sau A (la nivelul lexiconului, cel puţin) dar nu şi lui P.

în SCD proprietăţile de subcategorizare sunt exploatate *ab initio,* la nivelul de organizare a lexiconului, pe baza *trăsăturii funcţionale* PRED de *predicaţionalitate,* asignată sau nu, unora din categoriile sintactice majore N, V, A. Observaţii lingvistice

Empirice ne-au convins, încă de la începuturile cristalizării SCD [22], funcţională şi predicativă adecvată ar trebui să reprezinte punctul oricărei teorii lingvistice, atât din motive teoretice cât şi computaţionale din abordările actuale (cum ar fi [27]-[32]) aduc o susţinere puternică p ideile esenţiale din SCD, în special folosirea intensivă a pr ^funcţionalităţii descrierilor lexical-semantice ale categoriilor lingv ^*procesarea automată* a LN cât şi în organizarea *bazelor de cunoştinţe*

[19] propune următoarea specificare a *Principiului Proie* K(PMP) [Eng: *Principie of Maximal Projection],* ca un pas important intensivă a trăsăturilor predicaţionale (funcţionale) ale categoriilor l îh SCD. Propunem aici

O specificare a P M P (formă actualizată):

Proprietăţile de subcategorizare ale categoriilor sintactice ma **depind de trăsătura lor lexical-semantică PRED**(icativity), **cu val** EXIST, **şi de trăsătura lor morfo-semantică** TENS(e), cu **valoril** **INFI**(nite).

Trăsătura PRED, atribuită categoriilor majore N, V, A la ni primeşte două valori: valoarea ACT, pentru acele categorii car r *predicaţionale* (în literatură este folosit adesea termenul *"deverbale* ; EXIST, pentru acele categorii N, V, A cu caracter *existenţial,* predicaţional. Trăsătura TENS primeşte valorile FINI(te) pentru a „^categoriei V care posedă un timp sau aspect finit, personal, şi val pentru toate celelalte categorii şi situaţii. Exemple:

[Eng: *boy,.pencil*\ Rom: *băiat, pix]* PRED:= EXIST; şi TE

[Eng: *attempt, showing, proved;* Rom: *încercare, arătând, demonstr*

PRED:=ACT;şi TE

[Eng: are; Rom: *sunt]* PRED:= EXIST; şi TE

[Eng: *gives;* Rom: *dă]* PRED:= ACT; şi TE

într-un grup verbal VG reprezentând un compus la un ti "pozitive" de trăsături, cum sunt ACT sau FINI sunt *moştenite* de la de către întreaga sintagmă VG, sau pot fi obţinute *cumulativ* prin sintactică.

Specificarea PMP de mai a funcţiei *proiecţiei maximale* e SCD deoarece îh multe LNs, inclusiv în română, *calitatea dev* ţională, deci funcţională) a categoriilor lexicale tradiţional *non-verba* A trebuie descoperită cât mai devreme posibil şi asignată la nivel ^xemplu, în engleză, deşi pentru substantivele care *'verbali* valoarea trăsăturii lor TENS este INFI, aceste substantive posedă, PRED, aceeaşi valoare ACT sau EXIST pe care o au verbele

substantivele (sau gerunziile) în "-/ng", şi astfel posedă *aceleaşi* proprietăți de subcategorizare ca ale verbului de origine.

3.2. Ipoteze de lucru şi aspecte caracteristice ale FX-bar schemei

Continuând ideile de bază ale schemelor AX-bar din [19], propunem, pentru SCD, FX-bar schema generală din Fig. 3.2.1. Muchiile din stânga conțin *noduri* cu rol *funcțional* sau *relațional:* marcheri, cuantificatori, specificatori, modificatori (eventual adjuncți). Pentru a obține reprezentări sintactice şi semantice corecte, nodurile funcționale se aplică (recursiv) *nucleelor* Xk şi CLk, k = 0, 1, 2, iar nucleele, cu rol funcțional (predicațional, X1) sau relațional (eventual X2), au ca argumente clauze infinite (complemente, X1) sau finite (X2). Precizăm că la acest nivel nu se poate face distincția dintre complemente COMPLi (argumente obligatorii) şi adjuncți ADJCTj (argumente opționale). în mod normal, în Fig. 3.2.1., ADJCTj sunt "amestecați" printre ARGj ,'la nivel sintactic nefiind discernabili de complementele obligatorii ale unui nucleu predicațional. Poziția funcțională (la stânga nodului X1) a nodurilor ADJCT poate rezulta doar în urma unor calcule semantice şi pragmatice suplimentare, din care se obține rolul tematic al argumentelor ARGs ale lui X1.

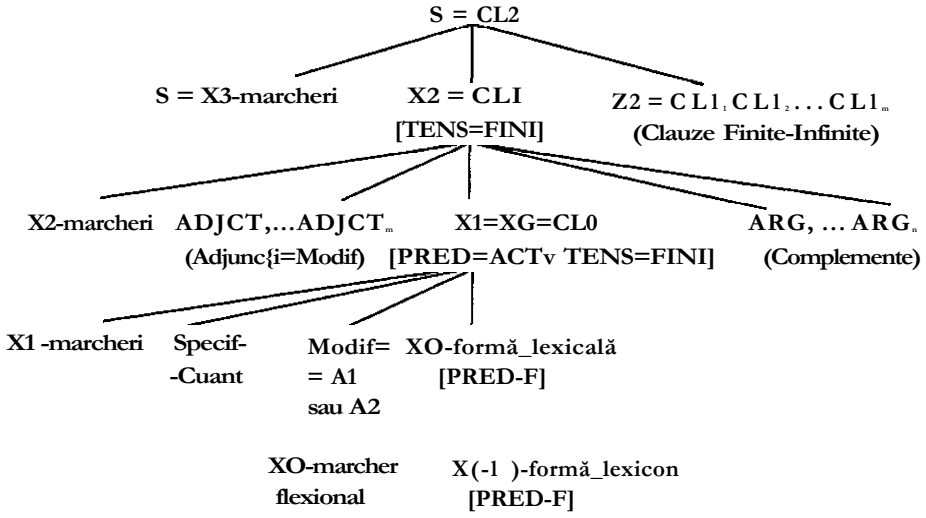


Figura 3.**2.1**. Schema (funcțională) FX-bar generală

(*) Aspecte specifice ale schemei FX-bar propuse: (*1) Sunt permise un număr arbitrar de *argumente* (sau *sateliți In* sensul [10], [31]), toate notate cu ARGs. In SCD, ARGs sunt formate din *complemente* obligatorii (COMPLs) şi din *adjuncți* (ADJCTs), sau *complemente opționale.* ADJCTs pot fi reprezentați la nivel sintactic tot ca argumente ale nucleului, însă la nivel semantic ADJCTs au rol de *modificatori* ai nucleului. Notația "A-poziție" din teoriile Chomskyene, care înseamnă ARG-poziție, nu trebuie confundată cu notația noastră pentru categoria A = adjectiv-adverb. în teoriile şi notația lui Chomsky, COMPLs sunt în A-poziție (ARG-poziție), în timp ce ADJCTs nu. SCD se situează pe o poziție sintactică similară cu HPSG [16], care utilizează lista SUBCAT pentru a codifica toate sintagmele pe care le subcategorizează un nucleu semantic, adică atât COMPLs cât şi ADJCTs (sau ARGs din SCD). (*2) Sintagmele AG = A0 sau A1, sau AP = A2 sunt *postulate* de către SCD ca fiind de tipul categoriei *funcționale* Modif, manifestate prin categoriile A (de nivel XO, şi aplicabile la nivel XO), ADJCTs (de nivel X1, şi aplicabile la nivel XO şi X1), şi clauza relativă (de nivel X2, şi aplicabilă la nivel XO şi X1). (*3) Categoria generică Specif (sau Spec), în care intră cuvintele şi sintagmele ce desemnează *cuantificatori* de toate tipurile (generalizați), determinatori (în particular), este postulată de către SCD ca fiind o *categorie funcțională* ce poartă trăsături de natură *cuantificațională* la nivel lexical (în particular, *negația* la nivelul X1), inclusiv articularea (hotărâtă sau nu), suprapunându-se deci uneori peste X1-marcheri de trăsături funcționale cum este *acordul.* Relațiile (funcționale) de *acord* sunt esențiale pentru *coeziunea locală* şi *globală* în cadrul strategiei SCD: acordul dintre XO-Modif şi XO-Specif cu nucleul XO (la nivel X1), acordul Nucleu-Subj (sau chiar Nucleu-COMPL) şi acordul COMPL-Pron$_{Em}$**fat** (Pronume emfatic) (la nivel X2), o anumită *corespondență* a timpurilor evenimentelor într-o clauză şi între clauze. Aceste tipuri de relații de acord, referință şi coreferință, coeziune, coerență, etc. sunt responsabile pentru o largă clasă de dependențe locale şi globale, inclusiv dependențe la distanță mare şi în extra-poziție. Accentul în componenta de *coeziune* a strategiei SCD (Segmentare-Coeziune-Dependență) cade pe mijloacele *sintactice* şi de *"suprafață",* mai curând decât pe cele semantice, încercând să găsim, să extragem, şi să utilizăm într-o măsură maximală informații de ordin superior, cum ar fi informația de discurs [34], pragmatică, semantică etc. (*4) Sintagma tradițională PP din teoriile lingvistice clasice, iar în SCD, grupul prepozițional PG (format dintr-un grup nominal NG care este precedat de o prepoziție sau o locuțiune prepozițională) este întotdeauna considerată un ARG (COMPL sau ADJCT) în FX-bar schemele al căror nucleu (lexical nevid sau vid) este N, V, A. Această ipoteză de bază asupra PG este justificată de SCD prin faptul că P *nu* este considerată o categorie majoră, adică o categorie de nivel X1 în schema FX-bar din Fig. 3.2.1. ci doar o categorie de nivel XO. Proprietățile de subcategorizare ale N, V, A (dar nu şi P) pot fi asignate *ab initio,* 'a nivel de lexicon, începând cu trăsătura lexicală PRED a categoriilor predicaționale. Categoria P poate primi proprietăți funcționale, cel mai adesea ca *marcher de caz,* uneori proprietăți *relaționale* (de exemplu [Eng: *on;* Rom: *asupra]),* dar nu şi Proprietăți de subcategorizare. (*5) Subiectul (Subj) în SCD, lexical nevid sau vid (°RO), este considerat ca un argument special al proiecțiilor maximale ale categoriilor X = N, V, A într-o clauză finită (de nivel X2) sau infinită (de nivel X1). (*6) In ipotezele (*5) şi (*2) de mai sus, categoria lingvistică tradițională VP este dizolvată într-un grup verbal VG (finit sau infinit), înconjurat (de cele mai multe ori

urmat) ca nucleu de ARGs şi formând o clauză finită, respectiv infinită. (*7) *Teoria limitării* şi multe probleme majore legate de TBarr [8], [9], [17] sunt explicitate şi rezolvate în cadrul realizat de SCD şi schemele FX-bar, în principal datorită delimitării clare a funcţiilor şi relaţiilor care se aplică cuvintelor şi sintagmelor, a reprezentării lor lexicale prin clasele de PS-Ms, şi a specificării domeniului lor de aplicare. Acest rol este realizat explicit în cadrul claselor si ierarhiilor de marcheri propuse şi utilizate de SCD [2], [6], [7], [45]. Trebuie să remarcăm că în lucrările sale cele mai recente [34], [35], Chomsky adoptă o tehnică similară de "limitare" a operaţiilor de *construire* [Eng: merge] şi *transformare* [Eng: *move]* doar la "domeniul" sintactic al unei *"faze"* [Eng: *phase],* o unitate textuală (care în general coincide cu clauza!) în care Chomsky propune următorul *principiu de impenetrabilitate "într-o fază* (clauză *n.n.)* F *cu nucleul* H, *domeniul lui* H *nu este accesibil la operaţii în exteriorul lui* F, *ci este accesibil numai* H şi *muchia sa* (nodul său ascendent)" [34]. Exact aşa este construită şi funcţionează schema FX-bar! De asemenea, fenomene de *teoria legării* [9], [8], [3], [16], *legăturile* [Eng: *linking]* din [27], mecanisme de *coeziune* (locală şi globală) şi *discurs* întâlnite în [36], [31], [33], etc. sunt mai uşor de pus în evidenţă şi de rezolvat în cadrul oferit de strategia lingvistică SCD si teoria FX-bar.

(•) Observaţii asupra ipotezelor de lucru pentru schema FX-bar din Fig. 3.2.1.: (41) Schema FX-bar este proiectată să lucreze în asociere cu un parser care este capabil să recunoască clasele de PS-Ms şi structurile sintagmatice considerate de strategia lingvistică SCD. Schema FX-bar este organizată pe *patru nivele* de proiecţie BAR = (H-3 (deasupra nivelului de lexicon, notat convenţional BAR = -1); *trei nivele* **X0-X1-X2** corespund proiecţiei dintre nivelul lexical (BAR = 0) si nivelul *clauzal,* al structurilor *uni-eveniment,* alte *trei nivele* **CL0**(=X1**)-CL1**(=X2**)-CL2** corespund proiecţiei dintre nivelul *clauzal minimal* CLO = X1 şi nivelul *frazei,* al structurilor *multi-eveniment.* Nivelele uni-eveniment X0-X1-X2 exprimă predicaţia clauzei (propoziţiei) simple în care sunt distribuite categoriile lexicale de bază şi sintagmele pe care le generează, în timp ce nivelele CL0-CL1-CL2 exprimă relaţiile logice şi predicaţionale (de ordinul doi) dintre clauzele simple. Schema FX-bar lucrează într-o manieră recursivă (top-down sau bottom-up), atât în situaţiile de analiză cât şi în cele de generare în care este antrenat parserul asociat, în strânsă cooperare cu strategia lingvistică SCD, cu clasele de PS-Ms şi ierarhiile lor si, mai ales, pe baza *meta-algoritmilor* SCD de analiză-generare [1],' [2], [6], [7]. Să mai observăm că FX-bar schema din Fig. 3.2.1. poate fi utilizată independent de aşa numita *ordine canonică* (sau *sistemică)* a cuvintelor si sintagmelor dintr-o clauză, specifică fiecărui LN [37], [38]. (^2) Valoarea ACT de trăsătură (funcţională) pentru categoriile N şi A (şi implicit V) este atribuită acestor categorii la nivel de lexicon atunci când ele corespund unor evenimente cu actanţi şi/sau stări multiple. Valoarea EXIST este implicit sau explicit introdusă de formele şi înţelesurile verbelor existenţiale (a *fi),* modale (a *trebui),* etc. (^3) Trăsătura (funcţională) TENS este similară cu categoriile virtuale I (INFL) şi T (Tense) din teoriile GB si TBarr ale lui Chomsky şi din schemele S-bar corespunzătoare (Fig. 1.3. şi Fig. 2.2.3.). Pentru un VG finit (TENS = FINI), structura V2 corespunzătoare devine clauza finită clasică.

Dacă sintagma XG (X1) este un grup a cărei categorie-nucleu X posedă valorile de trăsături PRED = ACT şi TENS = INFI, atunci XG devine noul nucleu al unei clauze infinite ce face parte dintr-o structură de nivel X2 (XP). (^4) Poziţia specială a *subiectului sintactic* (Subj) este considerată de către SCD atât o ARG-poziţie (asemănătoare, de fapt, cu o COMPL-poziţie) cât şi o Caz-poziţie. în concordanţă cu TBarr [8] şi cu HPSG [16], Subj primeşte poziţia specială a *primului element* din lista SUBCAT [16]. Aceasta este în esenţă o poziţie sintactică, iar Subj poate primi o funcţie tematică (Opoziţie) autentică doar ca rezultat al unor calcule sintactice şi semantice suplimentare! (^5) Aşa cum rezultă din schema FX-bar din Fig. 3.2.1., sintagmele AP şi PP din teoriile lingvistice clasice sunt segmentate de către marcherii SCD [7], [45] în sintagme mai mici XG, X = N, V, A. Aşa cum am precizat deja, SCD atribuie noilor sintagme următoarele roluri: AG = Modif, cu rol funcţional la nivelul de proiecţie X1, şi PG = ARG (COMPL sau ADJCT), ADJCT purtând de asemenea rol de Modif al nucleului de nivel X2. PG devine deci un NG P-marcat, iar orice categorie A are de la început reprezentarea (nesaturată) A(X), unde X = N, V, A este nucleul ([existent, viitor, sau lipsind pur şi simplu) al sintagmei de nivel X1 în care Modif = A. în mod similar, orice categorie Specif (determinator, cuantificator, etc.) joacă un rol similar, schema FX-bar impunând reprezentarea funcţională Specif(X), unde X este nucleul sintagmei. (^6) în ciuda anumitor asemănări (inerente) între schemele FX-bar şi versiunea MinP a teoriei X-bar, există diferenţe de bază în ce priveşte organizarea şi funcţionarea constructivă dintre schemele (F)X-bar din Fig. 3.2.1. şi Fig. 2.2.1. De exemplu, în schema FX-bar, fiecare element lexical se proiectează într-o categorie obiectuală sau funcţională (relaţională), aceasta este (coeziv şi ^recursiv) înconjurată de către Specif şi/sau Modif, iar dacă valoarea ACT a trăsăturii HPRED a nucleului este prezentă, atunci această valoare ACT este moştenită de către întreaga sintagmă al cărei nucleu a fost specificat sau modificat. Această sintagmă cu nucleu predicaţional îşi subcategorizează complementele (argumentele obligatorii COMPLs) şi adjuncţii ADJCTs (care modifică sintagma-nucleu). în schema X-bar din Fig. 2.2.1.,'se întâmplă tocmai invers deoarece *"The Head-Complement relation is the "most local" relation of an XP to a terminal Head* Y, *all other relations within YP being Head-Specifier (apari from adjunction, ...)"* [3: p. 53]. (^7) Deşi schema FX-bar generală a fost proiectată având în vedere în primul rând limba română, ea poate fi aplicată pentru a reprezenta, grafic şi logic, structuri sintactico-semantice ale LNs cu valori ale parametrilor gramaticali foarte diferite, cum ar fi engleză-germană sau franceză-germană. Distribuţia complementelor (argumentelor) în română (engleză, franceză) poate fi foarte diferită de cea din germană; de exemplu, într-o clauză al cărei verb principal din compusul său verbal VG se află în poziţie finală, sau pentru o categorie A (adjectiv-adverb) având valoarea de trăsătură PRED = ACT.

Ex. 3.2.2.R. */Paharul /spart //de Ion/ cu mingea /de fotbal/*

Ex. 3.2.2.E. */The glass /broken //by Ion/ with / the football /*

Ex. 3.2.2.G. */Das/von Ion/mit/dem Fuflball //zerbrochene //Glass/*

După cum am remarcat în (t1), schema FX-bar poate fi utilizată [incJependent de regulile structurilor sintagmatice şi ordinea lor (din română sau

germană), aceasta deoarece principiile rămân aceleaşi şi diferă numai anumiți parametri şi valorile lor pentru LNs distincte: în română (şi engleză, franceză) argumentele succed o categorie A ce reprezintă un nucleu predicațional, în timp de în germană ele îl pot precede. Dacă un nucleu V al unei clauze are valorile de trăsături PRED = ACT şi TENS = FINI, atunci distribuția ARGs este similară cu cea din română, cu posibile (şi probabile) diferențe impuse de *ordinea sistemică,* strict dependentă de LN, a ARGs (a se vedea [37} dar şi [27]).

Dacă se încearcă utilizarea formei FX-bar ca "schelet" pentru un automat (sau gramatică formală) de analiză şi generare a LN, un asemenea automat ar trebui să mimeze atât forma generală a schemei FX-bar cât şi regulile gramaticale de analiză-generare. Partea din automat care reflectă cele *patru* nivele de organizare a structurilor LN în schema FX-bar ar trebui să fie independentă de limbaj (cel puțin pentru o largă clasă de limbaje europene), în timp ce (sub)partea constituentă care recunoaşte structurile lingvistice pe fiecare nivel individual Xk (k = 1, 2, 3) trebuie să fie dependentă de limbaj (acest fapt este binecunoscut şi parametrizat). Reprezentarea schemei FX-bar pentru Ex.3.2.2.G. este aceeaşi cu reprezentările FX-bar pentru Ex.3.2.2.R.-E., şi similară cu figura pentru Ex.4.1.2.R.-E.

4. Exemple de aplicare a schemelor FX-bar

Vom expune câteva exemple de aplicare^ a schemelor FX-bar la reprezentarea sintagmelor, clauzelor şi frazelor. în exemplele prezentate, categoriile gramaticale pentru care PRED = ACT sau TENS = FINI vor fi *subliniate,* iar PS-Ms care se aplică sintagmelor Xk (k = 0, 1, 2) sunt reprezentați grafic în text prin apariția unuia sau mai multor semne 'slash' /. Să notăm că schemele (augmentate) AX-bar din [19], deşi oarecum asemănătoare în spirit sunt efectiv scufundate în schema FX-bar generală, diferențele substanțiale constând în forma unitară a FX-bar schemei şi în criteriile sintactice şi logico-semantice mai clare, pe baza cărora clasele de PS-Ms şi ierarhiile lor sunt explicit propuse şi aplicate în funcționarea schemei FX-bar.

Care este relația dintre exemplele de FX-bar scheme şi formulele logice ataşate după reprezentarea grafică? *Prima formulă* este o reprezentare uzuală a LN, care foloseşte limbajul logicii predicatelor, reprezentare mai apropiată de exprimarea în LN, conținând toate variabilele ce codifică referințele-coreferințele, dar (pentru simplitate) fără cuantifcatorii corespunzători. *A doua formulă* este traducerea mai completă a primei formule în limbajul de programare logică Prolog, folosind tehnici clasice de reprezentare a cunoştințelor de LN în Prolog. Pe o scală ascendentă a măsurii în care sintagmele LN ar fi analizate, *schema* FX-bar poate fi văzută ca un prim rezultat al procesului de parsare (analiză), *prima formulă* ar urma procesului de parsare, incorporând fenomenele de referință (şi coreferință, rezoluție a anaforei, etc), iar a *doua formulă* ar reprezenta o rafinare a primei formule. Formulele de tipul doi reprezintă de asemenea atât un stadiu final al procesului de analiză a frazei cât şi punctul de pornire în procesul de generare a

une i fraze (conform cu abordarea [39], [6] a generării automate a LN, de [31], de exemplu).

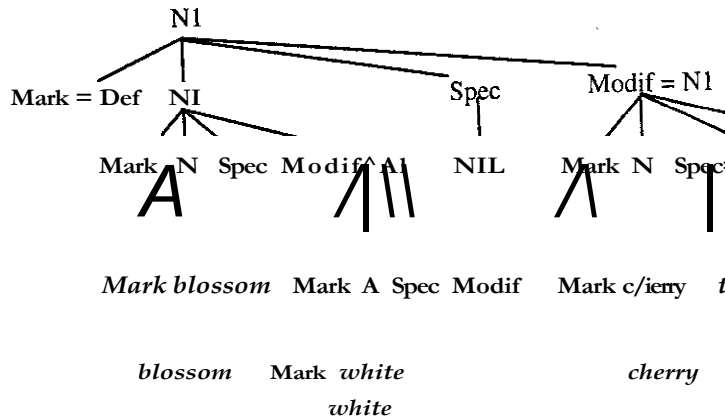
Este important să remarcăm că schema FX-bar propusă principal, relațiile de *dependență* dintre diferitele categorii, sintagme dintr-o frază, împreună cu marcherii corespunzători care controlează, comportamentul lor distribuțional. Deoarece am văzut în ce măs argumentelor este (parametric) dependentă de limbaj în schemele FX-b pot codifica nu numai situații în care argumentele succed (situația obiş care ele preced nucleul lor semantic (Ex.3.2.2.), dar şi în care aceluiaşi nucleu sunt interschimbabile. Deci aceleiaşi scheme FX- atribui mai multe formule logice corespunzătoare "echivalente".

4.1. De la text la scheme FX-bar

, \ Strategia SCD propune următoarele scheme FX-bar pentru mai jos. Deşi muchiile ale căror noduri sunt Modif sau Specif su dreapta nucleului corespunzător (pentru convenience grafice), ele tre ca având rol funcțional (situate la stânga şi aplicându-se nucleului X1 unii adjuncți, la nivel X2. Diferențele dintre codificarea formei pentru e pentru română sunt nesemnificative (cu excepția unor aspecte sup acord, care sunt puse în evidență). Forma codificată a textului engleză este un argument suplimentar pentru versatilitatea scher propuse.

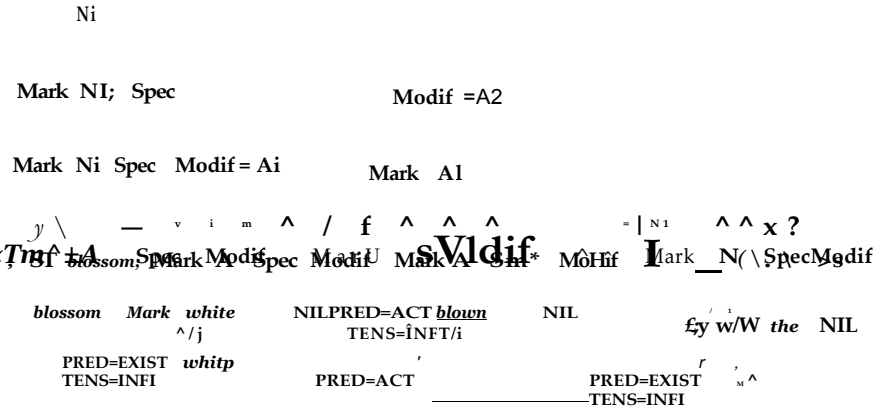
~Ex. 4.1.1.R. */floare albă/de cireş /*

Ex. 4.1.1 .E. */ the cherry / white blossom /*



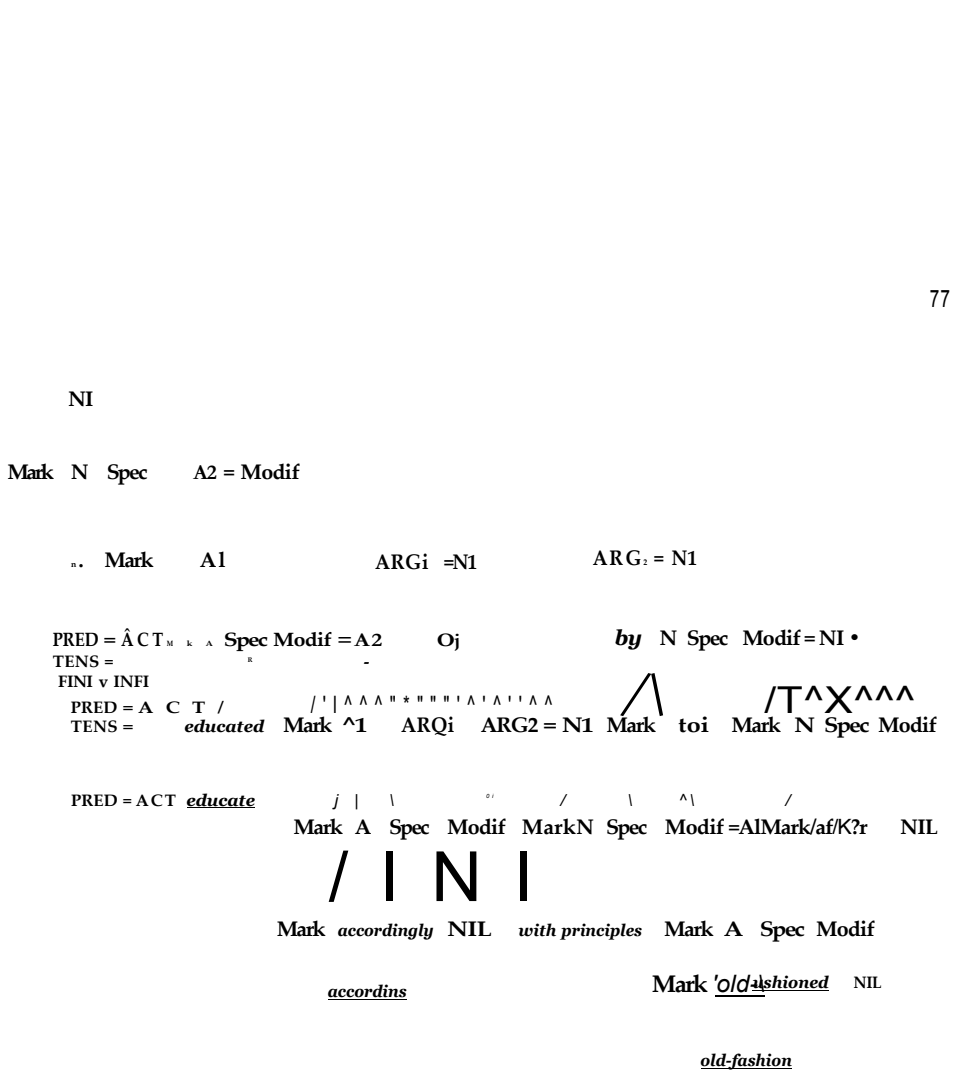
LR.4.1.1.R. de(cireş)(albă(floare(X)));
LR.4.1.1.E. quant(indef, X, white(blossom(X)), cherry(X)).

Ex. 4.1.2.R. */floare albă/ //bătută//de vânt/*
Ex. 4.1.2.E./tf7e *white blossom/ //blown//bv the wind/*
object, = O,; eventj = e,



LR.4.1.2.R. albă(floare(X)) A bătută(de(vânt(Y)), X);
LR.4.1.2.E. quant(indef, X, white(blossom(X)),
quant(indef, Y, by(the(wind(Y))), blown(Y, X))).

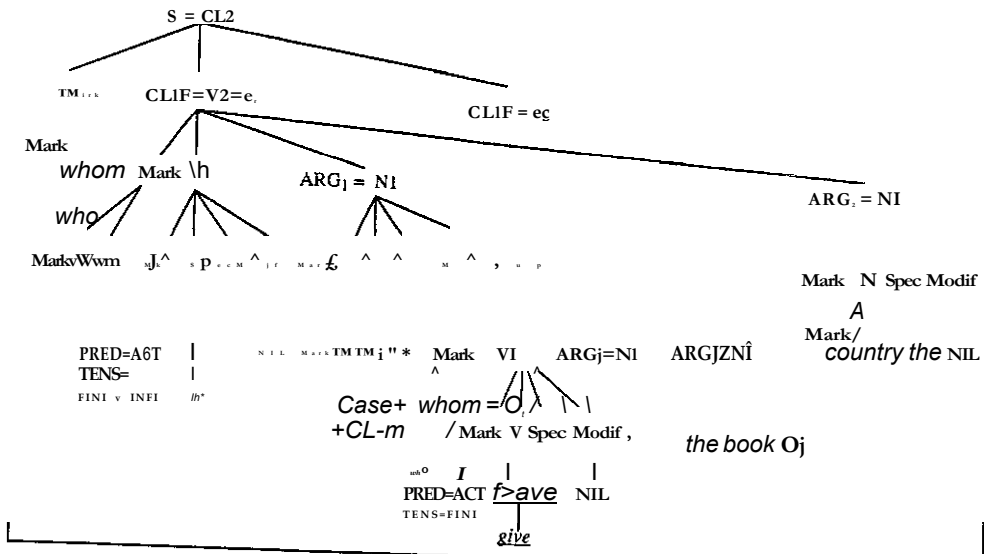
Ex. 4.1.3.R. *//educat//[de tatăl său//corespunzător//cu vechile principii/*
Ex. 4.1.3.E. *II educated///[bv his father//accordinalv// with old-fashioned principles/*



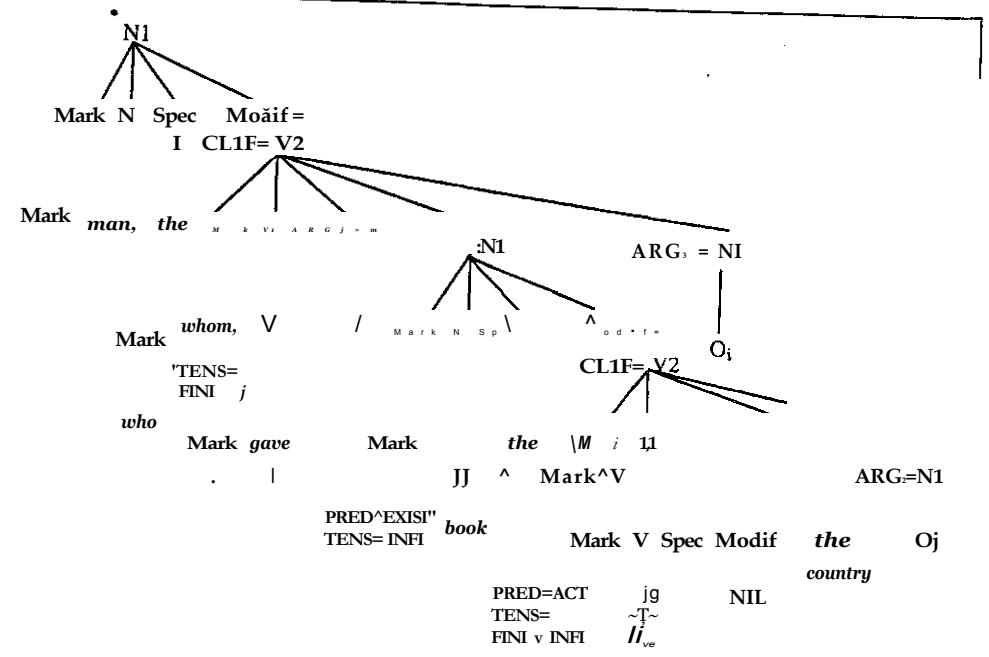
LF.4.1.3.R. corespunzător X, cu(vechile(principii(Y))))
(educaţ(X, de(său(tatăl(Z)))));
LF.4.1.3.E. quant(indef, X, educated(X, by(his(father(Z)))), quant(indef, Y, with(old(principles(Y))), accordinqlv(X, Y))).

Ex.4.1.4.R. *II Omul, // căruia, // PRO, / h* -am *dat // cartea //* PROj *a părăsit // ţara.//*
Ex.4.1.4.E. *II The man\ //whomVl\ /pave//the book//PRO\ left //the country.//*

4.1.4.E. Readinq 1 *(left* **= pastjense(feave))**



4.1.4.E. Reading2 (*left* = past_participle(*leave*))

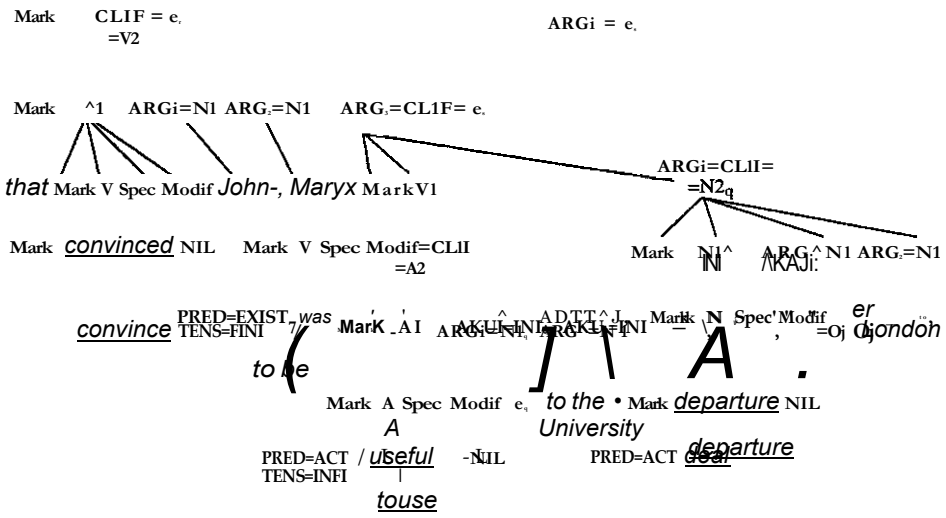


4.1.4.E. Reading3 ([Eng: *leff*\ - [Rom: *stânga]).*.

LF.4.1.4.R. a-părăsit(omul(X) A am-dat(Y, cartea(Z), X), ţara(T);
LF.4.1.4.E. quant(def, X, and(man(X), quant(def, Y, I(Y), quant(def, Z, book(Z), qavef Y, Z, X)))), quant(def, T, country(T), jeft(X, T))).

Ex. 4.1.5.R. *//lon\ //a convins^ //pe Mariaj //că //deplasarea* /\e/j /la Lodra / /a fost utilă //e_k / Universității. //*
Ex. 4.1.5.E. *// John\//convinced I/Mary^ //that //her^ //departure* // to London //was useful//e_k / to the University. //*

objectj = Oj; eventj = ej



LF.4.1.5.R. a-convins(ion, -o(pe(mariaj)), că(a-fost-utilă(deplasarea(ei(Xj), la(londra)), universității(Y)))));
LF.4.1.5.E. convinced(john = X, mary = Y, quant(def, X, her(X), departuref X, to(london)) = E, quant(def, Z, university(Z), was-useful(E, Z)))).

4.2. Observaţii generale

(*1) NU este scopul prezentei lucrări să arate cum sunt obţinute ^prezentările FX-bar ale structurilor LN (într-o manieră mai mult sau mai puţin 3'goritmică), ci doar să propună schema FX-bar generală ca un mecanism esenţial ^reprezentare a informaţiei lingvistice, să sugereze cum lucrează, şi să explice

raţiunile introducerii acestui mecanism. Teoria FX-bar este integrată ca[4] o componentă importantă a strategiei lingvistice SCD, însă ea poate fi utilizată şi în alte contexte computaţionale, cu condiţia de a include ingredientele necesare, şi anume, clasele de PS-Ms, ierarhiile acestor clase, o taxonomie funcţională (predicaţională) şi relaţională a categoriilor majore şi a marcherilor, un algoritm (în particular, algoritmii SCD) de obţinere a structurilor de dependenţă, etc. Aspecte mai detaliate ale SCD au fost prezentate în [1], [2], [6], [7], [45]. (*2) Funcţionarea corectă a schemelor FX-bar expuse arată clar cât de necesară este utilizarea (intensivă) a trăsăturilor *predicative* şi *funcţional-relafionale* pentru fiecare categorie lexicală. Din experienţa noastră în ce priveşte analiza şi generarea automată a limbii române [6], considerăm că accentul pus pe trăsăturile funcţionale ale categoriilor gramaticale, cuplat cu punerea în evidenţă a PS-Ms, reprezintă elemente-cheie în utilizarea cu succes a teoriilor X-bar curente în procesarea automată a LN şi în cadrul unor teorii lingvistice moderne (UG, FG, HPSG, etc). (A3) Punerea în valoare a trăsăturilor *funcţionale* (în particular, predicaţionale) ale categoriilor majore N, V, A, şi a celor *relaţionale ale* claselor *de marcheri* (marcheri numiţi în literatură şi *"cue phrases"* [Rom: *sintagme indicatoare]* [28], [31], sau *conective* [29], [30] etc), deşi esenţiale, nu poate rezolva toate problemele. De exemplu, asignarea dependenţelor corecte în juxtapunerea de NGs este o problemă binecunoscut de dificilă, imposibil de rezolvat complet doar la nivel sintactic. Există însă în prezent un puternic curent către acest tip de abordări, aceasta deoarece ele reflectă mult mai adecvat structura reală a textului de LN (cel puţin pentru o clasă largă de LNs europene). Aceste abordări pot diferi substanţial în instrumentele şi tehnicile de parsare, însă principiile rămân foarte similare (de exemplu, [19], [29], [31], [33], etc (*4) PS-Ms (marcherii de structuri sintagmatice) joacă un rol fundamental în delimitarea structurilor sintactice şi semantice, şi stabilirea dependenţelor corecte între aceste structuri, SCD a pus accentul încă de la începuturi pe acest aspect [22]. Se remarcă în prezent o întreagă mişcare către reconsiderarea rolului esenţial al marcherilor, în special la nivel de *discurs* şi în analize complexe ale marilor unităţi textuale (regăsirea informaţiei, rezumare automată, planificare şi generare automată de text, etc). Strategia SCD, cu componenta ei de teorie FX-bar, încearcă să pună la lucru întreaga paletă de PS-Ms, de la nivel *lexical* şi de *coeziune* (locală), până la nivel de *discurs (coeziune* şi *coerenţă* globală), punând accentul pe sintaxă (nivelul de *"suprafaţă",* [Eng: *shallow])* şi pe un nivel minimal de semantism. în funcţie de problema de LN ce trebuie rezolvată, acest nivel poate fi amplificat în mod corespunzător, (A5) Cuplarea schemelor FX-bar cu: (a) clasele de marcheri SCD şi cu ierarhia lor ce corespunde celor *patru* nivele de proiecţie lingvistică din FX-bar [7], [45]; *(b) o* taxonomie bazată pe predicaţionalitate a categoriilor majore N, V, A; *(c)* exploatarea maximală a trăsăturilor *funcţionale* (predicaţionale) şi *relaţionale* ale tuturor categoriilor lexicale şi nelexicale (deci şi ale PS-Ms); *(d)* o schemă X-bar simplă şi unică, apelată recursiv pe cele patru nivele ale sale, pornind de la *lexicon* (convenţional, BAR = -1) şi până la nivelul de *discurs* al frazei multi-eveniment (BAR = 3), aceste aspecte reprezintă principalele diferenţe (şi noutăţi) dintre teoria FX-bar şi teoriile X-bar precedente, (A6) Schema FX-bar poate fi de asemenea

asociată cu un automat dependent de limbaj (pentru o largă clasă de LNs), care începe să lucreze pentru fiecare frază, primeşte *on-lirie* cuvânt cu cuvânt, şi se opreşte odată cu semnul de punctuaţie final al frazei. Pentru valori adecvate ale parametrilor de LN cum sunt *ordinea cuvintelor (argumentelor)* şi *direcţia proiecţiei lingvistice* pentru categoriile majore şi pentru marcheri, schema FX-bar poate reprezenta corect dependenţele structurilor lingvistice (inclusiv pentru Ex.3.2.2.G).

5. Problema X-bar teoriei actuale

Mai este necesară X-bar teoria sau nu? Este teoria X-bar pe moarte sau nu? Care este valoarea teoretică şi, mai ales, practică a (sub)teoriei X-bar în cadrul teoriilor lingvistice şi al tehnologiilor actuale ale LN? Cum trebuie să percepem în mod corect X-bar teoria atunci când, în aceeaşi carte a lui Chomsky, găsim următoarele două pasaje:

(Chomskyl): *"The concepts of X-bar theory are therefore fundamental. In a minimalist theory, the crucial properties and relations will be stated in the simple and elementary terms of X-bar theory." [3, p. 172],*

(Chomsky2): *"Standard X-bar theory is thus largely eliminated in favor of bare essentials." [3; p. 246].*

Subliniem că aceste citate nu sunt extrase din text astfel încât să nu aibă relevanţă în context, cu intenţia de a provoca confuzie. Dimpotrivă! De asemenea, scopul nostru nu este de a căuta o posibilă incoerenţă ci de a pune în evidenţă noua poziţie a lui Noam Chomsky, între 1992 şi 1995. încercăm să deschidem o discuţie pe această temă deoarece considerăm că există o problemă, şi că ea este de o reală importanţă.

în această secţiune urmărim cinci obiective: **(A)** Să enunţăm problema X-bar teoriei. **(B)** Să rezumăm soluţiile existente în momentul de faţă. **(C)** Să stabilim rolul X-bar teoriei în interiorul contextului teoriilor lingvistice şi să sugerăm posibile dezvoltări. **(D)** Să specificăm poziţia FX-bar schemelor propuse privitor la dilema eliminării complete a X-bar teoriei şi, în special, relaţia noii FX-bar teorii conturate în contextul strategiei lingvistice SCD. (E) Câteva concluzii şi perspective.

(A) Să considerăm următoarea *problemă:* reflectă *teoria X-bar* o *realitate lingvistică* a LNs, şi dacă da, prin ce *mijloace* această realitate lingvistică ar putea fi *cel mai bine* reflectată? Proiecţia categoriilor lingvistice este un fapt lingvistic de netăgăduit. Chomsky şi alţi distinşi lingvişti au fost în completă eroare în ultimii 25-30 de ani? Credem că nu. Problema este dacă teoria X-bar poate încă să mai fie un *bun model,* sau vehicul, care să exprime acest *fapt,* şi cu ce preţ de utilitate. *Principiul Proiecţiei Extinse* [3, p. 55] şi *Principiul Proiecţiei Maximale* (propus în [19] şi secţiunea 3.1.) au ca scop să stabilească forma şi marginile cele mai probabile ale unităţilor textuale obţinute în cadrul procesului de proiecţie a categoriilor lingvistice.

(B) Ipoteza (Chomskyl) de mai sus dă un răspuns afirmativ la această întrebare în timp ce (Chomsky2) reprezintă, aparent, opusul acestui răspuns. Abordarea din [3, Cap. *Categories and Transformations]* pentru ipoteza (Chomsky2) este că disoluția schemelor X-bar, deci a proiecției categoriilor lingvistice, poate fi înlocuită cu succes prin folosirea proprietăților de funcționalitate, predicativitate, tipologie şi transformare intrinseci acestor categorii, deşi aceste proprietăți sunt reprezentate în [3] cu acelaşi aparat X-bar pe care îl combat! în cadrul unei teorii a *"structurii sintagmatice pure",* operațiile unui sistem computațional al LN *"construiesc recursiv obiecte sintactice",* iar *"categoriile sunt construcții elementare rezultate din proprietățile elementelor lexicale",* cu condiția *"să nu fie adăugate obiecte noi în cursul procesării, înafară de rearanjări ale proprietăților lexicale"* [3]. Rezultatul pare să fie spectacular: dispar nivelele de proiecție (în sensul teoriei X-bar), astfel spus, nu se face nici o deosebire între elementele lexicale şi nucleele proiectate din ele, în timp ce *"teoria structurilor sintagmatice poate fi eliminată în întregime, se pare, pe baza celor mai elementare ipoteze"* [3, p. 294].

Nu ar fi pentru întâia oară când teoria lingvistică încearcă să renunțe la (sub)teoria X-bar. Chomsky sugerează că nivelele de proiecție lingvistică pot fi înlocuite de către *"proprietățile* (funcționale *n.n.) ale elementelor lexicale".* Acesta este chiar cazul *gramaticii funcționale* (FG) [25] în care, formal, lipseşte teoria X-bar. Dar chiar şi în gramatica funcțională a lui Dik, conținutul ascuns al teoriei X-bar este scufundat de fapt în cele *patru nivele* de structuri ierarhice ale *functorilor* şi *operatorilor* ce se aplică pe categoriile şi structurile cu care FG lucrează la fiecare nivel sintactic. O situație specială avem în SCD, unde nivelele de proiecție a categoriilor lingvistice sunt recuperate pe baza unei funcționalități ierarhice a elementelor lexicale, iar FX-bar schema propusă poate fi utilizată (recursiv) ca un invariant sintactic constructiv al structurilor sintagmatice în cadrul proceselor de analiză şi generare automată a LN (limbii române).

Schema FX-bar propusă (Fig. 3.2.1.) poate fi considerată ca un compromis, o negociere, între (Chomskyl) şi (Chomsky2), deoarece (Chomsky2) se prezintă fără mecanisme concrete pentru a-şi susține ipoteza: în timp ce teoriile X-bar clasice nu mai pot fi utilizate ca instrumente operaționale pentru a reflecta o viziune exclusiv funcțională (şi relațională) asupra sintaxei, teoria FX-bar propusă poate face acest lucru.

(C) Poziția noastră privind problema **(A)** asupra teoriei X-bar poate fi rezumată astfel: (C1) Proiecția categoriilor gramaticale este un fapt lingvistic. (C2) Acest fapt poate fi corect reflectat prin *"nuclee"* şi *"nivele* (bar) c/e *proiecție"* în interiorul schemelor X-bar, dar şi prin proprietățile funcționale *"intrinsec?* ale categoriilor lexicale şi gramaticale. (C3) Teoria X-bar include deci o componentă de adevărată construcție lingvistică, iar ingredientul său de bază este confecționat din relațiile funcționale stabilite între elementele lexicale (şi nelexicale) conținute în cadrul schemelor X-bar. (C4) Atunci când proprietățile funcționale ale categoriilor lexicale nu sunt evaluate şi exploatate corespunzător, teoria X-bar este inconsistentă şi produce dificultăți de calcul şi rezultate incorecte. (C5) Acestea

sunt consecințele unui aspect mult mai general, şi anume că teoria X-bar nu trebuie să fie văzută ca o teorie gramaticală singulară, construită pentru sine, ci ca un dispozitiv component al unui *mecanism lingvistic teoretic şi computațional mai generat, ale* cărui principii să guverneze teoria X-bar. Axiomatica (bazele constructive ale) teoriei X-bar trebuie să fie un *rezultat* al bunei ei funcționări, pe fenomenele concrete de limbaj, şi nu invers! (C6) *Ad limitum,* se poate concepe că mecanismul lingvistic teoretic menționat mai înainte poate funcționa şi fără includerea dispozitivului reprezentat de teoria X-bar, aşa cum încearcă teoria MinP să propună în [3, Cap. *Categories and Transformations]* (dar folosindu-se în explicare tot de aparatul de reprezentare al teoriei X-bar), precum şi în cazul FG [25].

(D) Considerăm că schemele (funcționale) FX-bar propuse furnizează un (sub)sistem necesar şi folositor în cadrul oricărei teorii sintactice asupra LN, inclusiv (şi în special) pentru strategia lingvistică SCD. O condiție esențială pentru schemele FX-bar este ca ele să reflecte corespunzător proprietățile *funcționale* şi *relaționale* ale tuturor categoriilor *lexicale* şi *gramaticale.* Exemplele 4.1.1.-4.1.5 arată cum sunt construite schemele FX-bar, cum se obțin (prin apel recursiv pe nivele) structurile sintagmatice complexe ale LN, şi cum acestea rămân închise la operatorul de compunere (adjuncție) pe baza principiilor şi regulilor SCD.

Schimbând perspectiva, prin definirea teoriei FX-bar ca o componentă a strategiei lingvistice SCD, şi parafrazând formalismul bine-cunoscut al gramaticilo TAG [Eng: *tree adjoining grammar],* strategia SCD poate fi văzută şi ca o teorie a evaluării şi adjuncției de FX-bar scheme. Este doar o mostră a rolului important pe care teoria X-bar îl poate încă juca în cadrul teoriei şi tehnologiei LN.

(E) Un *element original* propus de schemele FX-bar în peisajul teoriilor X bar cunoscute este rolul lor dublu ce îl pot juca în cadrul strategiei SCD (şi n numai): Schemele FX-bar pentru X = N, V, A, CL (CL = clauză) trebuie conceput ca un set de invarianți sintactici (dinamici) ce pot fi folosiți **(1)** la *reprezentare* informației lingvistice la nivel de lexicon (în mod similar cu structurile de trăsătu lingvistice [18], dar într-o manieră mai simplă şi mai regulată), şi (2) la *procesare (analizarea şi generarea) automată* de text în LN (inclusiv, şi mai ales, pentru limb română), de la structurile sintagmatice simple până la cele de discurs.

Derivarea de *automate* şi *gramatici formale bazate* pe schema FX-ba pentru analiza LN, ar fi o consecință normală şi o provocare a prezentei propune Modul *recursiv, ascendent* şi *incremental* (prin apelul de funcții şi relații cu r lingvistic multiplu), dar şi *descendent* (bazat pe sateliții nucleelor semantice utilizarea la maximum a *contextualității* marcherilor de toate tipurile poate reprezen *o* motivație naturală pentru cercetarea relației dintre strategia SCD (cu componen ei de *teorie* FX-bar), şi modelele generative oferite de către *gramaticile contextua Marcus* [41], [42], un formalism *context-dependent* puternic, destinat parsării dar analizei semantice şi de discurs (articularea *topic-focus* [37]) a LN. *Gramatic contextuale Marcus* aparțin unei serii de formalisme care includ *gramatici* TAG [4 *dramatici orientate-nucleu* [15], [16], *gramatici indexate, gramatici X-bar, gramat* context-free *marcate* [44] etc, formalisme ce realizează o modelare mai realistă comportamentului sintactic, semantic şi discursiv al LN.

Referințe bibliografice

[1] N. Curteanu (1990). *A Marker-Hierarchy-based Approach Supporting the SCD Parsing Strategy.* Research Report no. 18, Institute of Technical Cybemetics, Bratislava.

[2] N. Curteanu (1994). *From Morphology to Discourse Through Marker Structures in the SCD Parsing Strategy. A Marker-Hierarchy Based Approach.* Language and Cybemetics, Akademia Libroservo, Prague, 61-73.

[3] Noam Chomsky (1995). *The Minimalist Program.* The MIT Press, Cambridge, Massachusetts.

[4] N. Chomsky (1970). *Remarks on Nominalizations.* In R. Jacobs and P. Rosenbaum (eds.), Readings in Transformational Grammar, Ginn and Co., Boston, 184-221.

[5] T. Stowell (1981). *Origins of Phrase Structure.* Ph.D. Dissertation, Dept. of Linguistics and Philosophy, MIT, Cambridge.

[6] N. Curteanu, G. Holban (1996). *Strategia lingvistică SCD aplicată la analiza şi generarea limbii române.* Limbaj şi Tehnologie (Dan Tufiş, Ed.), Academia Română, Bucureşti, p. 169-176.

[7] N. Curteanu, C. Linteş (2002). *Segmentation Algorithms for Clause-Type Textual Units,* Research Report, Institute of Theoretical Informatics, Romanian Academy.

[8] Noam Chomsky (1986). *Barriers.* The MIT Press, Cambridge.

[9] Noam Chomsky (1981). *Lectures on Government and Binding.* Foris, Dordrecht.

[10] Simon C. Dik (1989). *The Theory of Funcțional Grammar.* Foris Publishers, Dordrecht.

[11] Cari Pollard, Ivan Sag (1994). *Head-Driven Phrase Structure Grammar.* The University of Chicago Press, Chicago & London.

[12] Gerald Gazdar, E. Klein, G. Pullum, I. Sag (1985). *Generalized Phrase Structure Grammar.* Harvard University Press, Cambridge, Massachusetts.

[13] Peter Sells (1985). *Lectures on Contemporary Syntactic Theories.* CSLI, Stanford, California.

[14] Stuart Shieber (1986). *An Introduction to Unification-Based Approaches to Grammar.* CSLI, Stanford, California.

[15] Cari Pollard, Ivan Sag (1987). *Information-based Syntax and Semantics.* CSLI, Stanford, California.

[16] Cari Pollard, Ivan Sag (1994). *Head-Driven Phrase Structure Grammar.* The University of Chicago Press, Chicago & London.

[17] E.P. Stabler Jr. (1992). *The Logical Approach to Syntax: Foundations, Specifications and Implementations of Theories of Government and Binding.* The MIT Press, Cambridge, Massachusetts.

[18] N. Curteanu, G. Holban (2000). *A Set-Theoretic Approach to Linguistic Feature Structures and Unification Algorithms* (I, II). Computer Science Journal of Moldova, 8(2): 116-149, 8(3): 223-246.

[19] Neculai Curteanu (1988). *Augmented X-bar Schemes.* COLING'88 Proceedings, Budapest, 130-132.

[20] Neculai Curteanu, A. Todiraşcu, G. Holban (1997). *Teorii sintactice ale limbajului natural.* Raport de cercetare, Institutul de Informatică Teoretică, Academia Română, Iaşi, 66 p.

[21] Alain Lecomte (1998). *Multimodal Logic for Syntax.* Logica Trianguli, 2: 49-72.

[22] Neculai Curteanu (1983). *Algoritmi de analiză sintactică a frazei şi propoziției româneşti. INFO-IAŞI'83,* p. 533-549.

[23] M. Moortgat (1997). *Categorial Type Logics.* Handbook of Logic and Language, Elsevier.

[24] E.P. Stabler Jr. (1997). *Derivational Minimalism.* Logical Aspects of Computațional Linguistics, LNCS no. 1328, Springer-Verlag, Berlin.

[25] Simon Dik (1989). *The Theory of Funcțional Grammar.* Foris Publishers Dordrecht.

[26] Robert Kasper (1993). *Adjuncts in the Mittelfeld.* în "German Grammar in HPSG" (J. Nerbonne *et al.,* Eds.), CSLI, Stanford, California.

[27] Denis Bouchard (1995). *The Semantics of Syntax. A Minimalist Approach to Grammar.* The Univ. of Chicago Press, Chicago & London.

[28] Julia Hirschberg, D. Litman (1993). *Empirical Studies on the Disambiguation of Cue Phrases.* Computațional Linguistics 19(3): 501-530.

[29] Jacques Jayez, C. Rossari (1999)..*Pragmatic Connectives as Predicates. The Case of Inferential Connectives.* în "Predicative Forms in Natural Languag and in Lexical Knowledge Bases" (P. Saint-Dizier, Ed.), Kluwer Academi Publishers, Dordrecht.

[30] Patrick Saint-Dizier (Ed.) (1999). *Predicative Forms in Natural Language an in Lexical Knowledge Bases.* Kluwer Academic Publishers, Dordrecht.

[31] Daniel Marcu (2000). *The Theory and Practice of Discourse Parsing an Summarization.* The MIT Press, Cambridge.

[32] V. Raskin, S. Nirenburg (1999). *Lexical Rules for Deverbal Adjectives.* "Breadth and Depth of Semantic Lexicons", Kluwer Academic Publisher Dordrecht.

[33] O. Popârda, N. Curteanu (2002). *L'evolution du discours juridique frangais analyse par la strategie linguistique SCD.* In "Representation du Sens Linguistique" (D. Bouchard, Ed.), LINCOM Studies in Theoretical Linguistics, LINCOM EUROPA, Munchen.

[34] Noam Chomsky (2000). *Minimalist inquiries: the framework.* în R. Martin *et al.* (Eds) "Step by step. Esseys on Minimalist Syntax in Honor of Howard Lasnik", MIT Press, Cambridge, p. 89-155.

[35] Noam Chomsky (2001). *Derivation by phase:* în M. Kenstowicz (Ed.) "Ken Hale: a life in language", MIT Press, Cambridge, p. 1-52.

[36] Jane Morris; G. Hirst (1991). *Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text.* Computațional Linguistics 17(1): 21-48.

[37] Eva Hajicova, H. Skoumalova, P. Sgall (1995). *An Automatic Procedure for Topic-Focus Identification.* Computațional Linguistics, 21(1): 81-94.

[38] P. Sgall, E. Hajicova, J. Panevova (1986). *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects.* Kluwer Academic Publishers, Dordrecht.

[39] S. Shieber, F. Pereira, G. Van Noord, R. Moore (1990). *Semantic Head-Driven Generation.* Computațional Linguistics 16(1): 30-41.

[40] Şteven Abney (1996). *Part-Of-Speech Tagging and Parțial Parsing.* în "Corpus-Based Methods in Language and Speech", (K. Church *et al.,* Eds.), Kluwer Acad. Publishers, Dordrecht.

[41] Solomon Marcus (1997). *Contextual Grammars and Natural Language.* în Cap. 5 (Voi. 2) din "The Handbook of Formal Languages", G. Rozenberg, A. Salomaa, Eds., Springer-Verlag, Berlin, 215-235.

[42] Gheorghe Păun (1997). *Marcus Contextual Grammars.* Kluwer Academic Publishers, Dordrecht.

[43] Michele Abrusci, Christophe Fouquere, Jacqueline Vauzeille (1999). *Tree Adjoining Grammars in a Fragment of the Lambeck Calculus. Computațional Linguistics,* 25(2): 209-236.

[44] Philip Miller (1999). *Strong Generative Capacity. The Semantics of Linguistic Formalism.* CSLI Publications, Stanford, California.

[45] D. Gâlea, N. Curteanu, C. Linteş (2002). *Algoritmi de segmentare a textului în unități de tip cauzal.* (în prezentul volum)

Teoria HPSG. Studiu de caz: acordul încrucişat

Ana-Maria BARBU
RACAI, Calea 13 Septembrie nr.13, Bucureşti
abarbu@racai.ro

1. Introducere

Oricât ar fi de mare entuziasmul creat de performanțele realizate c
calculatorul, care cuprinde deopotrivă şi domeniul prelucrării limbajului natura
rezultate temeinice nu se pot obține dacă acestea nu sunt fundamentate p
îndelungi şi profunde analize teoretice. Nu putem aspira la obiective majore
ingineria lingvistică, precum analizarea şi generarea de texte, construirea
verificatoare ortografice şi gramaticale sau chiar de traducătoare automate, da
se ignoră particularitățile inerente ale obiectului în studiu, anume ale limbaju
natural în general, şi a limbii de aplicație, în special. Or aceste particularități su
oferite, sub un aspect sau altul, tocmai de teoriile gramaticale. Experiența
dovedit că eşecurile din ingineria lingvistică au avut ca posibile surse eşecurile
descrierea corespunzătoare a fenomenelor de limbă, dar şi succesele, la rând
lor, s-au datorat în parte acurateței, exactității, şi nu în ultimul rând caracteristici
computaționale ale unui model gramatical teoretic.

Iată de ce alegerea unei teorii lingvistice adecvate, cu scopul de a scrie
baza acesteia o gramatică computațională a unei limbi particulare, în speță a lim
române, este un act de primă însemnătate.

După anii primelor dezvoltări ale gramaticii generative, sintaxa form
este, de aproape două decenii, repusă în discuție ca obiect de studiu auton
distinct în aceiaşi timp de cel al lexicului şi cei al sensului. Mai multe cure
teoretice, cunoscute sub numele generic de "gramatici de unificare" sau "grama
bazate pe constrângeri", s-au născut din această reconsiderare a sintaxei. E
vorba de modele recente (cele mai vechi datând de la începutul anilor '
dezvoltate în cea mai mare parte în Statele Unite, şi în generai apro
necunoscute publicului român. Aceste modele se pretează scrierii de grama
pentru calculator, dar ambiția lor este mai întâi de a constitui teorii lingvistice
sine stătătoare. Autorii lor se înscriu pe linia programului gramaticii genera
chomskyene din 1957, de la care preiau grija pentru o formalizare operator
sintaxei, dar se disting suficient de modelul actual al Şcolii de la Cambridge (n
Government and Binding) pentru a prezenta teorii alternative. Printre punc

comune ale gramaticilor de unificare, se află pe de o parte atenția acordată unei articulări mai explicite a lexicului, sintaxei şi semanticii, pe de altă parte accentul pus pe descrierile lingvistice şi recurgerea la un stil de analiză sintactică mai "concret", care limitează recurgerea la elemente "vide" (nerealizate concret) şi care restrânge numărul etapelor intermediare în producerea unui enunț.

în acest articol vom prezenta pe scurt una dintre teoriile lingvistice amintite, anume "Gramatica sintagmatică ghidată de centru", denumită abreviat HPSG după numele său din engleză "Head-driven Phrase Structure Grammar". Apoi vom ilustra modul în care poate fi aplicată această teorie în reprezentarea unui fenomen mai special de limbă română prin aceea că presupune dependențe încrucişate de acord. Este vorba de structuri relative de tipul *băiatul a cărui soră cântă* unde articolul genitival a se acordă cu substantivul *soră,* iar pronumele relativ *cărui* se acordă cu substantivul *băiatul.*

2. Teoria lingvistică HPSG

2.1. Scurt istoric

Modelul gramaticii sintagmatice ghidate de centru (engl. *Head-driven Phrase Structure Grammar,* sau HPSG) a fost conceput la începutul anilor '80 de Cari Pollard şi Ivan Sag cu scopul de a permite o integrare mai explicită a diferitelor nivele de analiză lingvistică: fonetic, sintactic şi semantic. El a luat naştere în principal din Gramatica Sintagmatică Generalizată (GPSG) şi din lucrările lui C. Pollard despre *Head Grammar* [1], dar autorii lor s-au inspirat deopotrivă din numeroase alte teorii. Ei au preluat de la modelul chomskyan al Guvernării şi Anaforicității (GB) noțiunea de modularitate şi recurgerea la principii foarte generale (Principiul anaforicității, al controlului^etc). De la Gramatica Funcțională de Unificare (FUG) [2] au împrumutat reprezentarea uniformă a elementelor lexicale, a sintagmelor şi regulilor gramaticale sub formă de structuri de trăsături. S-au inspirat de la Gramatica Lexical Funcțională (LFG) pentru îmbogățirea cadrelor de subcategorizare şi a noțiunii de regulă lexicală. Au luat de la gramaticile categoriale ideea de saturare progresivă a predicatelor şi recurgerea la o ierarhie de funcții gramaticale (cf. [3]). S-au inspirat, în sfârşit, dintr-un punct de vedere mai formal, din lucrări de logică şi informatică asupra tipurilor şi moştenirii.

Teoria este prezentată în cele două lucrări ale lui C. Pollard şi Ivan Sag: [4] şi [5]. Majoritatea exemplelor privesc limba engleză şi tratează fenomene variate: fenomene de acord, construcții infinitivale, anafore, construcții relative şi comparative. Fenomenele de control sunt totodată dezvoltate în [6], iar o analiză a anaforelor este propusă în [7]. Primele lucrări au conferit de la bun început o dimensiune multilingvală acestei teorii prin abordări privind germana ([8], [9]), catalana ([10]), japoneza ([11]), dar şi coreana ([12]), franceza ([13]) şi italiana ([14]).

C. Pollard şi I. Sag preiau din modelul GPSG noțiunea de gramatică sintagmatică, cu distincția între o componentă ierarhică (scheme DI -de dominanță imediată) şi o componentă liniară (principii de precedență liniară), precum şi recurgerea la principii foarte generale de partaj şi de propagare a trăsăturilor. Totuşi ei se separă de modelul original în câteva puncte. Structurile sintagmatice sunt în întregime exprimate în termeni de structuri de trăsături, cu introducerea unui atribut Ramuri. Structurile de trăsături sunt la rândul lor organizate în ierarhii de tipuri, comportând fiecare trăsături predefinite. Modelul HPSG oferă astfel anumite simplificări în raport cu GPSG: întregul arsenal de reguli DI este redus la şase scheme de bază; metareguliie sunt eliminate în favoarea regulilor lexicale. S-a urmărit deosebirea clară între ceea ce ține de domeniul constrângerilor universale şi ceea ce ține de descrierea unei limbi particulare. Principiile de coocurență a trăsăturilor din GPSG, care amestecă constrângerile universale şi cele specifice unei limbi date, au fost suprimate.

2.2 Organizarea generală a HPSG

2.2.1 *Caracteristici specifice gramaticilor de unificare*

Se poate considera că gramaticile de unificare, sau gramaticile bazate pe constrângeri, reprezintă noile teorii sintactice ale anilor '80. Este vorba de modele care urmăresc o articulare explicită între lexic, sintaxă şi semantică. Proprietățile lingvistice corespunzătoare sunt concepute ca "informații" asociate morfemelor, sintagmelor sau construcțiilor, combinate prin operații variate, dintre care unificarea ocupă un rol central. Această concepție "integratoare" este unul dintre atuurile lor pentru tratarea automată a limbajelor naturale. Un alt avantaj este că ele se bazează pe modele logice sau matematice (gramatici de constituenți, structuri de trăsături), pentru care au fost definite metode de programare. Ele sunt în general rezultatul unui compromis între expresivitatea lingvistică (grija de a facilita exprimarea diferitor principii lingvistice adăugându-se variante notaționale sau operatori) şi eficacitate (notații concentrate, puține operații).

Aici, ne vom rezuma să punctăm trăsăturile lor comune cele mai pregnante, dintre care:

- reabilitarea descrierilor de suprafață;
- reînnoirea descrierilor sintactice prin definirea de trăsături complexe;
- definirea de principii generale de bună formare a enunțurilor;
- integrarea lexicului, sintaxei şi semanticii.

Gramaticile de unificare îmbogățesc aparatul formal al gramaticilor de constituenți cu un număr de noțiuni importante. în acest capitol ne vom limita la prezentarea principalelor noțiuni utilizate pe parcursul lucrării, pentru detalii putând fi consultate S. Shieber [21] sau H. Uszkoreit [38].

2.2.1.1 *Structuri de trăsături*

Structurile de trăsături (engl. *feature structure)* sunt primitive ale teoriilor sintactice bazate pe unificare şi reprezintă ansambluri de trăsături, numite şi complexe de trăsături (engl. *feature complexes* sau *feature bundles),* care pot fi reprezentate sub formă de matrice. O **trăsătură** este o pereche atribut-valoare, valorile putând fi simboluri atomice sau trăsături. Trăsăturile cu valoare non atomică conduc la structuri de trăsături care prezintă îmbricări.

Spre exemplu, cuvintelor *acest* şi *câine* li se asociază o trăsătură Cat cu valoare atomică (pentru categorie) şi o trăsătură complexă Acord care ia ca valoare conjuncția a două trăsături Num (pentru număr) şi Gen:

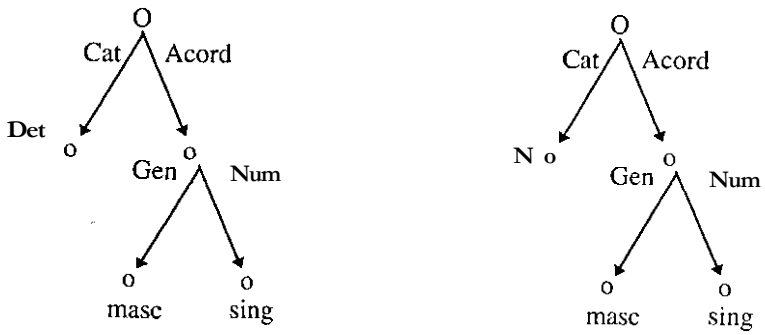
(1)

acest	câine
Cat = Det	Cat = N
Acord = Gen = mase, Num = sing	Acord Gen = mase, Num = sing

O structură este rău formată când conține de două ori acelaşi atribut (la acelaşi nivel de imbricare) cu o valoare diferită.

Şi alte reprezentări de structuri de trăsături (sau structuri atribut-valoare) sunt posibile, fiind echivalente formal. Cele mai utile, pentru implementarea informatică, sunt cele care utilizează grafuri orientate: arcuri care poartă nume de trăsături şi punctează spre noduri care sunt etichetate cu valoarea trăsăturii (dacă e vorba de trăsături cu valoare atomică) sau sunt puncte de plecare pentru alte arce (pentru trăsături cu valoare non atomică). De pildă, pentru exemplele de mai sus vom avea următoarele reprezentări:

(2)



In termeni de grafuri, echivalentul interdicției ca un acelaşi atribut să apară de doua ori la acelaşi nivel cu valori diferite este interdicția ca'două arcuri care poarta aceeaşi etichetă să puncteze, plecând din acelaşi nod, către două noduri

diferite (ceea ce e o restricție generală asupra grafurilor ce corespund automatelor deterministe).

Structurile de grafuri pot fi ciclice sau non ciclice. Acestea din urmă se numesc **grafuri aciclice orientate** (engl. *Directed Acyclic Graph* sau DAG), denumire adesea folosită pentru a desemna structurile de trăsături.

în lucrul cu structuri de trăsături complexe se impun unele distincții, de pildă, între structurile identice şi structurile cu valori partajate (sau reentrante). Cele din urmă sunt identice şi vor rămâne astfel indiferent de modificările suferite ulterior, ceea ce nu se întâmplă cu primele. în exemplul ce urmează structura de trăsături A comportă două atribute cu valori identice Acord şi Num. în structura B, cele două atribute Acord sunt coindexate (prin indicele 1), ceea ce face ca ele să partajeze în mod egal trăsătura [Num = sing].

(3).

A:	B:
Det = [Acord=[Num = sing]]	Det = [Acord = | 11 [Num = sing]]
Nume = [Acord = [Num = sing]]	Nume = [Acord = | 11]

Dacă se unifică fiecare din aceste structuri cu structura C de mai jos rezultatul nu va fi acelaşi:

C: [Det = [Acord = [Gen = mase]]

(4) C ^ A:

Det = [Acord = [Num = sing, Gen = mase]]

Nume - [Acord = [Num = sing]]

(5) C ^ B :

Det = [Acord = 111 [Num = sing, Gen = mase]]

Nume = [Acord = \| 11]

După unificare, trăsătura Acord îmbricată sub atributul Nume va avea şi o trăsătură Gen specificată în cazul lui C u B, dar nu şi în cazul C u A.

în termeni de grafuri, reprezentarea unei structuri reentrante ca B est următoarea:

(6)

B:

Nume — Det

Acord — Acord

Num

sing

2.2.1.2 *Extensiune şi unificare*

Se defineşte o relaţie de **extensiune** între structuri de trăsături după cum urmează:

O structură de trăsături A este o **extensiune** a unei structuri de trăsături B (notându-se A D B) dacă şi numai dacă:

- toate trăsăturile cu valoare atomică prezente în B sunt prezente şi în A cu aceeaşi valoare,

- pentru orice trăsătură <f> cu valoare non atomică, valoarea lui <f> în A este o extensiune a valorii lui <f> în B.

De exemplu, structura de trăsături asociată cuvântului *câine* în (1), reluată în (7) stânga, este o extensiune a structurii din (7), dreapta, dar reciproca nu este adevărată pentru că structura din dreapta nu are trăsătura [Num = sing] prezentă în cea a cuvântului *câine:*

(7)

Cat = N

Acord = Gen = mase; Num =.sing

Cat = N

Acord :➡ [Gen = mase]

Dacă numărul de atribute nu este limitat se poate obţine o infinitate de structuri care sunt extensii ale unei structuri date. Relaţia inversă a extensiei se numeşte **subsumare** , A subsuma B dacă şi numai dacă B este o extensie a lui A.

Pe baza acestei relaţii de ordine parţială putem defini o structură de latice, cu o limită superioară şi o limită inferioară. Este de notat că aici nu există o relaţie de ordine strictă pentru că orice structură este o extensie a ei înseşi (ADA). Structura care le subsumează pe toate celelalte (pentru care toate celelalte sunt extensiuni) este. structura vidă (notată T), pe care o putem interpreta ca disjuncţia tuturor cuplurilor atribut-valoare ale gramaticii. Dacă dorim să plasăm o limită inferioară, structura care va fi o extensie a tuturor celorlalte (care este subsumată de toate celelalte) va fi cea care conţine conjuncţia tuturor cuplurilor atribut-valoare posibile (notată 1) adică o structură "falsă" sau rău formată.

Această relaţie de ordine parţială e folosită pentru a defini unificarea. Această operaţie a luat naştere din cercetările în logică şi informatică (limbajul

Prolog). Definită la început ca procedură de rezolvare pentru logica predicatelor de ordinul întâi, cf. [15], ea a fost introdusă în lingvistică de A. Colmerauer, [16], apoi de M. Kay, [17], pentru a testa, fuziona şi propaga trăsături sintactice. Ea este definită în felul următor:

Unificarea a două structuri de trăsături A şi B (notată A u B) este structura minimală care este în acelaşi timp o extensiune a lui A şi a lui B. Dacă o astfel de structură nu există, unificarea "eşuează" (ceea ce e notat cu 1).

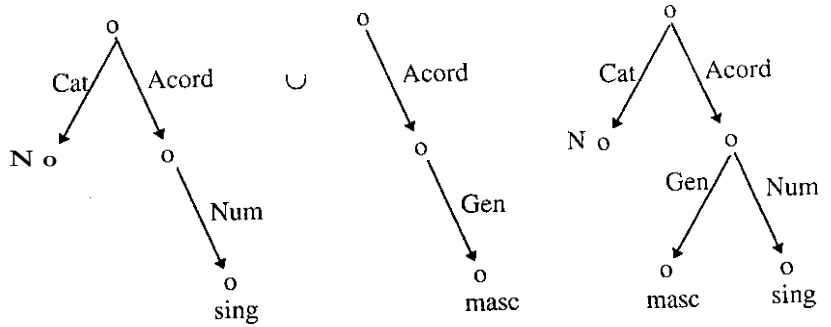
Altfel spus, unificarea verifică compatibilitatea dintre două structuri de trăsături şi produce o structură rezultantă care este cea mai mică structură ce conţine toată informaţia din prima structură şi toată informaţia din a doua structură.

Unificarea este o operaţie idempotentă (A u A = A), comutativă (A u B = B u A) şi asociativă (A u (B u C) = (A u B) u C), spunem de asemenea că este declarativă (dacă A = A' şi B = B' atunci A u B = A ' u B') şi monotonă (A U B D Ş Î A U B D B ; dacă A z) B atunci V C A U C D B U C) , ceea ce vrea să spună că relaţiile de extensiune sunt conservate prin unificare. Colocvial spus, unificarea adaugă informaţie, fără să o scadă.

în termeni de grafuri, echivalentul operaţiei de unificare este fuziune definită pentru automatele cu număr finit de stări. Pentru exemplul din (8a) se obţine reprezentarea grafică din (8b):

(8)

Cat - N; Acord = [Num = sing] U [Acord = [Gen = masc]]= Cat = N; Acord = Gen = mase, Num = sing



Anumiţi operatori pot fi adăugaţi structurilor de trăsături, cei mai utili fi negaţia (notată ~ sau * pentru trăsături cu valoare atomică) şi disjuncţia (not prin acolade sau semnul /). Folosirea negaţiei permite să se renunţe la anum disjuncţii. Există de exemplu echivalenţă între următoarele două ecuaţii, da considerăm că atributul Mod are 8 valori posibile în română (indicativ, conjunc imperativ, prezumtiv, infinitiv, gerunziu, supin, participiu):

[Mod * inf] [Mod = ind/conj7prez/imp/ger/sup/part].

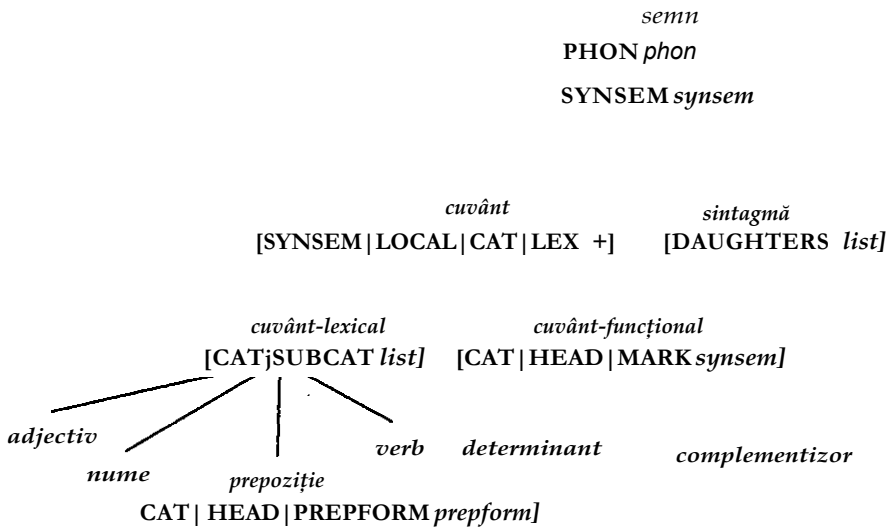
în secțiunea următoare vom trece la descrierea caracteristicilor specifice ale teoriei HPSG care o fac distinctă de toate celelalte teorii bazate pe unificare. Trebuie spus de la bun început că autorii modelului HPSG au preluat o mulțime de caracteristici ale teoriilor apărute anterior, inclusiv de la gramatica generativă, tocmai din dorința de a aduna într-un singur formalism tot ce e mai adecvat pentru reprezenatrea lingvistică în general. Pentru o paralelă detaliată între HPSG şi alte teorii bazate pe constrângeri a se vedea [18].

2.2 Caracteristici specifice HPSG

în HPSG, structurile de trăsături, utilizate în LFG pentru reprezentarea funcțiilor gramaticale, iar în GPSG pentru reprezentarea categoriilor, sunt sistematizate pentru a include atât structurile de constituenți cât şi regulile gramaticale. Ele corespund la ceea ce se numeşte un semn lingvistic, adică un cuvânt, o sintagmă sau o regulă, conținând informații fonetice, sintactice, semantice şi discursive. Structurile de trăsături sunt cât se poate de adecvate pentru organizarea într-o notație comună a informațiilor lingvistice eterogene.

Spre deosebire de celelalte teorii lingvistice bazate pe unificare, HPSG utilizează ierarhizarea tipologică. Fiecare structură de trăsături este încadrată într-un anumit tip pentru care sunt predefinite anumite constrângeri şi care îşi are locul într-o ierarhie de tipuri. în cadrul ierarhiei funcționează relația de moştenire a constrângerilor tipurilor superioare asupra descendenților lor. Un exemplu de ierarhie de tipuri este dată în (9).

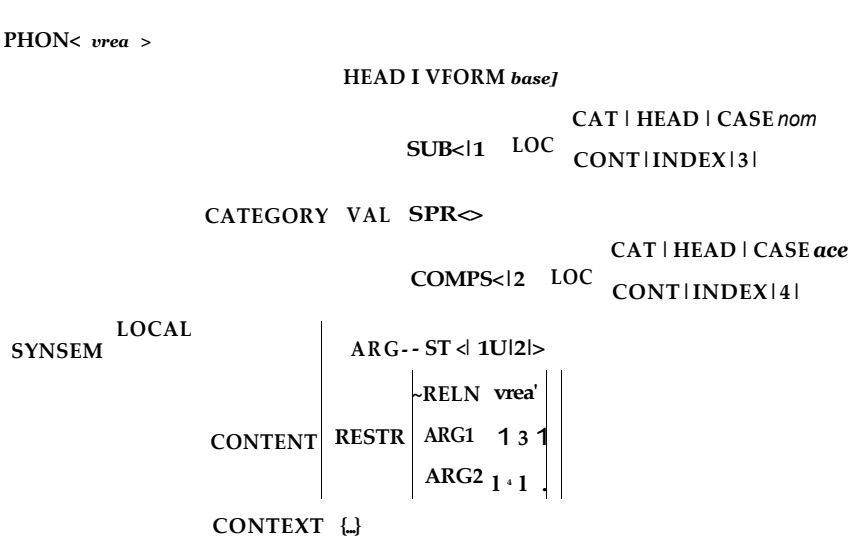
(9)



Pentru fiecare tip sunt definite anumite trăsături specifice (sau anumite constrângeri) care se adaugă constrângerilor moştenite de la tipurile din care descind. Trebuie adăugat că într-o ierarhie de tipuri sunt permise moştenir multiple, adică sunt permise tipuri care au mai mulți părinți.

Cel mai general tip în HPSG este "semnul" (în engleză *sign).* El conține informație fonologică (prin trăsătura PHON) şi informație sintactico-semantică (prin trăsătura SYNSEM). Semnul, la rândul lui, poate fi un cuvânt sau o sintagmă, după cum se vede în (9), mai sus. Sintagma are spre deosebire de cuvânt o trăsătură în plus, numită DAUGHTERS (adică ramuri-surori) care , 3 r /aloare o listă c semnele combinate în sintagmă. Un exemplu de semn lexical împreună c descrierea trăsăturilor specifice acestuia este dată în (10) pentru verbul a *vrea.*

(10)



Combinarea cuvintelor în sintagme se face pe baza unor reguli exprima la rândul lor sub formă de structuri de trăsături tipizate, purtând numele de sche de Dominanță Imediată (scheme DI). Asupra regulilor acționează suplimen principiile, care, la rândul lor, sunt exprimate prin constrângeri asupra anum trăsături. în cele ce urmează vom prezenta pe scurt principiile şi schemele DI bază. Dintre principii, ne rezumăm prezentarea la următoarele:

a. Principiul Trăsăturilor Centrale

b. Principiul de Subcategorizare

c. Principiul Semantic

a. Principiul Trăsăturilor Centrale

Pentru majoritatea sintagmelor se defineşte un atribut HEAD ("centru"), inclus în trăsătura CATEGORY (CAT), a cărui valoare trebuie să fie partajată cu cea a atributului HEAD din semnul ramurii-centru HEAD-DTR a sintagmei. Principiul Trăsăturilor Centrale poate fi exprimat prin descrierea următoare (notând valoarea partajată prin indicele [1]):

(11)
"SYNSEM | CAT | HEAD [1]

DAUGHTERS | HEAD-DTR | SYNSEM | CAT | HEAD [1]

Semnul HEAD-DTR poate fi sintagmatic sau lexical,

b. Principiul de Subcategorizare

Atributul SUBCAT are ca valoare o listă care este actualizată progresiv, pe măsură ce sintagma se "saturează", în sensul că atunci când complementele sunt realizate, ele sunt eliminate din lista SUBCAT a sintagmei respective. O sintagmă se numeşte saturată (sau completă) când valoarea listei SUBCAT este vidă. Principiul de Subcategorizare poate fi enunțat astfel:

Valoarea listei SUBCAT a ramurii HEAD-DTR a unei sintagme trebuie să corespundă concatenării listei L1ca valoare a atributului SUBCAT al sintagmei şi a listei L2 a semnelor ce aparțin ramurii de complemente COMPS-DTR (sau, mai precis, nu lista semnelor, ci a trăsăturilor SYNSEM a acestor semne).

Acesta poate fi reprezentat prin structura de trăsături următoare (notând prin simbolul © concatenarea listelor):

(12)
["SYNSEM | CATEGORY | SUBCAT L1

DAUGHTERS HEAD-DTR | SYNSEM | CAT | SUBCAT L10L2
COMPS-DTR L2

două schime DI: Subcategorizare pot fi descrise următoarele

cam ^ L ^ ^ ^ o Sintagmă S3tUrată CU ramură Comp.emente: ***head-***

(13)
SYNSEM | CATEGORY | SUBCAT <>

DAUGHTERSf HEAD TRISYNSEM I CAT I SUBCAT < X >
COMPS-DTR < X >

2. Schema DI pentru o sintagmă non saturată cu ramură Complemente: ***head-compl***

(14)
SYNSEMj CATEGORY| SUBCAT < X >

DAUGHTERS HEAD-DTR | SYNSEM| CAT| SUBCAT < X,Yl,Y2...Yn>
COMPS-DTR < Yl,Y2...Yn>

3. Schema DI pentru o sintagmă cu ramură Adjunct: ***head-adjunct***

Modificatorii (adjective atributive, adverbe, complemente circumstanțiale) sunt introduşi într-o ramură specială numită ramura Adjunct (sau ADJCT-DTR). Modificatorii selecționează categoria pe care o modifică (N' pentru adjective, V sau GV pentru adverbe). Această selecție se face printr-un atribut MODIF, care are ca valoare o structură de trăsături SYNSEM. Pentru o sintagmă centru-adjunct bine formată trebuie să aibă loc unificarea valorii trăsăturii MODIF a adjunctului cu valoarea trăsăturii SYNSEM a centrului. Astfel adjectivele pot selecționa numele pentru care sunt atribute, iar adverbele pot selecționa verbele respective, adică se poate preciza în intrarea lor lexicală trăsăturile Categorie, Conținut, Index etc. ale numelui sau verbului aşteptat. Descrierea unei sintagme cu Adjunct este următoarea:

(15)
DAUGHTERS HEAD-DTR | SYNSEM |1|
ADJCT-DTR | SYNSEM | CAT | HEAD | MODIF 111

c. Principiul Semantic

Principiul semantic reglementează propagarea trăsăturilor semantice, adică cele două trăsături CONTENT şi CONTEXT. Se urmăreşte pe de o parte ca sintagmele să partajeze valoarea trăsăturii CONTENT din ramura centrului cu trăsătura proprie CONTENT, iar pe de altă parte să determine "ridicarea" la nivelul sintagmelor superioiare a eventualilor cuantificatori şi a variabilelor care le pot corespunde.

HPSG face apel la noțiunea de centru semantic, acesta fiind identic cu centrul sintactic, în afara cazului sintagmelor cu adjunct. în acest caz, centrul Sintactic este categoria modificată, dar centrul semantic este modificatorul (care joacă rolul de predicat semantic). Principiul Semantic poate fi exprimat astfel:

Valoarea atributului CONTENT a categoriei dominante este identică cu valoarea atributului CONTENT a categoriei care este centru semantic (ramura Adjunct sau, implicit, ramura HEAD).

O altă schemă DI, *head-functor,* propusă de Allegranza în [19], reprezintă modificare a schemei *head-adjunct* cu scopul de a satisface exigențele de

reprezentare a determinatorilor într-un grup nominal. Determinatorii sunt trataţi ca functori aplicaţi centrului. Ei selectează centrul prin atributul ARG-SLOT şi marchează sintagma rezultată cu anumite trăsături specifice determinatorului respectiv prin partajarea valorii atributului MARKER între ramura Functor şi nodul mamă. Descrierea acestei scheme este dată mai jos.

4. Schema DI pentru o sintagmă cu ramură Functor: ***head-functor***

(16)
$$\begin{bmatrix} \text{SYNSEM} \mid \text{LOCAL} \mid \text{CAT} \mid \text{MARKER } 1 \\ \text{DAUGHTERS} \begin{bmatrix} \text{FUN-DTR} \mid \text{SYNSEM} \mid \text{LOCAL} \mid \text{CAT} \begin{bmatrix} \text{HEAD} \begin{bmatrix} \text{MARKER } 1 \\ \text{ARG-SLOT } 2 \end{bmatrix} \end{bmatrix} \\ \text{HEAD-DTR} \mid \text{SYNSEM} \mid 2 \end{bmatrix} \end{bmatrix}$$

Cu aparatul formal oferit de HPSG, în secţiunea care urmează, dăm spre exemplificare analizarea unei structuri concrete din limba română. Structura propusă conţine un centru nominal modificat de o propoziţie relativă al cărei element de relaţie este în cazul genitiv precedat de articolul genitival. Această structură este interesantă prin faptul că prezintă un fenomen, acela de acord încrucişat, care pare să scape reprezentărilor gramaticilor independente de context. Avantajul teoriei lingvistice discutate aici, însă, oferă o soluţie pe cât de unitară, pe atât de elegantă, după cum sperăm să reiasă din cele ce urmează.

3. Structuri relative cu acord încrucişat

în limba română structurile care prezintă acord încrucişat sunt propoziţiile relative în care pronumele relativ este precedat de articolul genitival, ca în exemplul de mai jos.

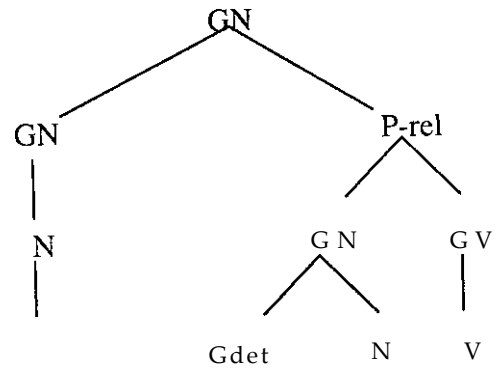
(17) băiatul a cărui soră cântă

Acordul este încrucişat prin aceea că pronumele relativ propriu-zis se acordă cu substantivul determinat de propoziţia relativă, *băiatul,* iar articolul genitival *al* se acordă cu subiectul relativei, soră, după următoarea schemă:

(18) **băiatul a cărui soră cântă**

Structura internă a acestui grup nominal este reprezentată în arborele de mai jos.

(19)



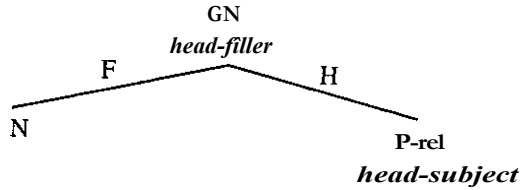
băiatul cărui soră cântă

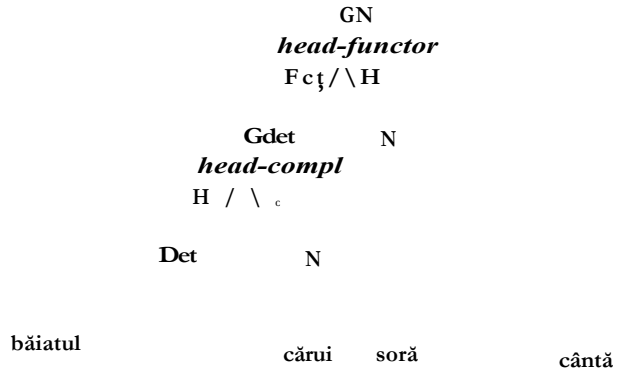
După cum se vede în acest arbore, exemplul din (17) este format dintr-un substantiv centru, *băiatul,* modificat de o propoziţie relativă al cărei subiect, a *cărui soră,* cuprinde elementul de relaţie care face legătura dintre numele amintit şi propoziţia relativă.

Dacă ne-am limita descrierea la regulile independente de context sugerate în arbore, nu am putea da seamă de fenomenul de acord încrucişat pe care-l discutăm aici. Acest lucru este însă posibil dacă folosim o gramatică HPSG, beneficiind de avantajele oferite de mecanismul unificării şi de reprezentările prin structuri de trăsături.

Aplicând schemele de dominanţă imediată şi principiile specifice teoriei HPSG, arborele de mai sus poate fi adnotat cu regulile HPSG aplicate, în felul următor (unde am folosit ca notaţii funcţionale H=centrul sintagmei, C=complement, Fct=functor, F=filler).

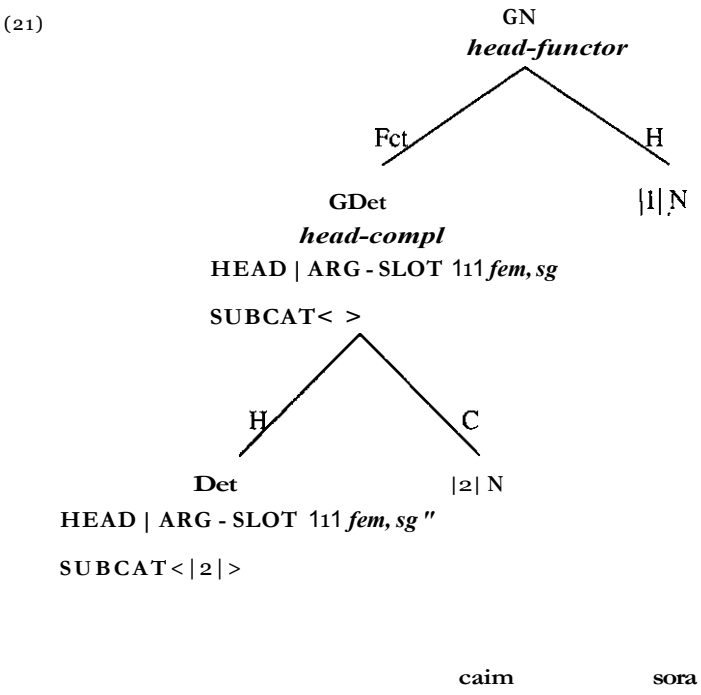
(20)





Fenomenul de acord încrucişat presupune pe de o parte acordul determinatorului a cu substantivul *soră,* iar pe de altă parte acordul pronumelui relativ *cărui* cu substantivul *băiatul.* Primul acord amintit se face relativ banal. Intrarea lexicală a determinatorului a, în calitatea sa de functor, specifică în valoarea atributului său central ARG-SLOT ce trăsături de acord trebuie să aibă substantivul pe care urmează să-l modifice. Când detrminatorul a se combină cu complementul său *cărui,* principiul trăsăturilor centrale face ca această informaţie să fie percolată la nodul mamă GDet. Mai departe, schema DI head-functor verifică dacă trăsăturile de acord ale GDet unifică cu cele ale centrului său nominal. Acest mecanism este ilustrat în arborele de mai jos.

(21)



Al doilea tip de acord, în schimb, ridică anumite dificultăţi p se realizează într-un arbore local, adică nu se realizează între ram aceluiaşi nod. Prin urmare, trăsăturile de acord ale pronumelui r percolate până la nivelul nodului P-rel (din (19)) pentru a putea fi unificare de regula head-filler cu trăsăturile de acord c substantivului determinat.

Mecanismul din teoria HPSG care dă seama de propagare anumitor trăsături se numeşte mecanismul dependenţelor la distar fenomenelor de limbă precum interogaţiile, topicalizările şi, cum-est construcţiile relative. Aici ne vom ocupa numai de tratarea rel celelalte fenomene a se vedea [5].

Ideea principală a acestui mecanism este că pronumele re intrările lor lexicale informaţii despre numele la care se referă. Intr Pronumelui relativ din exemplul nostru va conţine, prin urmare, info **(22).**

PHONOLOGY < *cărui* >

(22)

```
                                  HEAD noun[ genitive]
                CATEGORY
                                  SUBCAT o
LOCAL   CONTENT nproflNDEX 111 [ masculine, singular ]
                                                  RELN human
                                                  _INSTANCE|1|
SYNSEM
                                  QUE{}
                INHERITED         REL{|1
                                  SLASH{
NONLOCAL
                                  QUE{}
                TO - BIND         REL{}
                                  SLASHU
```

Valoarea trăsăturii NONLOCAL | INHERITED indică acele trăsături care vor fi supuse Principiului Trăsăturilor Nonlocale. Aceste trăsături pot fi specifice elemetelor interogative, definite prin atributul QUE, elementelor dislocate, date de atributul SLASH sau pot fi specifice elementelor relative indicate prin atributul REL. După cum se observă în (22), acest ultim atribut are în cazul de față valoare non-vidă, coindexată cu conținutul semantic de masculin-singular al pronumelui.

Potrivit Principiului Trăsăturilor Nonlocale, formulat în (23), valoarea atributului nonlocal INHERITED ("moştenit") este trecută din nod în nod spre vârful arborelui până va întâlni o ramură soră ale cărei trăsături locale unifică cu cele moştenite.

(23) **Principiului Trăsăturilor Nonlocale**

Pentru fiecare trăsătură nonlocală, valoarea atributului INHERITED a nodului mamă este egală cu reuniunea valorilor atributului INHERITED ale ramurilor fiice mai puțin valoarea atributului TO-BIND a ramurii centru.

Atributul TO-BIND, practic, opreşte propagarea trăsăturilor moştenite în momentul în care se realizează elementul căutat, adică elementul care a făcut necesară această propagare. De exemplu, trăsăturile de acord ale pronumelui relativ, în exemplul nostru *cărui,* se propagă la nivelul propoziției relative până când este realizat substantivul la care se referă acest pronume, adică *băiatul.*

Regula care asignează o valoare atributului TO-BIND în momentul în care are loc unificarea trăsăturilor locale ale unui element cu trăsăturile moştenite pe

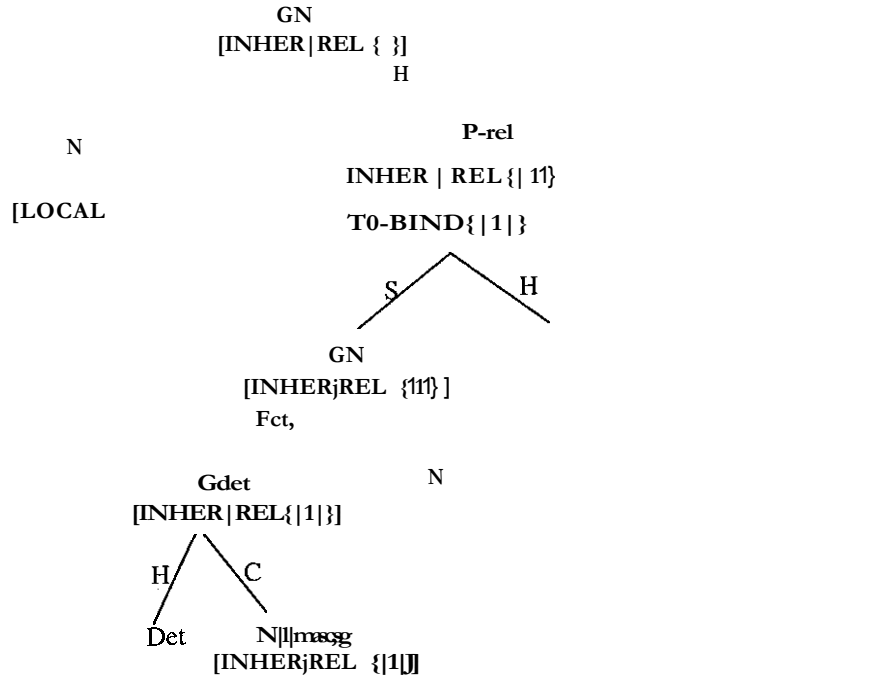
PQ rpntnl pste o sch emă de dominanță imediată numită *head-filler (filler* ar „Tal*. î " d ? e p » -ceTa ce vine sâ completeze *o* lipsâ") si es,e descnsa ,n (24).

(24) Schema DI ***head-filler***

```
                FILLER-DTR | SYNSEM | LOCAL 111                    Î
                                                  HEAD | VFQRM finite
                                  LOCALI CATEG
                                                  SUBCAT o
DAUGHTERS
                HEAD-DTR  SYNSEM
                                                  INHERITED | REL{|1!,.-}
                                  NONLOCAL
                                                  T0-BIND | REL{|1|}
```

în Târ*it dacă aplicăm Principiul Trăsăturilor Nonlocale şi schema DI *head-*Z^ M ? U care if-vem în vedere se realizează în man.era ilustrată în arborele de mai jos.

(25)



în concluzie, acordul încrucişat avut în vedere presupune, pe de o parte, un acord local, cel dintre articolul genitival şi substantivul determinat, în cazul nostru subiectul propoziției relative, iar pe de altă parte un acord la distanță, cel dintre pronumele relativ şi substantivul determinat, exterior propoziției relative. Primul tip de acord se face pe baza Principiului Trăsăturilor Centrale şi a acordului banal dintre functor şi centrul său, pe când cel de al doilea tip de acord face uz de Principiul Trăsăturilor Nonlocale şi de schema de Dominanță Imediată *head-filler*.

4. Concluzii

Analiza oferită aici pune în lumină faptul că un fenomen dificil precum acordul încrucişat poate fi tratat într-o manieră relativ simplă şi elegantă cu ajutorul unei teorii lingvistice adecvate, cum este teoria Head-driven Phrase Structure Grammar.

Prin aparatul formal şi adecvarea lingvistică pe care le oferă această teorie, descrierea fenomenelor limbii române devine incontestabil mai unitară, mai explicită şi mult mai riguroasă. Acestor avantaje li se adaugă încă unul, extrem de important, acela al adecvării teoriei pentru implementarea informatică. Este deschis astfel drumul pentru construirea de gramatici computaționale ale limbii române şi dezvoltarea componentei informatizate a acesteia.

Aplicațiile informatice ale teoriei HPSG sunt, de altfel, în plină dezvoltare şi nu am dori să încheiem înainte de a aminti câteva aspecte în acest sens. ^

Modelul HPSG a făcut parte încă de la origine dintr-un sistem de tratare automată a englezei dezvoltat în laboratoarele de cercetare Hewlett Packard din Palo Alto ([20]). Apoi, au fost propuse diferite implementări, unele bazate pe sistemul PATR ([21]), altele realizate direct în Prolog ([22], [23]). Dintre implementările de sisteme de gestiune a structurilor de trăsături tipologizate şi cu moştenire, se poate cita sistemul *Typed Feature Structure* (TFS) al lui M.Emele şi R. Zajac [24] şi sistemul ALE al lui B. Carpenter [25].

Teoria HPSG a inspirat deopotrivă noul formalism european ALEP, a cărui implementare (în Prolog) presupune un mecanism de gestionare de gramatici şi lexicoane, un analizor, un generator şi un modul de transfer pentru traducerile automate. Este de altfel utilizat în mai multe centre de cercetare universitară (precum DFKI la Saarbrucken, *Centerfor Cognitive Science* în statul Ohio, CSLI la Stanford) sau industriale, în special la ATR în Japonia (pentru traducerea automată englezo-japoneză pentru stabilirea de întâlniri prin telefon).

O altă aplicație informatică a acestei teorii, pe cât de recentă, pe atât de importantă este cea cuprinsă în proiectul Verbmobil, [26], care s-a ocupat cu traducerea bidirecțională, în timp real, a textelor vorbite în trei limbi (germană, engleză şi japoneză).

Head-driven Phrase Structure Grammar este o teorie care s-a impus incontestabil în ligvistica modernă atât prin numeroasele sale aplicații informatice, cât şi prin "generalitatea" aparatului său care o face adecvată pentru numeroase limbi ale lumii, aşa cum se poate vedea din impresionanta bibliografie electronică HPSG oferită de pagina www.dfki.de/lt/HPSG. Nu trebuie trecute cu vederea lucrările de limba română dezvoltate în acest cadru, dintre care le amintim pe cele ale lui Ionescu ([27]-[33]), Monachesi ([34]-[36]) şi Barbu ([37]) la care s-ar cuveni să se adauge multe altele spre afirmarea limbii române în lingvistica internațională.

Referințe bibliografice

[1] Polard, C. - *Generalized Context-Free Grammars, Head Grammars and Natural Language. Teză* de doctorat. Universitatea din Stanford, 1984.

[2] Kay, Martin - "Funcțional Grammars", *Actes 5° annual meeting of the Berkeley Linguistics Society,* Berkeley, 1979, pp. 142-158.

[3] Oehrle, Richard; Bach, Emmon; Wheeler, Deirdre (eds.) - "Categorial Grammars and Natural Language Structures", Dordrecht: Reidel, 1988.

[4] Pollard, C; Sag, I. - *Information-based Syntax and Semantics,* CSLI, University of Chicago Press, 1987.

[5] Pollard, C; Sag, I. - *Head-driven Phrase Structure Grammar,* CSLI, University of Chicago Press, 1994.

[6] Sag, I.; Pollard, C. - "An integrated theory of complement control"', *Language,* 67:1, 1991, pp. 63-113.

[7] Pollard, C; Sag, I. - "Anaphors in English and the scope of binding theory", *Linguistic Inquiry,* 23:2, 1992, pp. 261-303.

[8] Pollard, C. - "On head non-movement", Actele Colocviului *Discontinuous constituency,* Tilburg, 1990.

[9] Nerbonne, J.; Netter, K.; Pollard, C. (eds.) - "German grammar in HPSG", CSLI, University of Chicago Press, 1993.

[10] Balari, S. - "Feature structures, linguistic information and grammatical theory" Teză de doctorat, Universitatea Autonomă din Barcelona, 1993.

[11] Gunji, T. - *Japanese Phrase Structure Grammar,* Reidel, 1987.

[12] Chung, C. - "Korean auxiliary verb constructions without VP modes", *Harvar Workshop on Korean Linguistics,* V; în C. Pollard, I. Sag (eds.), *Readings i HPSG,* 1993.

[13] Miller, P.; Sag, I. - *French clitic movement without clitics or movement,* LSA Meeting, Los Arigeles, 1993.

[14] Monachesi, P. - "Oject clitics and clitic climbing in Italian HPSG grammar", *Actes 6° European ACL,* Utrecht, 1993, pp. 431-437.

[15] Robinson, J. - "A machine-oriented logic based on the resolution principie", *Journal of the ACM,* 12, 1965, pp.23-44.

[16] Colmerauer, A. - "Les grammaires de metamorphose", Universite d'Aix Marseille, 1975, reluat în L. Bolc (ed.) *Natural Language Communication with computers,* Springer, Verlag, 1978.

[17] Kay, M. - "Funcțional grammars", *Actes 5° annual meeting of the Berkeley Linguistics Society,* Berkeley, 1979, pp. 142-158.

[18] Abeille, A. - *Les nouvelles syntaxes. Grammaires d'unification et analyse du frangais,* Armând Colin, Paris, 1993.

[19] Allegranza, V. - "Determiners as Functors: NP Structure in Italian" în S. Balari & L. Dini (eds.) *Romance in HPSG,* CSLI, Stanford, 1998.

[20] Proudian, D.; Pollard, C. - "Parsing Head-driven Phrase Structure Grammar", *Actes 23°ACL,* Chicago, 1985, pp. 167-171.

[21] Shieber, S. - *An Introduction to unification-based theories of grammar,* CSLI, University of Chicago Press, 1986.

[22] Oliva, K. - "Simple parser for an HPSG-style grammar implemented in Prolog", *Actes13°COLING,* Helsinki, vol.3,1990, pp.434-436.

[23] Carpenter, B. - "The generative power of Categorial grammars and Head-driven Phrase Structure grammar with lexical rules", *Computațional Linguistics,* 17:3, 1991, pp. 301-314.

[24] Emele, M.; Zajac, R. - "Typed-unification grammars", *Actes 13° COLING,* Helsinki, vol.3, 1990, pp. 293-298.

[25] Carpenter, B. - "The Logic of typed Feature Structures with Applications to Unification Grammars, Logic Programs and Constraint Resolution", Cambridge University Press [Implementarea sistemului ALE], 1992.

[26] Wahlster, W. (ed.) - *Verbmobil: Foundations of Speech-to-Speech Translation,* Springer, Berlin, 2000.

[27] Ionescu, E. - "A Type of SOV Construction in Romanian", *"Cahiers de Linguistique Theorique et Appliquee",* tomes XXXII-XXXIII, 1995-1996, 19-39.

[28] Ionescu, E. - "Accusative Weak Pronouns in Romanian", *Cahiers de Linguistique Theorique et Appliquee,* tomes XXXII-XXXIII, 1995-1996, 40-52.

[29] Ionescu, E. - "Accusative Clitic Doubling in Romanian", *Cahiers de Linguistique Theorique et Appliquee* tomes XXXII-XXXIII, 1995-1996, 53-73.

[30] Ionescu, E. - "Accusative Clitic Climbing in Romanian", *Cahiers de Linguistique Theorique et Appliquee,* tomes XXXII-XXXIII, 1995-1996, 74-87.

[31] Ionescu, E. - "A Quantification-based Approach to Negative Concord in Romanian" in Geert-Jan M. Kruijff and Richard T. Oehrle (editori), *Proceedings of Formal Grammar Conference Utrecht*,1999, p. 25-36.

[32] Ionescu, E. - *Pro-Drop: An HPSG Account without Lexical Rules,* "Bucharest Working Papers in Linguistics", voi. I, nr.1, 1999, 117-124.

[33] Ionescu, E. - *On the Status of PE in the Direct Object Construction in Romanian,* Romanian Journal of Information Science and Technology, volume 4, numbers 3-4, 2001, p. 293-310.

[34] Monachesi, P. - "The morphosyntax of Romanian cliticization" în P.-A. Coppen, H. van Halteren, & L. Teunissen, eds., *Proceeding of Computațional Linguistics in The Netherlands 1997,* pp. 99-118, Amsterdam-Atlanta:Rodopi.

[35] Monachesi, P. - "Linearization properties of the Romanian verbal complex" în *Proceedings of WECOL* 98, Tempe, 1999.

[36] Monachesi, P. - "Clitic Placement in the Romanian verbal complex", în B. Gerlach and J. Grijzenhout (eds.) *Clitics in Phonology, Morphology and Syntax,* LA 36, Amsterdam: John Benjamins Publishing Company, 2000.

[37] Barbu, A.M. - "Romanian determiners:order and classification" în *Revue Roumaine de Linguistique,* XLIII, nr.5-6, pp.299-315, Bucureşti, 1998.

[38] Uszkoreit, H. - "From Feature Bundles to Abstract Data Types: New Directions in the Representation of Linguistic Knowledge, in H. Blaser *Natural Language at the Computer,* Berlin: Springer, 1989.

După 10 ani de experiență terminografică: noul model de date terminologice al TermRom

Dan MATEI, Institutul de Memorie Culturală
Piața Presei Libere, nr. 1, C.P. 33-90, 713411, Bucureşti
dan@cimec.ro

A. Preambul

Din 1991 − când a fost înființată — Asociația Română de Terminologie (TermRom) a desfăşurat o activitate terminografică materializată într-o bază de date proprie (accesibilă, în parte, pe web la www.cimec.ro/tr/) şi într-o serie de publicații specifice. Formatul terminografic utilizat — descris în [1] −, derivat din formatul standard MicroMATER (ISO 6156), se bazează pe un model de date (relativ) complex, serializat pe două nivele: nivelul conceptului şi nivelul termenului. Practica terminografică (ce se traduce prin prelucrarea unei mari diversități de date terminologice) ne-a revelat o tensiune între complexitatea datelor reale şi insuficienta complexitate a modelului folosit. în plus, necesitatea transferului de date între aplicații diverse a scos la iveală utilitatea consemnării cu o granularitate sporită a elementelor înregistrării terminologice. Mai mult, "entuziasmul" cu care ISO revizuieşte standardele terminologice în ultimii ani[1], cu alte cuvinte, relativa instabilitate a standardelor din acest domeniu, îndeamnă la o şi mai fină granularitate, pentru a spori şansele de compatibilitate cu normele de transfer viitoare. Pe de altă parte, pe măsura acumulării experienței, era din ce în ce mai limpede că modelul de date folosit ar trebui să acomodeze o mai mare diversitate şi complexitate de metadate bibliografice, ca şi o fină şi flexibilă tratare a metadatelor "administrative", de gestionare a colecției terminologice (vezi şi [2]).

Aceste considerente au dus la elaborarea unui model de date obiectual, care, pe lângă cerințele expuse mai sus, să fie şi suficient de abstract ca să permită o serializare convenabilă (pentru transfer de date), — probabil bazată pe XML, de exemplu în formatul MARTIF [3] — şi să nu ceară elaborarea de aplicații informatice de o complexitate excesivă.

[1] *Atât ISO 12200 cât şi ISO 12620 sunt în revizie (deşi ambele datează doar din 1999), iar ISO 16642, este încă nedefinitivat. Desigur, această stare a lucrurilor probează şi faptul că domeniul nu este încă bine "aşezat".*

B. Modelul

Clasă

Generalizare

Dependență

Asociere

Convențional, modelul este împărțit în secțiuni ("pachete" [packages], în terminologia UML). La nivelul cel mai de sus, se disting secțiunea (așa zis) funcțională şi secțiunea administrativă.

B.1. Secțiunea funcțională

în fig. 1 se prezintă clasele funcționale esențiale şi asocierile lor. Practic, orice element al modelului este o 'înregistrare'. Cu alte cuvinte, 'înregistrare' este clasa generică. Existența unei clase generice oferă — pe lângă gruparea proprietăților comune tuturor elementelor — şi posibilitatea de a avea un identificator unic pentru fiecare înregistrare din baza de date ce implementează acest model.

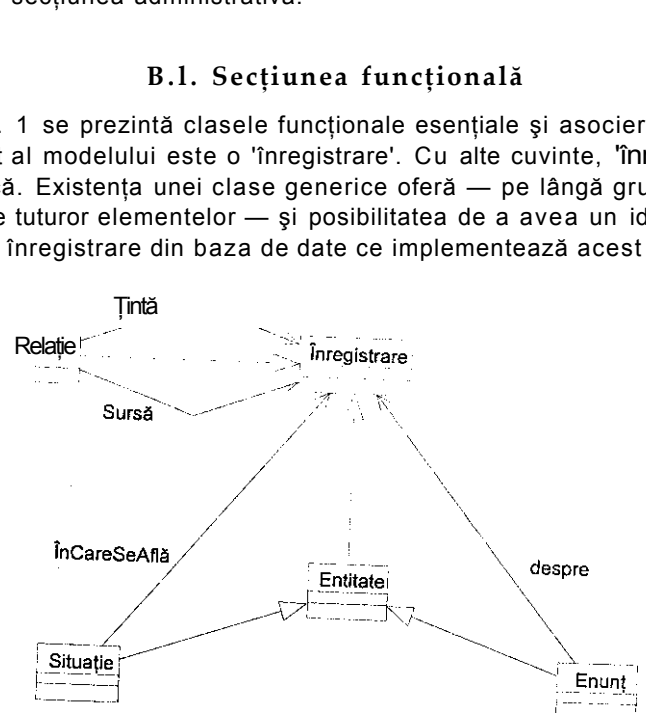


Figura 1 - Secțiunea funcțională (generică)

Clasa 'înregistrare' are două subclase: 'entitate' (care grupe ce au o existență autonomă) şi 'relație' (care grupează asocieri înregistrări). Se observă că sunt acceptabile chiar şi relațiile bina lucru folositor şi în practică.

Reificarea relațiilor binare între înregistrări simplifică m constituie o manieră flexibilă de a consemna o mare varietate de elementele modelului. O relație R poate avea două caracteristici utile în cadrul modelului:

a. simetria: dacă x este în relația R cu y, y este în relația R
b. tranzitivitatea: dacă x este în relația R cu y şi y este în r este în relația R cu z.

Pentru fiecare instanță a clasei 'relație', aceste caracteristic se consemnează ca un atribut al tipului respectiv de relații (nerepre model)[2]. Consemnarea acestor proprietăți ale relațiilor poate fi fo pentru programele care ar exploata baza de date.

Pentru a se rezolva (relativ) simplu şi flexibil asocierile înregistrări, s-a introdus subclasa 'situație' a clasei 'entitate'. După figură, o instanță (sau mai multe) a clasei 'situație' se asociază c clasei 'înregistrare', iar obiectul 'situație' este conectat cu oricâte alt instanțe banale ale clasei 'relație'. în practică, cele mai frecven acestui tip de obiect sunt ca încarnări de contexte şi evenimente. în două subclasă a clasei 'entitate' este 'enunț'. Acest tip de obiect consemna atribute ale unei înregistrări care n-au fost aprioric prevă cu alte cuvinte el găzduieşte mențiuni pentru care se doreşte un simplelor note, şi anume care se doresc a fi colocabile şi/sau indexa

în continuare se prezintă doar subsecțiunile secțiunii funcțio de interes în contextul acestui volum.

B.1.1. Secțiunea terminologică

Fig. 2 prezintă entitățile (i.e. subclasele clasei 'entitate') d nologică.

O categorie de relații importantă în terminologie — este cea a r tranzitive şi asimetrice.

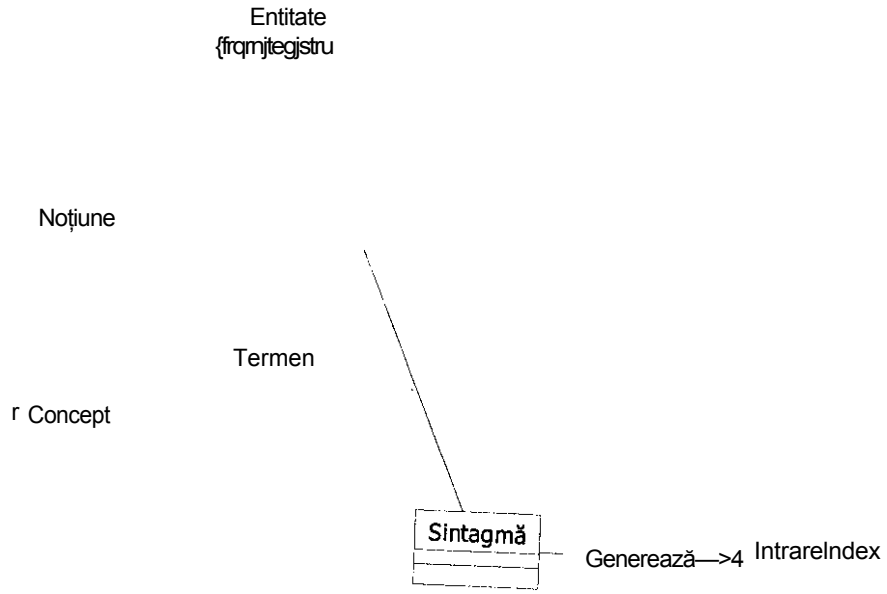


Figura 2 - Secțiunea terminologi-ca

Principala clasă a acestei subsecțiuni este 'noțiune'. Instanțele ei consemnează noțiunile vehiculate în baza de date terminologică, independent de limbă. Din rațiuni practice, şi anume din necesitatea de a cuprinde în baze de date terminologice şi materialul organizat de obicei în tezaure terminologice, s-a decis să se cuprindă în modelul de date nu doar conceptele pure, ci şi unități semantice mai largi, precum cele desemnate de termenii compuşi într-un tezaur (sau ceea ce ISO 12620 numeşte 'unități frazeologice' [A.2.1.18]). Clasa acestor unități conceptuale care cuprinde conceptele şi unitățile semantice mai largi este clasa 'noțiune'. Distincția fină între 'noțiune' şi 'concept' este formulată în logică astfel [5]:

Noțiune: formă logică fundamentală care reflectă însuşirile caracteristice necesare şi generale ale unei clase de obiecte.

Concept: noțiune care reflectă însuşirile esențiale ale unei clase de obiecte[3].

Aşadar, o noțiune care nu e concept cuprinde mai mulți factori semantici, deci poate fi factorizată[1].

[1] d e **n *** , M r e n u s u n l c o n c e p , e ; b | o n a · i m w t o t _

A doua subclasă a acestei secțiuni este 'termen'. Instan nează doar "denumirile" conceptelor (A.1. în ISO 12620). consemnează ceea ce au în comun o familie de expresii lingvistic un concept[5]. Expresiile lingvistice propriu-zise sunt consemnate în 'sintagmă'[6]. Din pricina faptului că un termen poate fi exprimat expresii lingvistice (flexiuni, variante ortografice etc), s-a pre "termenului" de expresiile sale lingvistice, în felul acesta nu ne definiției pentru 'termen', din ISO 12620 (A.1): "a designation of a in a special language by a linguistic expression".

Se poate observa în figură faptul că sintagmele generează în fapt, o sintagmă poate genera — prin inversare/permutare — m index, dacă terminograful decide că asta ar fi în folosul utilizatorilo sintagmei la fiecare "factor" semnificativ. Exemple:

Sintagma	**Intrări de index**
efect Doppler	efect Doppler Doppler, efect
pseudofonetism	pseudofonetism fonetism, pseudo-
completivă indirectă anticipată	compietivă indirectă anticipată indirectă anticipată, completivă anticipată, completivă indirectă

Clasa 'relație' este vitală pentru consemnarea asocierii modelului. Pentru a ilustra modul în care se consemnează informa esențială, în fig. 3 s-au reprezentat tipurile de relații esențiale care o parte, conceptele cu termenii care le designează, iar pe de sintagmele care-i exprimă. De asemenea, se vede cum o "situ această ilustrare — implică (cel puțin) un loc, o perioad caracterizează designarea.

O regulă simplă, pragmatică de a distinge o noțiune care este conce este, ni se pare: noțiunea care e concept şi-ar găsi locul într-un dicționar hu e, nu.

[6] *Exemple de "familie de expresii lingvistice" sunt: a) clădire, clădiri; b) e expresivi, expresive.*

în acest context, 'sintagmă' desemnează — printr-un abuz de limbaj - a cuvinte.

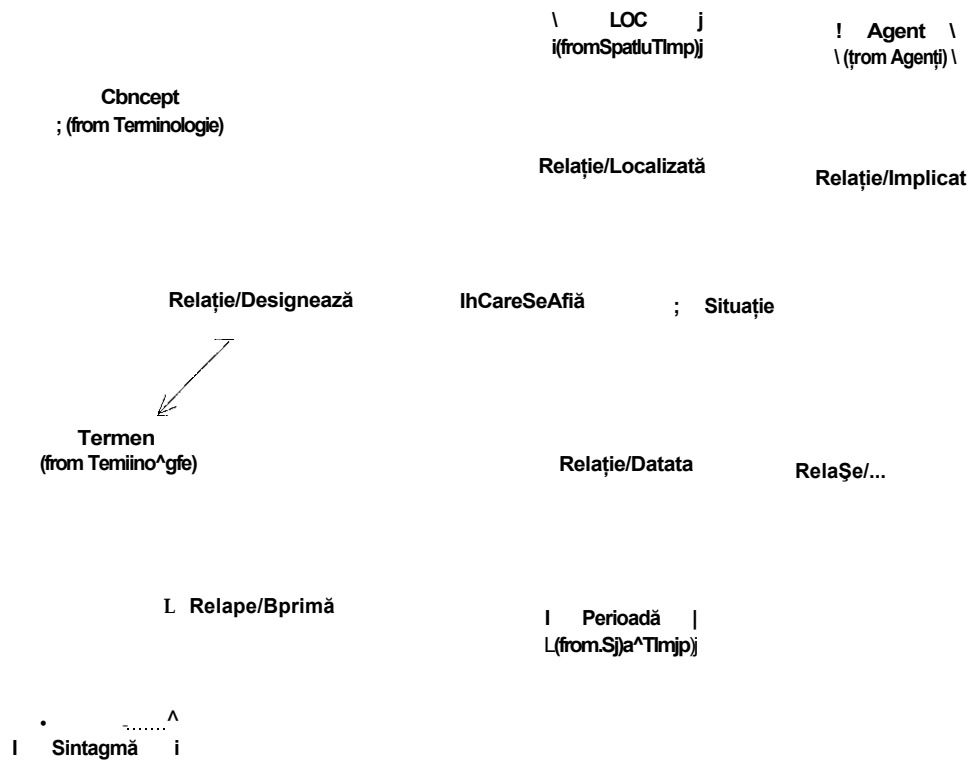


Figura 3 - Ilustrare a reprezentării informației terminologice

într-o astfel de schemă, se pot reprezenta cu acuratețe cazuri precum:

a) Concept: *mic arbust cu flori roşietice din familia ericaceae ...*

- Relație/designează:
  - Termen (ştiințific) [latină]:
    - Relație/exprimă:
      - Sintagmă: *Kalmia latifolia*
- Relație/designează:
  - Situație/context:
    - Relație/localizează:
      - Loc: *nordul Statelor Unite*
  - Termen [engleză]:
    - Relație/exprimă:
      - Sintagmă: *mountain laurel*
- Relație/designează:
  - Situație/context:
    - Relație/localizează:
      - Loc: *sudul Statelor Unite*
  - Termen [engleză]:
    - Relație/exprimă:
      - Sintagmă: *calico bush*
- Relație/designează:
  - Situație/context:
    - Relație/localizează:
      - Loc: *sudul Statelor Unite*
  - Termen [engleză]:
    - Relație/exprimă:
      - Sintagmă: *sheep's bane*
- Relație/designează:
  - Termen [română]:
    - Relație/exprimă:
      - Sintagmă [s.m.sg.]: *laur de munte*
    - Relație/exprimă:
      - Sintagmă [s.m.pl.]: *lauri de munte*

b) Concept: comandant de călărime

- Relație/designează:
  - Situație/context:
    - Relație/localizează:
      - Loc: Moldova
    - Relație/localizează:
      - Loc: Țara Românească
    - Relație/datează:
      - Perioadă: sec. XVII-XVIII
  - Termen [română]:
    - Relație/exprimă:
      - Sintagmă [s.m.sg.]: *serdar*
    - Relație/exprimă:
      - Sintagmă [s.m.pl.]: *serdari*

c) Concept: boier de rang mijlociu

- Relație/designează
  Situație/context:
  Relație/datează:
  Perioadă: *sec. XVIII-XIX*
  Termen [română]:
  Relație/exprimă:
  Sintagmă: *serdar* [s.m.sg.]
  Relație/exprimă:
  Sintagmă: *serdari* [s.m.pl.]

Tot ca o ilustrare, în fig. 4 se prezintă modul cum se consemnează etimologia unui termen, cu ajutorul clasei 'situație': o situație de tip 'etimologie' se asociază cu termenul de bază, iar termenii din care acesta provine sunt asociați cu situația prin intermediul unor relații de tip 'provineDin'.

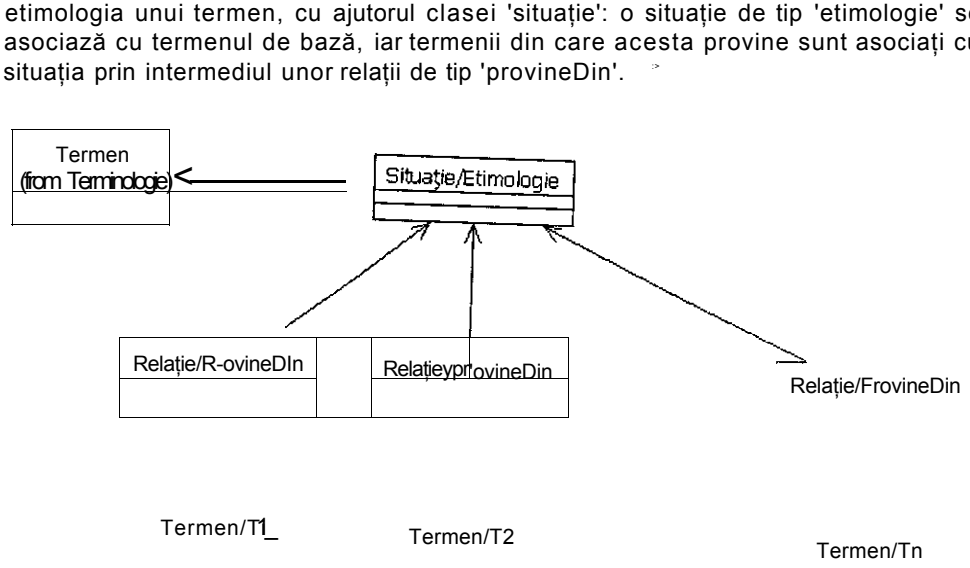


Figura 4 - Ilustrare a reprezentării etimologiei

De pildă:
Concept: fixat la vârf
Relație/designează:
Termen [română]:
Relație/exprimă:
Sintagmă: acrofix
Situație/etimologie:
Relație/provine din:
Termen [greacă]:
Relație/exprimă:
Sintagmă:
Relație/provinedin:
Termen [latină]:
Relație/exprimă:
Sintagmă:

6.7.2. *Secțiunea bibliografică*

Fig. 5 prezintă entitățile (i.e. subclasele clasei 'entita bibliografică, cu alte cuvinte este o secțiune de metadate. Secțiun deoarece o bună parte din multitudinea de date bibliografice sunt ajutorul relațiilor. Clasa esențială este 'ediție'; cea care con bibliografică a unei ediții citate.

Entitatea 'lucrare' consemnează metadatele specifice unei textuale, în cazul nostru), i.e. "abstractizează" ceea ce au în comun t lucrări. Utilitatea ei imediată este colocarea tuturor manifestăril indiferent de limbă sau ediție. O subclasă importantă a clasei 'lucrar 'serial'. Aici se consemnează şi periodicele, adică entitățile ce grup clasei 'NumărPeriodic', cu alte cuvinte publicațiile-gazdă ale arti asupra acestor clase şi a relațiilor între ele depăşeşte cadrul acestui a

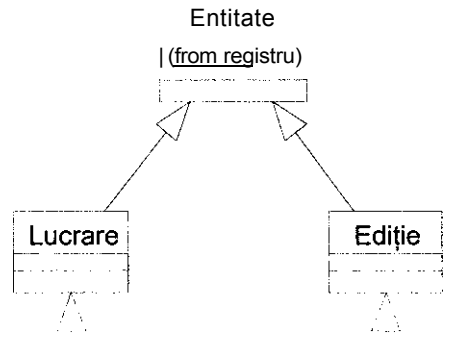


Figura 5 - Secțiunea bibliografică

B.2. Secțiunea administrativă

în fig. 6 se prezintă clasele de natură administrativă şi relațiile esențiale între ele. Rolul acestor clase este de a consemna modificările survenite în baza de date, în succesiunea lor. în acest fel se poate urmări geneza înregistrărilor şi se pot identifica responsabilitățile. în plus, deoarece se prevede şi stocarea datelor modificate, se creează premizele revenirii la stări anterioare ale bazei de date. în instanțele clasei 'intervenție' se consemnează fiecare modificare operată asupra unei înregistrări. Fiecare asemenea instanță este asociată — prin intermediul instanțelor clasei 'contribuție' — cu agentul (i.e. operatorul) care a produs-o. în plus o intervenție este asociată şi cu sursele ei documentare. Se observă cum clasa 'referință' poate avea ca instanțe atât referințe bibliografice (citând o ediție), cât şi referințe personale (citând o comunicare personală).

Clasa 'înregistrareArhivă' este foarte importantă, instanțele ei fiind chiar versiunile "desuete" (i.e. cele dinainte de modificări) ale atributelor înregistrărilor.

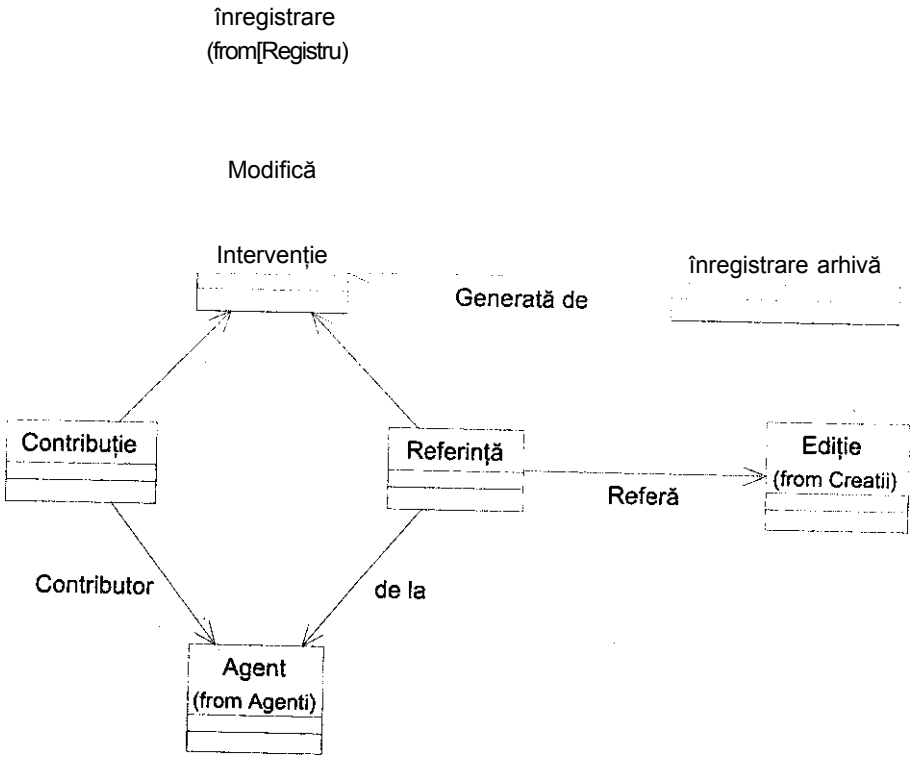


Figura 6 - Secțiunea administrativă

C. Remarci finale

Modelul prezentat pare suficient de flexibil pentru a satisface cerințele funcționale atât ale unei baze de date terminologice, cât şi a uneia lexicografice (mai ales datorită distincției între termeni şi expresiile lor lingvistice). El este şi suficient de abstract pentru ca schema unei baze de date ce l-ar folosi ca fundament să fie relativ comodă la implementare.

TermRom are în curs un proiect de elaborare a unei astfel de baze de date terminologice. După finalizarea acesteia, este de aşteptat un proces traumatic de convertire a bazei de date curente. Sporul de funcționalitate obținut va compensa însă efortul.

D. Referințe

[1] Matei, Dan. *Banca de date terminologice a TermRom şi problemele ei neologice,* în *Limbaj şi Tehnologie / Dan Tufiş - editor.* - Bucureşti: Editura Academiei Române, 1996'

[2] *ISO/CD 16642:1999, Computer applications in terminology - Metamodel for representing terminologica! data collections*

[3] *ISO 12200:1999, Computer applications in terminology - Machine-readable terminology interchange format (MARTIF) - Negotiate interchange*

[4] *ISO 12620:1999, Computer applications in terminology - Data categories*

[5] Chețan, Octavian, Radu Sommer. *Dicționar de filozofie / Coordonare ştiințifică Octavian Chețan, Radu Sommer.* — Bucureşti: Editura Politică, 1978

Probleme de reprezentare a date terminografice într-o bază de date rel

Sorin GHEȚARII
TERMOROM, Str. Meşterul Manole nr. 3
gsorin@fx.ro

Oriunde şi oricând se creează, comunică, înregistrea stochează, transformă sau refoloseşte informaţie sau cunoştinţ este implicată într-un fel sau altul şi terminologia. Comunicare domeniu a devenit un discurs specializat cu texte de specialita nenumărate forme. Atunci când se defineşte terminologia ca o mu de concepte şi denumirile lor într-un anumit domeniu, ea poate fi infrastructura cunoaşterii de specialitate. Scrierea textelor tehnice tehnică devin astfel imposibile fără o utilizare corectă a unor resur Deoarece producerea textelor tehnice implică frecvent m terminologiile multilingve de înaltă calitate au devenit bunuri mu găsit pe înfloritoarea piaţă a industriilor limbajelor şi cunoaşterii.

Există numeroase baze de date terminologice disponibile on-line sau pe CD-ROM (TERMIUM, EURODICAUTOM), pe di unor dicţionare electronice sau ca baze de date personale realiz de ingineri, specialişti în calculatoare, chimişti care lucrează traducătorii, autori de texte tehnice. Aceste baze de date sunt utiliz

- traducere asistată de calculator;
- scrierea de texte tehnice şi ştiinţifice asistată de calcu
- sisteme informatice (administrarea componentelor etc
- cercetări terminologice în lingvistică, filozofia şti tehnologiei etc.

Pentru asemenea obiective au fost dezvoltate aplic (programe de management al bazelor de date terminologice), une piaţa terminologică internaţională, altele ca prototipuri în cadrul cercetare.academică.

MARTIF este formatul standardizat pentru managem terminologice. Posibilitatea organizării terminologiei în baze de da diferite face nerealistă presupunerea ca s-ar putea cădea de ac anumit format de bază de date relaţională, aşa cum este SQL, c pentru schimburile terminologice. De aceea s-a mers pe linia

format la dispoziția publică fără obligațiii materiale şi care să fie independent de platforma de lucru. Rezultatul este MARTIF (Machine-Readable Terminology Interchange Format cunoscut şi ca ISO 12200. In ISO 12620 sunt descrise 150 de categorii de date, un număr imens care nu urmăreşte decât să le arate pe cele posibile şi modul în care acestea pot fi structurate. Categoriile MARTIF sunt împărțite în 10 secțiuni grupate în 4 clase. Acestea sunt:

- termen: cuprinde categoria de date termen (1);
- informație în legătură cu termenii: conține informația legată de termeni (2) şi informația privind gradul de echivalență;
- informație descriptivă: relație cu domeniul (4), descrierea conceptului (5), relații între concepte (6), categorii de date care leagă un concept de poziția sa în sistemul de concepte (7), note (8);
- informație administrativă: categorii de date care leagă un concept de un element al unui tezaur sau de o altă formă de documentare (9), categorii de date care cuprind informații administrative.

Un avantaj major al faptului că MARTIF este scris folosind cod SGML este acela că, deşi se poate aprecia că lectura codului nu este facilă, ea este totuşi posibilă ca urmare a faptului că nu face apel decât la caracterele ASCII. Un alt avantaj al sistemului MARTIF este acela că el acceptă referințe către alte documente chiar din interiorul documentului. Inițial MARTIF presupune că înainte de implementarea produselor software pentru importul sau exportul datelor programatorii sunt obligați să examineze sursele implicate. Pentru a asigura un acces aşa numit "orb" care să permită oricui să transfere baze de date terminologice din orice sistem spre sau dinspre MARTIF este necesară o standardizare suplimentară a categoriilor de date, domeniilor specifice etc.

Tabela ce urmează enumera acea parte a "elementelor" MARTIF care sunt de cea mai mare importanță pentru realizarea unei resurse terminologice Multilingve.

<termEntry>	**îvnrimTf** [illegible] ate terminologice pentru un concept " **S ^ n S S S ^ ? ^** [illegible] administrative codate lor sau, în cazul unei abordări bilingve sau multilingve, două sau mai mu te datele descriptive şi administrative asociate lor Atributele includ: type, care clasifică setul de date terminologice conform categoriile de date specificate de ISO 12620
<langSet>	**Limba,** in caarul unui element <termEntry> va fi folosit pentru a grupa **mai** multe <tig> şi <ntig> asociate unei singure limbi *S S S S T* * [illegible] - u **' S** care

1<tig>	Grup de informații terminologice; în cadrul unui elem <termEntry>, va conține elemente de informații asoci singur termen, fiecare dintre acestea funcționând la a cu alte cuvinte nu este permisă imbricarea între elem subordonate unui <tig>. Prezența atributului lang este obligatorie, în afara ca el este moştenit.
kntig>	Grup încuibat de informații terminologice; va fi folosi unui element <termEntry> dacă anumite elemente i sunt asociate mai curând cu elemente interne, decâ <tig>. Următoarele elemente vor fi folosite în cadrul <ntig> găzdui alte date terminologice: <termGrp>, <termN <descripGrp> şi <adminGrp>. Prezența atributului lang este obligatorie, în afara ca el este moştenit.
<term>	Va conține un termen format dintr-un singur cuvânt multe cuvinte, sau o desemnare simbolică privită ca
[<termGrp>	"Va conține un element <term> şi posibil, cel puțin m element încuibat în plus față de termen.
<termNote>	" Va conține informații legate de termen. Atributele includ: type care clasifică <termNote> conform categoriilo ;în ISO 12200.
[<termNoteGrp>	Va conține un element <termNote> şi posibil cel pu încuibat în plus față de informația legată de termen pentru a găzdui un nivel suplimentar de imbricare î elementului <termGrp>
<descrip>	Va conține informații descriptive precum definiția, c explicații descriind concepte şi termeni. Atributele includ: type, care clasifică <descrip> potrivit categoriilor d în ISO 12200.
<descripGrp>	
|<admin>	" Va conține date administrative. Atributele includ: type care clasifică <admin> în funcție de categori ite în ISO 12200.
<adminGrp>	

<date>
Va conține o singură dată de formatul YYYY-MM-DD, cu opțiunea notării dată-timp YYYY-MM-DDhh:mm:ss.
Atributele includ:
type, care clasifică <date> după categoriile specificate în ISO 12200.

<note>
Va conține o notă sau o adnotare drept comentariu legat fie de un întreg <termEntry>, un întreg <tig> sau <ntig> ori de unul din elementele <...Grp>.

<descripNote>
Va fi folosit în cazul informațiilor de tipul <note> folosite în cadrul <descripGrp> când conținutul notei este legat de o listă de opțiuni.

<adminNote>
Va fi folosit în cazul informațiilor de tipul <note> folosite în cadrul <adminGrp> când conținutul notei este legat de o listă de opțiuni

<ptr>[1]
Va consta dintr-un indicator către o altă locație din documentul curent.
Atributele includ:
type, care clasifică <ptr> conform Anexei A, A.12
target, care precizează destinația referirii, ca unul sau mai mulți identificatori SGML.

<ref>
Va defini o referire către o altă locație din documentul curent, în termeni de unul sau mai multe elemente identificabile. <ref> este asociat cu text suplimentar drept conținut al elementului, deci constă dintr-o etichetă-start cu o țintă integrată, urmată de textul asociat şi închisă de o etichetă-sfârşit.
Atributele includ:
type, care clasifică <ref> conform Anexei A.
target, care precizează destinația referirii ca unul sau mai mulți identificatori SGML.

<xref>[2]
Va defini o referință la un grafic, ilustrație, figură, tabel sau alt document extern sau fişier folosind o notație indicativă extinsă ca valoare a atributului țintă a <xref>, de ex. <xref target='documentIdentifier'>, unde valoarea 'documentIdentifier' este un cod de identificare pentru documentul țintă. Utilizatorul va documenta notația indicativă extinsă care este folosită incluzând un comentariu adecvat în elementul <encodingDesc> ale header DTD.
Atributele includ:
type, care clasifică <xref> conform Anexei A.
target, care precizează destinația referirii ca unul sau mai mulți identificatori SGML

[1] Mmcăt ZTJdoar "To I t S i Z IT ... ar drept conținut ... amantului, <xref> sunt toate c o n ^ ... integrată – Elementele <ptr>, <ref> şi

* Z ~ u r ^ ... t.ntite de <™f> **>»*> să fi accesibile Sllui-țintă pentru

<hi>[3]
Va fi folosit pentru a marca un cuvânt sau o frază ca[...] grafic în contrast cu textul înconjurător.
Atributele includ:
type, care clasifică <ref> conform Anexei A.
target, care precizează destinația referirii ca unul sa[...] identificatori SGML

<foreign>
[...] sau o fraza ca apărând a[...] cea a textului înconjurător.
Atributele includ:
lang, care identifică limba cuvântului sau frazei ma[...]

<refObjectList>
Va fi folosit în back-matter şi va conține unul sau m[...] obiecte back-matter, mai ales resurse comune ca: [...] bibliografice, date de responsabilitate, identificatori [...] namespace (URL-uri şi FPI-uri), material textual la[...] referiri dese, liste de locații geografice, fişiere exte[...] asemenea.
Atributele includ:
type, care clasifică <refObjectList> după categoriil[...] specificate în ISO 12620 Anexa A, A.11.4.1.

<refObject>[4]
Va conține o dată constând în general dintr-o resu[...] ca: date bibliografice, date de responsabilitate, ide[...] namespace (URL-uri şi FPI-uri), material textual la[...] referiri dese, liste de locații geografice, fişiere exte[...] asemenea. Datele bibliografice ar trebui să rezide[...] sau într-un document extern (caz în care se va fa[...] datele bibliografice din back matter folosind elem[...]
Atributele includ:
type, care clasifică <refObject> după categoriile d[...] specificate în ISO 12620 Anexa A, A.11.4.2. Dac[...] altfel, tipul <refObject> este moştenit de la <refOb[...] respectiv.

Notă - în managementul terminologiei o utilizare frecventă a <hi> se fac[...] termeni necesari, adică termeni folosiți într-o definiție, notă sau alt m[...] sunt definiți altundeva în resursa terminologică. Vezi de asemenea Anex[...]

[4] *Notă - Unele documente terminologice cuprind date bibliografice c[...] nediferențiat drept conținut al categoriei de date sursă (vezi ISO 126[...] Această practică încurajează redundanța şi efortul mărit pentru îngrijir[...] Informații ar trebui convertite în obiecte back matter (informații biblio[...] posibil.*

<itemSet>	Va fi folosit în back matter şi va conține unul sau mai multe obiecte individuale care în mod tradițional sunt grupate împreună, de ex. obiectele numele autorului si prenumele autorului vor fi grupate împreună într-un <itemSet> de tip=autor Atributele includ: type care clasifică <itemSet> în principal conform categoriilor de date listate în ISO 12620 Anexa B. Totuşi acest Standard Internațional nu specifică întregul spectru al categoriilor de date care pot fi folosite cu <itemSet>
<item>	Va conține un exemplu individual de informație back matter Atributele includ: type, care clasifică <itemSet> în principal conform categoriilor de date listate în ISO 12620 Anexa B pentru informații bibliografice Totuşi acest Standard Internațional nu specifică întregul spectru al categoriilor de date care pot fi folosite cu <item>
<itemGrp>	Va conține unul sau mai multe <item> împreună cu <otr> <ref> sau <note>. Atributele includ: type, care clasifică <item> în principal conform categoriilor de date listate în ISO 12620 Anexa B pentru informații bibliografice Totuşi acest Standard Internațional nu specifică întregul spectru al categoriilor de date care pot fi folosite cu <itemSet>

Din acest tabel au mai fost eliminate elementele (aproape la fel de numeroase) specifice informațiilor bibliografice. Instanțierea elementelor enumerat mai sus se face prin intermediul "categoriilor de date" standardizate de ISO 12620. Numărul acestora este de aproximativ 200. în cea mai amplă resursă terminologică (EURODICAUTOM) sunt în prezent prezente mai puțin de 20 astfel de categorii de date.

Uniunea Eropeană în activitatea sa este unul dintre utilizatorii majori ai procedurilor de translatare a textelor şi terminologiei. Aceasta se datoreşte parțial faptului că legislația sa este direct aplicabilă în statele membre şi de aceea ea trebuie să fie disponibilă în toate limbile de lucru oficiale. Ca rezultat, traducătorii Comisiei Europene produc mai mult de 1 milion de pagini pe an şi au de-a face cu cel puțin 6-7 milioane de termeni (în medie sunt 8 sau 9 termeni care ridică probleme pe fiecare pagină).

Unitatea pentru Terminologie a Comisiei Europene este destinată asigurării suportului lingvistic pentru toate limbile oficiale ale Uniunii Europene. Au fost elaborate glosare de specialitate, multe dintre ele în nouă limbi. Domeniile acoperite sunt tratatele importante cum ar fi cele de la Maastricht şi Roma, cele economice şi administrative (Taxa pe Valoarea Adăugată, buget) dar şi unele legate de subiectele centrale sau puternic inovatoare ale ştiinței şi tehnologiei (fizica plasmei, biotehnologie, minerit). Deosebit de rolul lor de resurse

terminologice şi de surse terminologice pentru domeniile de inovare, acest glosare documentează ceea ce se numeşte "Eurolect", adică frazele şi cuvinte care îşi au origina în cadrul Uniunii Europene şi pentru care nu există echivalen naționale.

Monitorizând toate modificările apărute ca urmare a unei evolu permanente a bazei de date EURODICAUTOM am constatat că, recent, a avut lo schimbarea suportului hardware şi odată cu aceasta pot fi observate următoarele

- Indicarea mult mai frecventă a referinței la documentul sursă termenului;
- Indicarea frecventă a referinței la documentul sursă al definiț acestuia;
- Indicarea documentului sursă şi pentru sinonime şi abrevieri;
- Utilizarea mai frecventă a notelor pentru adăugarea unor informa suplimentare asupra termenilor, acestea putând fi grupate astfel:
  - {NTE} explicații şi informații generale asupra termenilor;
  - {TXT} contextul (de cele mai multe ori un exemplu de utilizare termenului respectiv);
  - {GRM} informații gramaticale (gen, număr);
  - {USG} indicarea mediului în care este utilizat termenul: *"techn jargon";*
  - {REG} notă asupra unor utilizări locale speciale sau asu regionalismelor;
  - {DOM} indicarea unui domeniu sau subdomeniu care comp mentează clasificarea obişnuită folosită anterior şi care a răm încă prezentă.

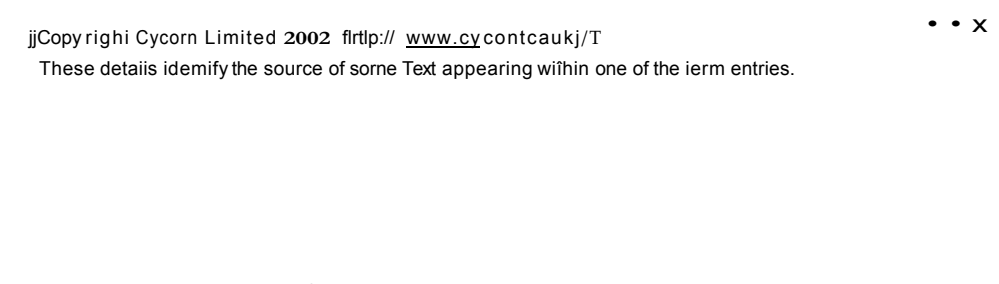
De asemenea se prevede ca în cel mai scurt timp să fie implement următoarele:

- afişarea tuturor caracterelor şi diacriticelor (ca şi a informa nelingvistice, dacă se cere);
- îmbunătățirea sistemului de clasificare a domeniilor;
- introducerea link-urilor interne şi externe.

Modelele de date terminologice orientate în exclusivitate către terminolo au avantajul de a fi relativ intuitive pentru terminolog. Transcrierea directă elementelor şi relațiilor dintre acestea într-o bază de date este din ce în ce dificilă şi mai riscantă.

Există încercări meritorii de realizare a unor interfețe "cuprinzătoare" pentru consultarea resurselor terminologice. Exemplele ' următoare sunt edificatoare în acest sens.

Primul exemplu ar putea provoca comentarii legate de complexitatea reală a înregistrării referințelor bibliografice cele mai obişnuite.



Al doilea, ne determină să luăm în considerare următoarele:

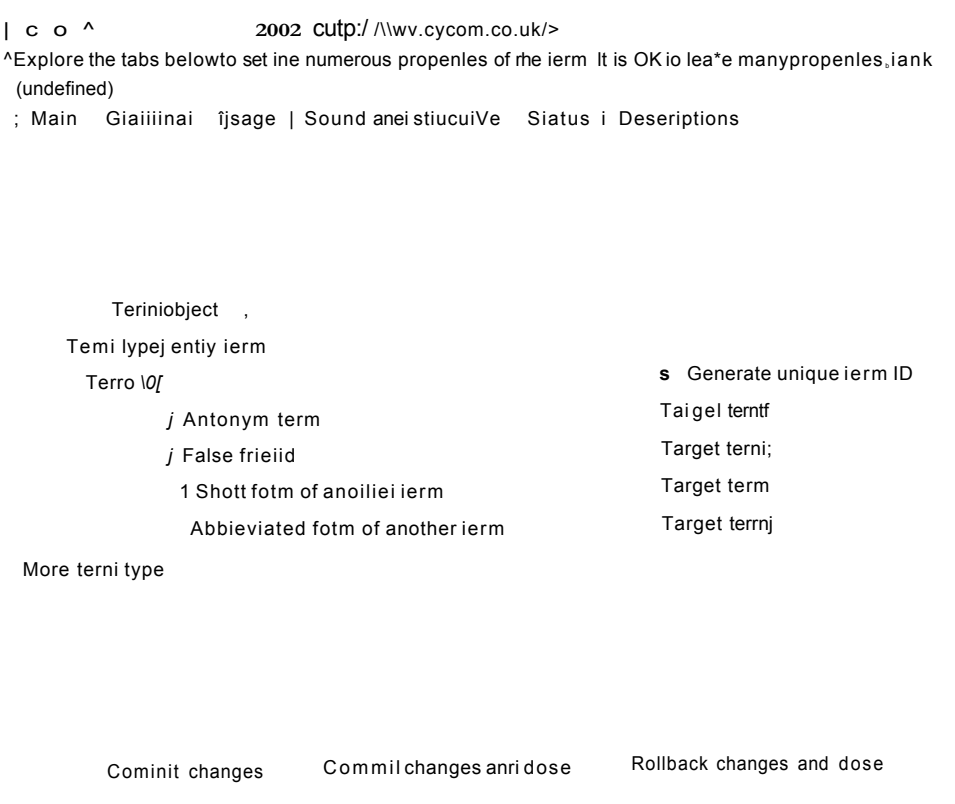
La nivelul Uniunii Europene numărul limbilor pentru care este necesar suport terminologic este atât de mare (şi sperăm încă în creştere) încât nu mai este posibilă multiplicarea tabelelor bazelor de date potrivit numărului de limbi de lucru. Din fericire, "balizarea" documentelor permite identificarea şi prelucrarea corect dependentă de limba în care au fost concepute acestea. Se vine astfel în sprijinul "globalizării" aplicațiilor informatice care sunt suport al resurselor terminologice multilingve dând posibilitatea acceptării, prelucrării şi prezentării

numeroaselor scrisuri, formate de date şi limbi existente. în acelaşi timp trebuie adaptată şi interfața utilizator potrivit locului şi culturii căreia îi aparține acesta printr-un proces nu mai puțin important de "localizare"

Multă vreme, prelucrarea automată a datelor a fost considerată satisfă cător realizabilă prin utilizarea setului ASCII de caractere. în prezent este însă absolut necesar ca:

- Utilizatorul calculatorului să poată tasta caractere şi simboluri (vest europene, est-europene, greceşti şi cirilice, cel puțin) folosind claviatură standard.
- Aplicația să prelucreze şi să afişeze sau să imprime şiruri de caracter formatate corect folosind seturi de caractere specifice fiecărei limbi.

Aceste cerințe pot fi realizate prin valorificarea calităților standardulu Unicode de codificare prin utilizarea unor coduri de 16 biți pentru reprezentare tuturor caracterelor pentu calculatoarele moderne care includ simbolurile tehnice s semnele speciale necesare imprimării textelor.



Cu alte cuvinte la nivelul seturilor de semne necesare unei resurse terminologice multilingve se poate conta pe serviciile standardului Unicode şi pe cele ale oricărei baze de date relaționale care acceptă Unicode.

Pentru indicarea formatelor de prezentare (fonte, punere în pagină, seturi de caractere) şi a limbii utilizate se face apel la balizare astfel încât la nivelul câmpului vom găsi şiruri de caractere Unicode balizate.

înscrierea datelor terminologice este facilitată de înscrierea lor în "categorii de date" bine definite (vezi ISO 12620). Dar numărul mare al acestor categorii şi mai ales incidența ridicată a aparițiilor neprevăzute dinainte a unora noi face imposibilă alocarea unui câmp de date fiecărei categorii de date. Aceeaşi observație poate fi făcută şi asupra relațiilor dintre diferitele categorii de date care reflectă direct relațiile dintre elementele MARTIF. O soluție este o abstractizare suplimentară a datelor terminologice după încadrarea lor succesivă în şiruri de caractere balizate, categorii de date, elemente MARTIF.

în centrul modelului de date se află un set de 13 entități (atomi):

[Entitate	Descriere
data category	o anumită clasă de informații terminologice (de exemplu: term, part of speech)
jdata category name	un nume agreat de utilizator (user-friendly), dependent de limbă, al unei anumite categorii de date (de exemplu, în română, "termen" pentru term)
data category index type	o strategie de indexare corespunzătoare unei anumite categorii de date (ISO 12620) (de exemplu: nu se indexează, se indexează ca valoare unică, se indexează cuvânt cu cuvânt)
	o anumită limbă, care dispune de o schema de codare uniformă care utilizează un singur set de caractere (de jexemplu: French, German, Italian),
	o combinație unică de caractere care poate fi utilizată pentru reprezentarea unei singure sau mai multor limbi (de
picklist	exemplu: ISO 8879-1. ISO 8859-2)
	o mulțime de valori posibile ale unor date terminologice aparținând unei anumite categorii de date (ISO 12620) (de exemplu, pentru categoria "parte de vorbire": noun, verb, adjective)
	o dată terminologică unică
	o dată (time stamp) care constituie valoarea unui element
	un număr care constituie valoarea unui element
	un membru al unei liste care reprezintă valoarea unui element
	;ir de caractere care constituie valoarea unui element
	în şir de caractere r.pr^ r o n ^ - : - ^ x—

Primele 6 "articole" sunt "meta-entități"; ele sunt create şi tabelel corespunzătoare sunt completate cu informații înainte de încărcarea oricărei dat terminologice în baza de date. Prin completarea acestor table se conturează şi s activează chiar modelul de date al bazei de date terminologice. Cu alte cuvinte ansamblul "meta-tabelelor" defineşte structura care impune condiții şi unifică datel terminologice de nivel molecular. Ele pot fi considerate atomi catalizatori a reacțiilor necesare combinării altor atomi în interacțiuni moleculare.

Celelalte 7 entități se încarcă direct prin proceduri de introducere a datelo sau prin import şi cuprind datele terminologice vizibile pentru utilizatorul bazei d date. Informațiile conținute de aceste entități pot fi validate la nivel molecula folosind interogări SQL standard. Majoritatea interogărilor formulate de utilizator bazei de date se concentrează aproape în întregime asupra informațiilor încărcat în aceste entități.

Elementul central al aplicației pentru întreținerea unei astfel de baze d date este componenta de tip *parser* pentru crearea, validarea şi prelucrare documentelor MARTIF în particular (fără a ignora documentele SGML, HTML XML). în mod obişnuit un *parser* este un modul software care examinează u document SGML prin confruntarea acestuia cu DTD-ul corespunzător. Rezultatu acestei examinări este de cele mai multe ori simplu: 'da' în situația în car documentul reprezintă o instanțiere validă a DTD-ului şi 'nu' în cazul contrar. D cele mai multe ori *parser-u*\ este capabil să 'normalizeze' documentul valida (aducându-l la o 'formă canonică') astfel încât facilitează formatarea, editarea s încărcarea documentului în baza de date.

Alături de *parser* şi legat de acesta se află un *editor structurat.* Pornind d la DTD acesta propune utilizatorului pas cu pas opțiunile de compunere, sa modificare a unui document în conformitate cu definiția tipului corespunzăto documentului. în cazul în care obiectivul este compunerea unui document SGML poate asigura completarea teg-urilor necesare.

De cele mai multe ori sistemele de management al bazelor de dat orientate spre text folosesc *fişiere inversate de indexare* a conținutului acestor pentru regăsirea informațiilor. Căutarea poate urmări apariția unui anume cuvân sau a unui model oarecare într-un document sau în o parte a acestuia Identificarea subdiviziunilor documentului se poate face folosind tocmai tag-urile c acesta este marcat, respectiv modul în care acestea au fost transcrise în relații dintre tabelele bazei de date.

în fine, o componentă deosebit de importantă este aceea care realizeaz funcțiile de *import-export a*\e datelor terminologice spre şi dinspre baza de date.

Terminologia calității

Realizarea unor resurse terminologice multilingve este de mai multă vreme în centrul preocupărilor Asociației Române pentru Terminologie (TERMROM). începând de anul trecut pe lista temelor având aceeaşi orientare se înscrie proiectul "Terminologie armonizată cu prevederile EURODICAUTOM în domeniul calitate şi standardizare". Proiectul a fost inițiat de Ministerul Educației şi Cercetării şi este finanțat în cadrul Programului CALIST.

Obiectivele principale ale acestui subprogram sunt:

- Asigurarea flexibilității necesare pentru a răspunde operativ la cerințele concrete de rezolvare a unor teme de cercetare care decurg din prioritățile stabilite prin strategiile guvernamentale adoptate pe domenii specifice, în procesul integrării României în U.E.
- Asigurarea condițiilor de dezvoltare şi armonizare a sistemului de standarde naționale în conformitate cu cerințele organismelor de standardizare europene şi internaționale;
- Asigurarea unei baze terminologice ştiințifice pentru elaborarea standardelor de calitate româneşti, precum şi în ceea ce priveşte condițiile de aplicabilitate a prevederilor standardelor internaționale şi europene adaptate ca standarde româneşti;
- Clarificarea condițiilor pe care trebuie să le îndeplinească produsele româneşti în vederea pătrunderii lor pe piața unică a Uniunii Europene şi produsele introduse în România.

Pentru realizarea obiectivelor proiectului au fost prevăzute următoarele activități:

- întocmirea unui Proiect Terminologic pentru definirea şi înregistrarea terminologiei domeniilor calitate şi standardizare utilizate în documentele oficiale ale Uniunii Europene, conform prevederilor EURODICAUTOM şi standardelor internaționale;
- Extragerea, traducerea şi structurarea terminologiei domeniilor calitate şi standardizare;
- Proiectarea, programarea şi implementarea unei Baze de date conform Proiectului Terminologic capabilă să gestioneze toate domeniile EURODICAUTOM;
- înregistrarea în baza de date a terminologiei domeniilor calitate şi standardizare;
- Elaborarea unei aplicații informatice de administrare a bazei de date terminologice şi de transfer de date terminologice conform formatului standard ISO pentru lucrul în rețea;
- Realizarea unui site web pentru promovarea Bazei de date terminologice şi punerea acesteia la dispoziția publicului.

A fost avizat Proiectul Terminologic, au fost stabilite cerințele pe care să le satisfacă suportul informatic, s-a constituit un fond de termeni specifici extraşi din EURODICAUTOM şi din Tezaurul rațional al CEI şi au fost demarate activitățile pentru realizarea unei baze de date relaționale EUROCAST pentru înregistrarea acestora.

Bibliografie

[1] **ISO 639:1988**
Code for the representation of names of languages

[2] **ISO 639-2:1998**
Code for the representation of names of languages - Part 2: Alpha-3 code

[3] **ISO 704:2000**
Terminology work - Principles and methods

[4] **ISO 860:1996**
Terminology work - Harmonization of concepts and terms

[5] **ISO 1087-1:2000**
Terminology work - Vocabulary - Part 1: Theory and application

[6] **ISO 1087-2:2000**
Terminology work - Vocabulary - Part 2: Computer applications

[7] **ISO 1951:1997**
Lexicographical symbols particularly for use in classified defining vocabularies

[8] **ISO 6156:1987**
Magnetic tape exchange format for terminological/lexicographical records (MATER)

[9] **ISO 10241:1992**
Preparation and layout of international terminology standards

[10] **ISO 12199:2000(E)**
Alphabetical ordering of multilingual terminological and lexicographical data represented in the Latin alphabet

[11] **IS012200:1999**
Computer applications in terminology - Machine-readable terminology interchange format (MARTIF) - Negotiated interchange

[12] **ISO/TR 12618:1994**
Computer aids in terminology - Creation and use of terminological databases and text corpora

[13] **IS012620:1999**
Computer applications in terminology - Data categories

[14] **IS015188:2001**
Project management guidelines for terminology standardization

SECȚIUNEA il

TEHNOLOGII ALE LIMBAJULUI SCRIS

Ro-Balkanet - ontologie lexicalizată, în context multilingv, pentru limba română

Dan TUFIŞ, Institutul de Cercetări pentru Inteligența Artificială,
Academia Română, Bucureşti
Calea 13 Septembrie nr. 13, 74311, sector 5

tufis@racai.ro

Dan CRISTEA, Facultatea de Informatică, Universitatea A.I.Cuza, Iaşi
Str. General Berthelot, nr. 16

dcristea@infoiasi.ro

Rezumat

Cerințele creării unei ontologii multilingve de tipul EuroWordNet sunt frecvent contradictorii şi dacă problemele de compatibilitate nu sunt considerate în etapele timpurii ale construcției, o armonizare tardivă se poate dovedi dificilă sau imposibilă. Mai exact, există două probleme majore de compatibilitate care trebuie avute în vedere şi anume: acoperirea conceptuală - în sensul că fiecare lexicon monolingv ar trebui să conțină lexicalizări ale aceluiaşi fond conceptual şi coeziunea interpretativă - în sensul că interpretarea relațiilor folosite în fiecare din ontologiile cuprinse în ontologia multilingvă trebuie să fie identică. în lucrare sunt discutate ambele aspecte şi prezentate soluțiile adoptate în vederea satisfacerii criteriilor de consistență şi coerență multilinguală a wordnet-ului pentru limba română.

1. Limbă, resurse lingvistice şi comunicare electronică

Cercetarea în domeniul tehnologiilor limbajului este un domeniu ce are deja istorie în ştiința calculatoarelor, dar, actualmente, motivațiile sale depăşesc sfera interesului pur ştiințific sau comercial. Păstrarea identității limbilor şi culturilor naționale în cadrul globalizant al societății informaționale şi a cunoaşterii readuce în actualitate avertismentul lui Alain Danzin [1]: ***"în era electronică, este esențial pentru supraviețuirea unei limbi ca ea să fie folosită în sistemele de informare electronică."*** Avansul ştiințific şi tehnologic obținut în cei 10 ani scurşi de la raportul prezentat de Danzin Comisiei Europene a condus la maturizarea unor teorii, tehnologii, metode şi la dezvoltarea altora noi, dar mai ales a permis

definirea unor standarde pentru realizarea unitară a ceea ce generic se numeşte *resurse lingvistice fundamentale* ale unei limbi. Caracterul multilingual al societății cunoaşterii, în care conceptul de "unitate prin diversitate" se referă în primul rând la prezervarea limbilor şi culturilor actuale, a generat o deosebită efervescență, puternic stimulată de organismele internaționale - în primul rând de Comisia Europeană - asupra cercetării în domeniul resurselor multilingve. Metodologic, tehnologia limbajului natural creează o distincție netă între prelucrări şi date, între "maşinăria software de prelucrare a limbajului" numită şi *lingware* şi cunoştințele lingvistice, numite cum arătam *resurse lingvistice,* necesare funcționării acestei maşinării. Dihotomia *lingware - resurse lingvistice,* susținută de standardele de reprezentare şi codificare a cunoştințelor lingvistice permite dezvoltarea independentă a celor două componente ale unui sistem de prelucrare a limbajului. Lingware-ul este independent de limbă şi intră tot mai pregnant în zona ingineriei software. El poate fi dezvoltat de specialişti de oriunde fără ca aceştia să fie preocupați de limba pentru care va fi folosit. Resursele lingvistice însă sunt de competența specialiştilor vorbitori nativi ai limbii respective. în condițiile în care aceste resurse lingvistice sunt realizate în conformitate cu standardele sau practicile internaționale, ele pot fi integrate în sistemele de comunicare electronică, nu doar pentru prelucrare monolingvă ci mai ales pentru prelucrări multilingve. Beneficiile alinierii la standardele internaționale în realizarea resurselor lingvistice sunt enorme, şi putem considera un exemplu foarte simplu. Să presupunem că suntem interesați de un anumit subiect şi, folosind imensul ocean informațional ce este Internet-ul, apelăm la un aşa numit "motor de căutare", un program a cărui funcționalitate asigură identificarea documentelor electronice ce conțin informații potențial relevante pentru subiectul nostru de interes. Acest gen de serviciu informațional este asigurat de "motoare de căutare" precum Google, Altavista, Excite şi multe altele. Documentele interesante din punctul nostru de vedere ar putea să fie scrise în limba engleză, franceză, germană, română sau orice altă limbă. Dar pentru a le regăsi pe toate, indiferent în ce limbă am formulat cererea noastră de regăsire, motorului general de căutare îi sunt necesare resursele lingvistice specifice limbilor în care documentele ar putea exista. Dacă aceste resurse lingvistice există pentru engleză, franceză, germană, italiană etc. şi ele sunt reprezentate în acelaşi format standardizat, rezultatul cercetării noastre documentare va fi o colecție de documente tratând subiectul de interes în oricare dintre aceste limbi. Un astfel de serviciu, numit regăsire documentară multilingvă este o realitate pentru toate limbile "mari", o calificare ce nu are acoperire în substratul cultural ci doar în ceea ce se numeşte "nivelul de informatizare al limbii". Procesul de informatizare a unei limbi naturale permite potențarea şi diseminarea ei prin mijloacele tehnologice ale societății informaționale.

2. Lexicalizarea abordărilor în tehnologia limbajului şi conceptul "wordnet"

Lexicul este fără îndoială cea mai importantă resursă lingvistică a unei limbi. Marea majoritate a cercetării actuale, atât în lingvistica formală cât mai ales în tehnologia limbajului, plasează componenta lexicală în centrul modelelor de limbă, sub influența a ceea ce a fost numită abordarea *lexicalizaiă* sau *lexicalistă* a studiului limbii. Nu este de mirare, deci, enormul interes pentru dezvoltarea de resurse lexicale multilingve. Studiul computațional al dicționarelor electronice natura informației ce trebuie inclusă în ele şi tipul de prelucrări pe care le poate facilita o anumită structurare a unui mare voium lexical a fost, fără îndoială fundamental influențat de proiectul WordNet, lansat în urmă cu mai mult de 25 de ani la Universitatea din Princeton sub conducerea reputatului psiholingvist George Miller. WordNet, resursă publică, este o uriaşă rețea semantică lexicală în care peste 100.000 de *înțelesuri* lexicalizate în limba engleză prin mai mult de 130.000 de cuvinte sunt asociate între ele prin relații semantice şi/sau lexicale [2]. Fondul lexical este distribuit în 4 rețele semantice corespunzând categoriilor gramaticale deschise: substantive, verbe, adjective şi adverbe. Noțiunea de *înțeles (meaning* este în WordNet echivalată cu cea de concept şi este reprezentată printr-o serie sinonimică în care fiecare cuvânt al seriei are asociat un număr ce identifică sensul în care cuvântul respectiv are înțelesul asociat conceptului. Seria sinonimică ce identifică un înțeles se numeşte *sinset.* Relațiile existente între sinseturi sunt de diferite tipuri, depinzând de categoria gramaticală a cuvintelor ce alcătuiesc un anumit sinset (antonimie/sinonimie, hiponimie/hiperonimie, holonimie/meronimie troponimie etc). Influența proiectului WordNet a fost enormă în domeniul tehnologiei limbajului (exprimată poate şi prin faptul că acum, în limbajul tehnic ce puțin, cuvintele "wordnet" şi "synset" au devenit substantive comune, importate prin calchiere în mai toate limbile), iar beneficiile acestui concept sunt atât de evidente încât Comisia Europeană, între 1996 şi 1998, a finanțat un proiect similar de mai anvergură numit EuroWordNet [3]. Acest proiect, extrem de ambițios şi-a propus nu numai realizarea concertată de wordneturi monolingve pentru limbile europene de circulație internațională (engleză, franceză, germană, italiană, olandeză spaniolă) dar a introdus o cerință fundamental nouă, anume corelarea multilingvă a celor 6 rețele semantice lexicale, astfel încât dintr-un sinset al unei limbi să se poată ajunge în echivalentul de traducere al oricăreia dintre celelalte 5 limbi. Față de relațiile originale din WordNet, EuroWordNet propune un inventar mult mai bogat (90) de relații cum ar fi cele tematice de tip cazual (Agent, Patient, Instrument, Location, Direction) sau cele corelând sensurile derivaților lexicali (XPOS-SYNONYMY: a adora - adorație).

Soluția tehnică pentru corelarea multilingvală a rețelelor semantice monolingve a fost definirea unui index interlingual (ILI), independent de limbă conținând reprezentări conceptuale ale înțelesurilor lexicalizabile în limbi

proiectului. Fiecare înțeles din oricare din limbile reprezentate în rețeaua semantică multilingvă este pus în corespondență, în general, cu un singur concept al indexului interlingual. Aceste corespondențe se realizează prin intermediul a 20 de tipuri distincte de relații binare. Sinseturile (seriile sinonimice) din două sau mai multe limbi care sunt puse în corespondență cu acelaşi concept din ILI sunt considerate echivalenți de traducere, natura echivalenței de traducere fiind definită de tipul relațiilor ce definesc corespondența dintre sinseturile respective şi conceptul comun.

Inițial, indexul multilingual a fost constituit ca o mulțime nestructurată a tuturor înțelesurilor lexicalizate în WordNet (cu alte cuvinte în engleză). Ulterior, prin dezvoltarea wordneturilor monolingve, ILI a fost îmbogățit şi cu reprezentări conceptuale cu lexicalizări ce nu se regăsesc în engleză.

O altă inovație a proiectului EuroWordNet a fost adoptarea unei mulțimi de primitive semantice, independente de limbaj, în termenii cărora aşa-numitele *concepte de bază* din ILI au fost asociate cu descrieri *ontologice.* Prin importul acestor descrieri la nivelul lexicalizărilor prin echivalenți de traducere (şi, prin moştenire, la hiponimii acestora) în fiecare dintre wordneturile monolingve, în EuroWordNet se poate vorbi de o ontologie lexicală multilingvă. O prezentare în detaliu a proiectului EuroWordNet se poate găsi în [4].

După 3 ani, proiectul EuroWordNet inițial a fost extins pentru o perioadă de încă doi ani (EuroWordNet II) şi a încorporat încă 4 limbi: bască, catalană, cehă şi estoniană. Proiectul EuroWordNet II s-a încheiat în anul 2000 cu realizarea unor nuclee a căror extensie a rămas în exercițiul financiar al autorităților naționale.

3. Limba română în contextul proiectului BALKANET, extensie a EuroWordNet

In septembrie 2001 a fost lansat proiectul european BALKANET (IST - 2000 - 29388), o continuare firească a proiectului EuroWordNet II care aduce alături de cele 10 limbi europene alte 5 limbi din zona balcanică: bulgară, greacă, română, sârbo-croată, turcă [5]. Ca şi în EuroWordNet, ontologiile lexicale monolingve sunt corelate printr-o mulțime de concepte interlinguale, corespondențele fiind stabilite cu ajutorul unor relații de echivalență complexe *(eq-synonymy, eq-near-synonymy, eq-has-hyperonym, eq-has-hypernym etc).*

Reprezentanții din România în acest proiect, care va dura trei ani, sunt Institutul Academiei Române de Cercetări pentru Inteligență Artificială din Bucureşti (coordonator Dan Tufiş) şi Facultatea de Informatică a Universității A.I.Cuza din Iaşi (coordonator Dan Cristea) şi în realizarea obiectivelor proiectului sunt implicați numeroşi specialişti, atât informaticieni cât şi lingvişti. Desigur, participarea românească în acest proiect şi angajarea față de obiectivele

proiectului nu s-au bazat numai pe entuziasm ci pe activități şi rezultate anterioare importante, pe *surse lingvistice* primare [6] de referință ale limbii române, implementate ca *resurse lingvistice* [6] în format standardizat şi pe o multitudine de programe de prelucrare dezvoltate de-a lungul a mulți ani de cercetare, în cea mai mare parte prin finanțare internațională.

3.1. Corpusuri

în cadrul proiectelor europene Multext-East şi TELRI [7], [8], [9], [10], [11] a fost creat un corpus paralel în 7 limbi, foarte detaliat adnotat, bazat pe romanul "1984" al lui Orwell şi un alt corpus paralel în 25 de limbi, bazat pe "Republica" lui Platon. Adnotarea folosită inițial a fost conformă cu standardul TEI (http://www.tei-c.org/),dar ulterior, odată cu cristalizarea standardului CES [12], corpusurile au fost re-adnotate (automat) în conformitate cu CES. Acestea sunt două corpusuri relativ mici (câte aproximativ 110.000 cuvinte în fiecare limbă) dar, datorită acurateței proceselor de etichetare şi de aliniere (validate manual), au fost extrem de folositoare pentru diverse aplicații, de la construirea modelelor lingvistice pentru etichetare morfo-sintactică [13], clasificare a documentelor [14], extragere de echivalenți de traducere [15], până la discriminarea automată a sensurilor [16]. Pe lângă corpusurile multilingve s-au construit alte două corpusuri monolingve mult mai mari: un corpus literar bazat pe diverse romane (conținând aproximativ 1.500.000 cuvinte) şi un corpus jurnalistic (conținând peste 100.000.000 cuvinte). Ambele corpusuri au fost segmentate, etichetate şi lematizate automat[1].

3.2. Dicționare explicative: WEB-LEX şi XML-LEX

Principalul dicționar pe care l-am folosit în analiza noastră este Dicționarul Explicativ al Limbii Române [17], referința lexicografică pentru limba română contemporană, dicționar realizat de Institutul de Lingvistică "Iorgu Iordan"[2] al Academiei Române. în urma analizelor statistice de frecvență în corpusurile menționate, au fost selectate şi introduse în format electronic cele mai frecvente 23.000 de cuvinte titlu din DEX. Acest nucleu DEX a fost convertit într-o bază de date lexicală în cadrul proiectului european CONCEDE *(CONortium for Central European Dictionary Encoding)* [11] şi al proiectului prioritar al Academiei WEB-LEX [18]. Ulterior, îmbogățit continuu prin culegere manuală din alte câteva dicționare explicative (DEX'84, DOOM, DLRM), la inițiativa unor tineri entuziaşti atât din țară cât şi din diaspora (vezi de pildă: http://dex.francu.com), WEB-LEX a fost corectat sub aspect sintactic-structural şi codificat într-un format standardizat respectând convențiile lexicografice utilizate de DEX şi, în măsura posibilului, conținutul său textual. Uneori, din considerente legate de consistența structurală

[1] *Multe dintre aceste resurse pot fi găsite pe situl Consorțiului de Informatizare pentru Limba Română (ConsILR) la adresa http://consilr.info.uaic.ro*

[2] *Noua sa denumire este* Institutul de Lingvistică "Iorgu Iordan-Al. Rosetti

s-au operat o serie de modificări asupra conținutului. De asemenea, o serie de erori evidente în sursa primară au fost corectate de specialişti avizați. Deşi mai bogat (în prezent WEB-LEX conține aproape 70.000 de intrări, față de cele circa 56.000 de intrări din DEX'96), influența DEX a fost fundamentală în dezvoltarea WEB-LEX. Pe de altă parte, eventualele critici asupra conținutului, acolo unde ne-am despărțit de DEX, în nici un caz nu trebuie puse în seama Institutul de Lingvistică "Iorgu Iordan-Al. Rosetti" ci a noastră. Din acest motiv, preferăm să ne referim la WEB-LEX ca la un dicționar ***de tip*** DEX şi nu ca variantă computațională a DEX-ului.

Codificarea conținutului WEB-LEX s-a realizat folosind limbajul de adnotare XML. Implementarea, ce explicitează toate convențiile tipografice precum şi informațiile implicite, a condus la un volum textual de date de circa 8-10 ori mai mare față de conținutul textual echivalent al DEX-ului. Adnotarea XML a fost realizată automat, cu ajutorul compilatorului **DIC** [18]. Compilatorul a fost generat automat folosind JavaCC®, pe baza unei gramatici LL(7) ce descrie structura formală a intrărilor în DEX. **DIC** poate fi folosit pentru a genera documente XML (conform cu DTD-ul CONCEDE) pentru orice dicționar ce foloseşte convențiile tipografice adoptate în DEX. în [19] sunt prezentate o multitudine de dicționare realizate sau aflate în curs de realizare la Institutul de Lingvistică "Iorgu Iordan-Al. Rosetti" şi presupunând că ele urmăresc convențiile tipografice şi lexicografice adoptate în DEX, toate aceste surse lingvistice de referință pentru limba română ar putea fi transformate, cu efort minim, în resurse computaționale fundamentale pentru prelucrarea automată.

Varianta codificată a dicționarului nostru este numită XML-LEX iar structura sa este descrisă de DTD-ul *(Document Type Definition)* pe care îl reproducem în figura 1, dezvoltat în cadrul proiectului CONCEDE.

```
<!-- CONCEDE project - Deliverable DR2.1: concede.dtd      ->
<'-- copyright CONCEDE project.consortium, 1999        -->
<!-- ENTITY DECLARATIONS                    ->
<!ENTITY % a.global'
    id      ID      #IMPLIED
    n       CDATA   #IMPLIED
    lang    IDREF   #IMPLIED'
<!ENTITY % a.text'
    %a.global;
    rend    CDATA   #IMPLIED
    wsd     CDATA   #IMPLIED'
<!ENTITY % basetags'
```



```
    (orth|pron|hyph|syll|stress|pos|gen|case|number|gram|tns|
    mood|q|source|gloss|usg|def|per|aspect|degree|voice|eg|
    etym|xr|trans|itype|subc)'          >
<!ENTITY % dictbase.seq  '#PCDATA | na'        >
<!-- STRUCTURAL ELEMENTS                  ->
<!ELEMENT dictionary    (body)          >
<!ATTLIST dictionary %a.global;
    type      CDATA     #IMPLIED
    version   CDATA     #REQUIRED
    xmkspace  (default | preserve)  'preserve'   >
<!ELEMENT body      (entry+)          >
<!ATTLIST body %a.global; type CDATA     #IMPLIED    >
<!ELEMENT entry
    (hw, (%basetags;|struc|alt|brack)*)       >
<!ATTLIST entry  %a.global; type CDATA    #IMPLIED    >
<!ELEMENT struc     (%başetags; j struc | alt | brack)*  >
<!ATTLIST struc  %a.global; type CDATA    #IMPLIED    >
<!ELEMENT trans     (%basetags; | struc | alt | brack)*  >
<!ATTLIST trans  %a.global; type CDATA    #IMPLIED    >
<!ELEMENT alt       (%basetags; | brack )*      >
<!ATTLIST alt   %a.global; type CDATA    #IMPLIED    >
<!ELEMENT brack     (%basetags;)*         >
<!ATTLIST brack  %a.global; type CDATA    #IMPLIED    s
<!-- CONTENT ELEMENTS                    ->
<!ELEMENT voice  (%dictbase.seq;)*        >
<!ATTLIST voice  %a.text;             >
<!ELEMENT tns    (%dictbase.seq;)*        >
<!ATTLIST tns    %a.text;             >
' <!ELEMENT syll   (%dictbase.seq;)*        >
<!ATTLIST syll   %a.text;             >
<!ELEMENT subc   (%dictbase.seq;)*        >
<!ATTLIST subc   %a.text;             >
```

```
<!ELEMENT stress (%dictbase.seq;)* >
<!ATTLIST stress %a.text; >
<!ELEMENT source (%dictbase.seq;)* >
<!ATTLIST source %a.text; >
<!ELEMENT pos (%dictbase.seq;)* >
<!ATTLIST pos %a.text; >
<!ELEMENT per (%dictbase.seq;)* >
<!ATTLIST per %a.text; >
<!ELEMENT number (%dictbase.seq;)* >
<!ATTLIST number %a.text; >
<!ELEMENT na (#PCDATA) >
<!ATTLIST na %a.text; >
<!ELEMENT mood (%dictbase.seq;)* >
<!ATTLIST mood %a.text; >
<!ELEMENT m (%dictbase.seq;)* >
<!ATTLIST m %a.text; >
<!ELEMENT lang (%dictbase.seq;)* >
<!ATTLIST lang %a.text; >
<!ELEMENT itype (%dictbase.seq;)* >
<!ATTLIST itype %a.text; >
<!ELEMENT hw (%dictbase.seq;)* >
<!ATTLIST hw %a.text; >
<!ELEMENT gram (%dictbase.seq;)* >
<!ATTLIST gram %a.text; >
<!ELEMENT gen (%dictbase.seq;)* >
<!ATTLIST gen %a.text; >
<!ELEMENT degree (%dictbase.seq;)* >
<!ATTLIST degree %a.text; >
<!ELEMENT case (%dictbase.seq;)* >
<!ATTLIST case %a.text; >
<!ELEMENT aspect (%dictbase.seq;)* >
<!ATTLIST aspect %a.text; >
```



```
<!ELEMENT hyph (%dictbase.seq;)* >
<!ATTLIST hyph %a.text; >
<!ELEMENT eg (source | q | gloss)* >
<!ATTLIST eg %a.global; >
<!ELEMENT pron (%dictbase.seq;)* >
<!ATTLIST pron %a.text; type CDATA #IMPLIED >
<!ELEMENT q
   (%dictbase.seq; | gloss |ptr [xptr | oref)* >
<!ATTLIST q %a.text; type CDATA #IMPLIED >
<!ELEMENT etym
   (%dictbase.seq; | gloss | lang | m |ptr |xptr | oref)* >
<!ATTLIST etym %a.text; type CDATA #IMPLIED >
<!ELEMENT xr (%dictbase.seq; | ptr |xptr )* >
<!ATTLIST xr %a.text; type CDATA #IMPLIED >
<!ELEMENT def (%dictbase.seq; | ptr |xptr |oref |usg)* >
<!ATTLIST def %a.text; type CDATA #IMPLIED >
<!ELEMENT gloss (%dictbase.seq; | ptr [xptr |oref )* >
<!ATTLIST gloss %a.text; type CDATA #IMPLIED >
<!ELEMENT orth (%dictbase.seq; | ptr |xptr |oref |usg)* >
<!ATTLIST orth %a.text;
   expansion NMTOKEN #IMPLIED
   extent (full | pref | suff | part) "full"
   type CDATA #IMPLIED >
<!ELEMENT usg (%dictbase.seq;)* >
<!ATTLIST usg %a.text;
   type (syn|hyper|colloc|comp|plev|accjlang|gram|obj|
        subj|verb|hint|geo|domjregister|time|style|
        hyponym | antonym | other) "other" >
<!ELEMENT oref EMPTY >
<!ATTLIST oref %a.text;
   target IDREF #IMPLIED
   fullform NMTOKEN #IMPLIED >
```

```
<!ELEMENT ptr    EMPTY
<!ATTLIST ptr  %a.text;
    corresp  IDREFS        #IMPLIED
    next   IDREF         #IMPLIED
    prev   IDREF         #IMPLIED
    type   CDATA         #IMPLIED
    resp   CDATA         #IMPLIED
    crdate  CDATA         #IMPLIED
    targType  NMTOKEN         #IMPLIED
    targOrder  (y | n | u)       v
    evaluate  (all | one | none)    #IMPLIED
    target  IDREFS         #REQUIRED    >
<!ELEMENT xptr   EMPTY              >
<!ATTLIST xptr  %a.text;
    corresp  IDREFS        #IMPLIED
    next   IDREF         #IMPLIED
    prev   IDREF         #IMPLIED
    type   CDATA         #IMPLIED
    resp   CDATA         #IMPLIED
    crdate  CDATA         #IMPLIED
    targType  NMTOKEN         #IMPLIED
    targOrder  (y | n | u)       "u"
    evaluate  (all | one | none)    #IMPLIED
    target  NMTOKEN         #REQUIRED
```

Figura **1:** DTD-ul Concede, utilizat la implementarea XML-LEX

Această structură de codificare a fost adoptată în implementarea unui număr mare de dicționare, reprezentând un standard "de facto" în lexicografia computațională actuală [20]. Detalii suplimentare privind semantica entităților folosite în codificare şi a atributelor acestora pot fi găsite în documentația tehnică a proiectului la adresa www.itri.bton.ac.uk/projects/ concede/. în tabelul de mai jos, sunt exemplificate reprezentarea tipografică (de tip DEX) şi reprezentarea codificată în XML.

DEX

ZA^, zale, s.f. 1. Fiecare dintre ochiurile unui lanț; p. gener. (la pl.) lanț. • Lănțişor de metal întrebuințat uneori ca podoabă. • Cusătură în formă de lănțişor, executată de obicei la broderii. 2. (La pl.) împletitură executată din inele mici de fier legate unul de altul; p. ext. armură făcută din această împletitură, cu care se îmbrăcau oştenii în antichitate şi în evul mediu, spre a se apăra de loviturile duşmanilor. [Var.: (reg.) zâlă.zea s.f.] - Cf. ngr. zâva.

XML-LEX

```
<entry type="homonym"  id="ZÂ.2">
 <hw>ZA</hw>
 <alt>
  <brack>

<gram>nominativ_feminin__singularjndefinit</gram>
    <orth>ZA</orth>
  </brack>
  <brack>
    <gram>nominativ_feminin_pluraljndefinit</gram>
    <orth>zale</orth>
  </brack>
 </alt>
 <pos>substantiv</pos>
 <gen>feminin</gen>
 <struc n="1">
  <alt>
    <def>Fiecare dintre ochiurile unui lanț</def>
    <brack>
     <usg type="hyper">prin generalizare </usg>
     <usg>la pl.</usg>
     <def>lanț.</def>
    </brack> '
  </alt>
  <struc  type="Sec">
    <def>Lănțişor de metal întrebuințat uneori ca
podoabă. </def>
  </struc>
  <struc  type="Sec">
    <def>Cusătură în formă de lănțişor, executată de
obicei la
        broderii.</def>
  </struc>
 </struc>
 <struc n="2">
  <usg>La pl.</usg>
  <alt>
    <def>împletitură executată din inele mici de fier
legate unul de altul</def>
    <brack>
     <usg  type="hyper">prin  extensiune</usg>
     <def>armură făcută din această împletitură, cu
care se îmbrăcau oştenii în antichitate şi în evul
mediu, spre a se apăra de loviturile duşmanilor.</de
    </brack>
```

```
  </alt>
 </struc>
 <struc type="Varianta">
  <alt>
   <brack>
    <orth> zală</orth>
    <stress>  zâlă</stress>
    <usg>reg.</usg>
   </brack>
   <orth> zea</orth>
  </alt>
  <pos>substantiv</pos>
  <gen>feminin</gen>
 </struc>
 <etym>
 Cf.
  <lang>ngr.</lang>
 zâva.
 </etym>
</entry>
```

Figura 2: Conținut primar şi codificarea echivalentă în XML (cf. CONCEDE.dtd)

în tabelul din Figura 2, sunt exemplificate reprezentarea tipografică (de tip DEX) şi reprezentarea codificată în XML. Menționăm că reprezentarea tipografică din coloană stângă a Figurii 2 s-a obținut automat, folosind un convertor XML de format, proiectat astfel încât rezultatul generării (interpretarea marcajului XML) să fie cât mai apropiat de aspectul dicționarului tipărit. Structura de dicționar, definită mai jos, este suficient de generală pentru a permite implementarea diferitelor tipuri de dicționare. în fapt, DTD-ul CONCEDE a fost utilizat pentru codificarea a două dicționare bilingve: un dicționar Sloven-Englez şi un dicționar Român-Francez.

Adnotarea XML fiind independentă atât de convențiile tipografice cât şi de limba dicționarului, este posibilă căutarea multi-criterială a informației în unul, două sau mai multe dicționare explicative ale unor limbi diferite. De pildă, o căutare multi-criterială ar putea fi parafrazată astfel:

Găseşte şi afişează toate intrările ce corespund ***substantivelor feminine, de origine neo-greacă*** *şi al căror cuvinte titlu* ***încep cu secvența de litere ZA.****O* astfel de căutare va avea ca rezultat tipărirea cel puțin a intrării corespunzătoare cuvântului titlu ZA[2]:

ZA[2]**,** zale, s.f. **1.** Fiecare dintre ochiurile unui lanț; *p. gener.* (la pl.) lanț. • Lănțişor de metal întrebuințat uneori ca podoabă. • Cusătură în formă de lănțişor, executată de obicei la broderii. 2. (La pl.) împletitură executată din inel mici de fier legate unul de altul; *p. ext.* armură făcută din această împletitură, c care se îmbrăcau oştenii în antichitate şi în evul mediu, spre a se apăra d loviturile duşmanilor. [Var.: (reg.) **zâlă,zea** s.f.] - Cf. ngr. zâva.

33. Alte dicționare, lexicoane; indexul interlingual

Unul dintre rezultatele proiectului Multext-East îl constituie un lexicon d forme ocurență (LFO), cu peste 450.000 de intrări, care conține triplete de tipu <cuvânt, lernă, cod__morfo-sintactic>. Acest lexicon va fi completat cu formel flexionare (generate automat) ale lemelor din XML-LEX nereprezentate în LFO Codificarea folosită este compatibilă cu recomandările Eagles (http://www.ilc.p cnr.it/EAGLES/home.html) pentru adnotarea morfo-sintactică şi este documentat pe larg în [10].

O altă resursă lexicală esențială a fost Dicționarul de Sinonime al Limb Române - DSLR [21], care a fost transpus în formă electronică la Facultatea d Informatică a Universității "A.I.Cuza" din Iaşi. Forma electronică a DSLR a fos convertită în format XML astfel încât aceeaşi interfață ce a fost dezvoltată pentr XML-LEX funcționează şi cu XML-DSLR.

Din corpusurile paralele menționate mai sus şi folosind programul c implementează metodologia noastră de extragere a echivalenților de traducer [22], [23], [24] s-a construit un dicționar bilingv Român - Englez (de asemene transpus în format XML). Acest lexicon bilingv a fost validat manual şi îmbogățit c noi intrări din diverse surse publice.

în sfârşit, o resursă extrem de valoroasă a fost şi Indexul Interlingual (IL al EuroWordNet, exportat în format XML cu editorul VisDic produs la Universitate Masaryk din Brno [25].

3.4. Alegerea nucleului lexical

Vom da câteva definiții ale unor noțiuni pe care le vom folosi în cele c urmează.

Când ne plasăm într-un context monolingv, vorbim despre *sensuri înțelesuri* şi *sinseturi.* Un cuvânt are unul sau mai multe *sensuri.* Un sens referă u *înțeles.* în EuroWordNet sensurile unui cuvânt sunt numerotate în funcție d frecvența lor, iar sensul unei leme este denotat adăugând numărul sensului l forma ortografică a acesteia. O mulțime de sensuri astfel specificate (ex. *action activityl, activitenessl)* care referă acelaşi înțeles este numit *sinset* şi constituie insuşi denotația înțelesului sensurilor din sinset. Cu alte cuvinte, un sinse reprezintă *lexicalizarea unui înțeles* în contextul monolingv curent.

Dacă abstractizăm noțiunea de *înțeles,* definită ca mai sus, astfel încât s nu mai facem referirea la un anumit context monolingv, vom vorbi despre *concep*

care sunt referite de *înțelesurile* lexicalizate în diferitele limbi. Aşadar, putem vorbi despre concepte care au sau nu realizare lingvistică într-o limbă sau alta. Un concept este un construct cognitiv, independent de limbă, care în EuroWordNet este totdeauna lexicalizat cel puțin într-una dintre limbi. Un concept este mai departe rafinat în termeni de distincții semantice elementare (trăsături semantice), deci putem vorbi despre gruparea conceptelor în funcție de trăsăturile lor semantice.

în EuroWordNet şi deci şi în BALKANET, ILI este definit ca o colecție nestructurată de intrări de forma: <ILI-index><descriere ontologică><glosă> {domeniu}. Indexul interlingual inițial a fost construit plecând de la versiunea 1.5 a Wordnet-ului şi deci glosele pentru fiecare concept au fost importate direct din sinsetul englezesc care se referata înțelesul conceptualizat în ILI.

Pentru a facilita o cât mai bună intercorelare a wordneturilor monolingve din cadrul proiectului şi pentru a înlesni extensia lor ulterioară, consorțiul proiectului a decis ca procesul implementărilor paralele sa fie centrat pe concepte (independente de limbă) selectate de comun acord, la momente succesive de timp.

O primă selecție a constituit-o mulțimea aşa-numitelor "concepte de bază" definite în EuroWordNet ca fiind acele concepte din ILI lexicalizate în limba engleză (în WORDNET) prin sinseturi plasate pe un nivel ierarhic cât mai sus şi, în plus, care au un număr mare de hiponimi direcți (tot în WORDNET). Rațiunea acestei decizii a constat în faptul că, aceste concepte fiind foarte generale şi totodată productive în definirea unor concepte mai particulare, este foarte probabil ca ele să fie lexicalizate în majoritatea limbilor de interes. Acest lucru a fost probat atât în EuroWordNet cât şi în BALKANET. Mulțimea *conceptelor de bază* (o motivație mai detaliată a selecției lor este prezentată în [4] în raport cu obiectivele EuroWodNet) conține 1.310 concepte, fiecăruia dintre ele fiindu-i ataşată o glosă explicativă şi o *descriere ontologică* (vezi [26]).

După implementarea, în toate cele 5 limbi ale proiectului, a nucleelor de ontologii lexicale corespunzând conceptelor de bază, s-a făcut o nouă selecție, de data aceasta, conținând 4.000 de noi concepte interlinguale.

Selecția a avut în vedere, pe de o parte maximizarea compatibilității cu EuroWordNet, iar pe de altă parte relevanța stocului lexical pentru fiecare limbă din perspectivă monolingvă. Primul criteriu a fost operaționalizat alegându-se acele concepte lexicalizate în cele mai multe limbi din EuroWordNet. Limita inferioară a numărului de limbi a fost fixată la 5, astfel încât după implementarea acestor concepte în BALKANET ele să fie lexicalizate în cel puțin 10 limbi.

Criteriul relevanței monolingve a condus la propunerea mai multor mulțimi candidate de concepte. Pentru fiecare limbă a proiectului au fost efectuate analize cantitative în context strict monolingv. Metodele de analiză au diferit de la partener la partener, în raport cu datele şi instrumentele disponibile pentru lim
După analiza acestor mulțimi, au fost incluse în mulțimea finală acele au apărut în cel puțin două propuneri. Mulțimea finală a conce ordonată după numărul de limbi din EuroWordNet ce le lexicaliz numărul de limbi din^ALKANET care le-au propus. Primele 4000 de în această listă au fost de comun acord alese ca țintă comună pen doua etapă a proiectului.

în continuare prezentăm metodologia folosită pentru limba r selecția fondului lexical în cadrul BALKANET. Analiza cantitativă asupra unui corpus foarte mare, format din mai multe romane şi dintr texte jurnalistice culese de pe web. Corpusul (conținând mai mu milioane de cuvinte) a fost supus unor prelucrări statistice, fiin lematizat automat, iar cuvintele care prezentau interes (substan verbe, adjective şi adverbe) au fost sortate în funcție de frecvența lo extras în acest fel o listă de mai mult de 30.000 de leme. în funcți acestora în textele din corpus, această listă a fost împărțită corespunzând celor mai frecvente 10.000 de leme (I), următoa frecvente 10.000 (II) şi restul (III). Frecvența dintr-un corpus este c mulți lexicografi un criteriu subiectiv. Printre cele mai puternice numără volumul şi reprezentativitatea textelor incluse în corpusul fol cantitativă. Luând în calcul faptul că din ce în ce mai multe texte s pe web, mărimea corpusului nu mai reprezintă o problemă semn reprezentativitatea rămâne în continuare un punct slab. Definirea e textelor care trebuie incluse într-o analiză cantitativă face obiectul polemici şi nu vom insista asupra ei. Având în vedere că datele n aproape în întregime din texte jurnalistice, problema reprezentativit îndreptățire ridicată. Dicționarul de Frecvențe al Cuvintelor Române publicat cu mult timp în urmă, bazat pe un *corpus balansat* de 500.0 (teatru, nuvele şi scurte povestiri, eseuri memorii şi corespo jurnalistice, literatură tehnică) conține cele mai frecvente 5.000 d dacă este foarte controversat, FDLW este încă folosit de mulți lingvi referință. Comparația pe care am făcut-o a arătat că mai toa inventariate de FDRW se găsesc şi în lista obținută de noi, chia aceleaşi scoruri de frecvență. Pe lângă frecvența în corpus am a două criterii mai puțin controversate şi care au putut fi operaționaliz resursele lingvistice disponibile şi instrumentele noastre de analiză Primul este numărul de sensuri pe care un cuvânt (împreună cu expresiile în care participă) îl are într-un dicționar. Al doilea es definiții de dicționar în care apare un anumit cuvânt. Al treilea cri încă în analiză, ar putea fi numărul de derivate lexicale ale unui cu pertinentă analiză din acest punct de vedere, o excelentă lucrare es

în această fază a proiectului BALKANET, ne-am concentrat atenția asupra substantivelor din limba română, iar datele experimentale raportate mai jos se referă doar la acestea. Având însă în vedere că procedurile tehnice nu depind de categoria gramaticală, metodologia şi procedura vor fi aceleaşi şi pentru verbe, adjective şi adverbe. Luând în calcul numai primele două clase de frecvență descrise mai sus (primele 20.000 cele mai frecvente din corpusul jurnalistic) am extras din XML-LEX mai mult de 8.000 de intrări de substantive şi substantive compuse (care însumează aproximativ 35.000 de sensuri) astfel încât productivitatea definițională PD (numărul de definiții în care participă un substantiv) să fie cel puțin 3. Lista a fost sortată în funcție de productivitatea definițională şi numărul de sensuri ale fiecărui cuvânt titlu.

Substantiv	Productivitate definițională	Număr de sensuri	$FRECV_{range}$
acțiune	2279	13	I
persoană	1979	9	I
parte	1882	94	I
formă	1286	21	I
obiect	1204	16	I
fapt	1044	11	I
rasism	**3**	1	II

Figura **3:** Ordonarea candidaților

Pentru toate aceste substantive am extras traduceri englezeşti din dicționarul de echivalenți de traducere. Procedurile pentru extragerea automată a echivalenților de traducere din corpusuri paralele ca şi procedura de discriminare a sensurilor sunt descrise pe larg în [22], [23], [29], [15], [16]. Fiecare substantiv din limba română a fost pus în corespondență cu lista tuturor conceptelor din ILI corespunzătoare traducerilor sale în engleză. Conceptele astfel identificate au fost sortate după rangul corelat al substantivelor româneşti de la care s-a pornit.

Interesant de remarcat este că dintre cele 4000 de concepte selectate în final prin armonizarea propunerilor tuturor partenerilor, circa 2600 s-au regăsit şi în primele 4000 de concepte ale ierarhiei noastre. Toate cele 4000 de concepte selectate de consorțiu se regăsesc printre primele 6000 de concepte ale ierarhiei noastre.

Toate substantivele reprezentând potențiale lexicalizări ale celor 4000 de concepte din cea de a doua selecție au fost automat puse în corespondență cu toate definițiile lor din XML-LEX. De asemenea, ele au fost corelate cu lexicalizările din limba engleză ale celor 4.000 de concepte. Prin intermediul dicționarului de echivalenți de traducere englez-român, fiecare concept a fost asociat cu

lexicalizarea din limba engleză (extrasă din WORDNET) şi cu potențialele lexicalizări în limba română.

Dicționarul de Sinonime al Limbii Române (DSLR), digitizat şi codificat în XML, a fost folosit pentru a extrage seriile sinonimice pentru cuvintele româneşti selectate. în XML-DSLR unii membri ai seriilor sinonimice sunt arhaisme sau regionalisme. Discuțiile preliminare au condus către ideea de a elimina toate cuvintele care fac parte din aceste clase (ne-am bazat pe cerința de a construi un nucleu lexical de uz general în limba română contemporană). Totuşi, pentru eventualitatea în care aceste cuvinte filtrate (împreună cu informațiile despre uz) vor fi necesare mai târziu, s-a asigurat recuperabilitatea lor. Seriile sinonimice româneşti au fost considerate ca posibile sinseturi şi adăugate la asociațiile descrise mai sus.

4. Instrumente software dezvoltate pentru proiectul BALKANET

Materialul lingvistic de bază descris în secțiunea anterioară, a fost asamblat prin intermediul unor programe unitare, astfel încât toată această informație este disponibilă într-o interfață "prietenoasă", prin care lexicograful alege echivalențele corecte de sens dintre cele potențiale. Această interfață este generată şi "personalizată" automat în funcție de mulțimea conceptelor interlinguale furnizată ca parametru de intrare unui generator de interfețe. Printr-un astfel de model arhitectural, a fost posibil ca sarcina construirii wordnet-ului pentru limba română să fie distribuită între membrii celor două colective româneşti participante la proiect şi judicios controlată. Pentru fiecare dintre aceştia s-a generat o interfață personalizată pentru o submulțime distinctă de concepte dintre cele agreate de consorțiul proiectului. Utilizatorul acestei interfețe, pe care generic îl numim în continuare *lexicograf,* va lucra în mod independent de ceilalți, construind, ca urmare a interacțiunii, fragmente ale wardnetului pentru limba română. La un moment dat, lexicograful alege un concept din mulțimea ce i-a fost repartizată căruia doreşte să-i ataşeze un sinset românesc. El are la dispoziție simultan sinsetul ce lexicalizează în limba engleză conceptul respectiv şi, pentru fiecare cuvânt englezesc din acest sinset, toate potențialele lui traduceri în limba română, aceste traduceri având ataşate toate definițiile conținute în XML-LEX. În plus, fiecare cuvânt românesc are ataşate toate seriile sinonimice din XML-DSLR în care el este prezent. Ceea ce trebuie să decidă lexicograful este (vezi figura 4)

a. care este cuvântul românesc a cărui definiție este cea mai apropiată de definiția conceptului lexicalizat în limba engleză;
b. care este cea mai bună serie sinonimică a acestui cuvânt;

c. care dintre definiţiile ataşate cuvintelor dintr-o serie sinonimică este cea mai adecvată pentru a fi aplicabilă tuturor cuvintelor din seria respectiva.

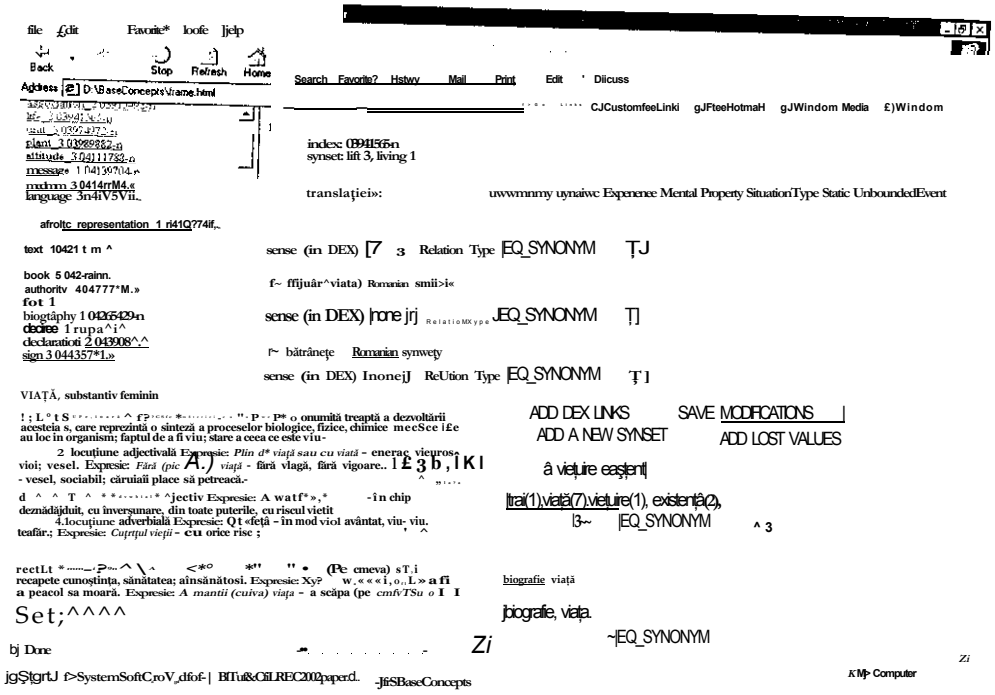


Figura **4:** Editorul pentru construirea sinseturilor

In majoritatea cazurilor, definiţiile extrase din XML-LEX corespunzând sinonimelor dintr-un sinset nu sunt identice, lexicografii alegând p S a mai apropiata de definiţia conceptului corespunzător (vezi figura 5)

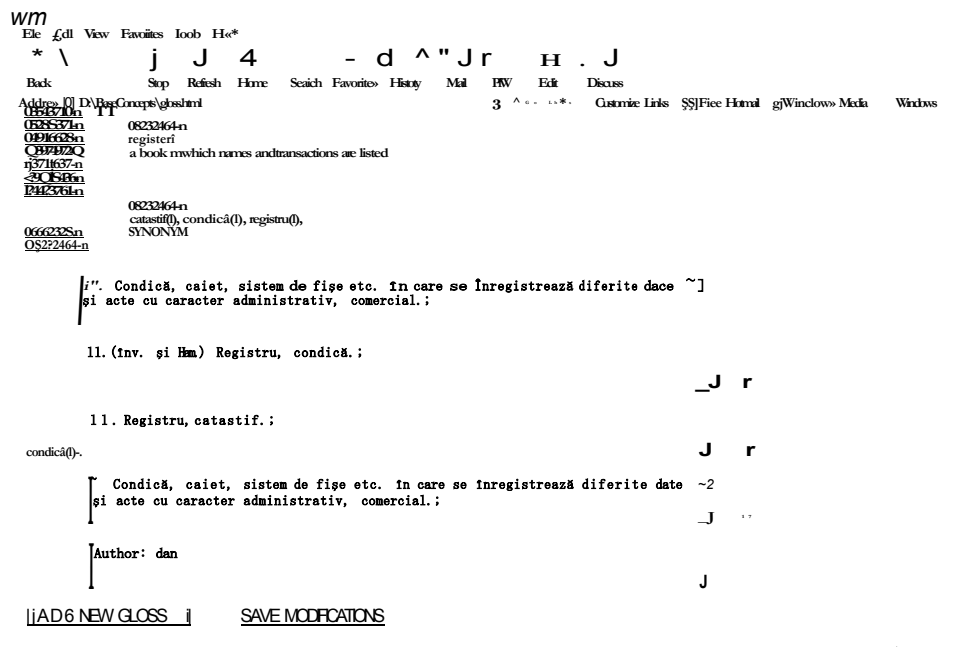


Figura **5:** Editorul pentru asignarea gloselor

Merită menţionat că în faza asocierii gloselor a devenit evidentă incorectitudinea alcătuirii unor sinseturi, ele fiind modificate. în alte cazuri Dicţionarul Explicativ al Limbii Române include în aceeaşi definiţie două sensuri care sunt demarcate în ILI ca două concepte diferite. în astfel de situaţii strategia generală a fost să se despartă definiţia românească şi să se ataşeze ca glosă partea relevantă.

Fragmente create de fiecare lexicograf sunt agregate în mod incremental în structuri din ce în ce mai complexe şi mai acoperitoare din punct de vedere lexical. Acest proces de agregare se realizează în mod centralizat, astfel încât corectitudinea structurilor rezultate să poată fi controlată şi, în cazul conflictelor, să se poată identifica şi corecta sursele de conflict (de exemplu: acelaşi sens pus în corespondenţă cu concepte diferite, sensuri diferite ale aceluiaşi cuvânt puse în corespondenţă cu acelaşi concept, literali fără identificatori de sens etc). Corectarea unor conflicte între două porţiuni ale structurii agregate poate să genereze conflicte între alte părţi ale sale. Pentru evitarea acestui pericol au fost proiectate mecanisme de control centralizat al unificării subseturilor de wordnet ce gestionează efectul global al oricăror modificări locale.

4.1. Importul relațiilor taxonomice; vizualizare sincronizată a mai multor wordneturi

Construcția sinseturilor şi punerea lor în corespondență cu conceptele interlinguale reprezintă doar una din cele două dimensiuni fundamentale ale procesului de construire a unei rețele semantice lexicale pusă în corespondență cu indexul interlingual, respectiv cea de implementare a nodurilor şi echivalarea acestora cu conceptele interlinguale. Cea de a doua dimensiune a procesului construcției rețelei o constituie definirea relațiilor (intraiinguale) între nodurile create şi echivalate în prima fază. Deosebit de importante sunt relațiile taxonomice care stabilesc o ierarhie de tip generic-specific între sinseturile unui wordnet.

Stabilirea relațiilor taxonomice între sinseturile wordnetului pentru limba română s-a făcut automat (urmată de validarea umană) în baza principiului "echivalenței ierarhice interlinguale" **[30].** în esență, acest principiu afirmă că:

1. dacă sinsetul S_{1LA} din limba LA şi sinsetul S_{1LB} din limba LB sunt echivalate cu acelaşi concept din ILI şi
2. dacă sinsetul S_{2LA} din limba LA şi sinsetul S_{2LB} din limba LB sunt echivalate cu acelaşi concept C_2 din ILI şi
3. dacă în limba A sinseturile S_{1LA} şi S_{2LA} sunt într-o relație ierarhică H^+ (H^+ denotă compunerea de un număr de ori cel puțin egal cu 1 a relației H, în cazul nostru: has-as-hypernym),atunci:

în limba B sinseturile S_{1LB} şi S_{2LB} sunt într-o relație ierarhică similară H^+ (deşi lanțurile de relații H pot fi de lungimi diferite în cele două limbi).

Principiul explicitează necesitatea ca interpretarea relațiilor folosite în ontologia multilingvă să fie similară, aşadar defineşte coeziunea interpretativă a relațiilor ontologice în toate limbile participante la proiect. Acest principiu este reprezentat schematic în figura 6:

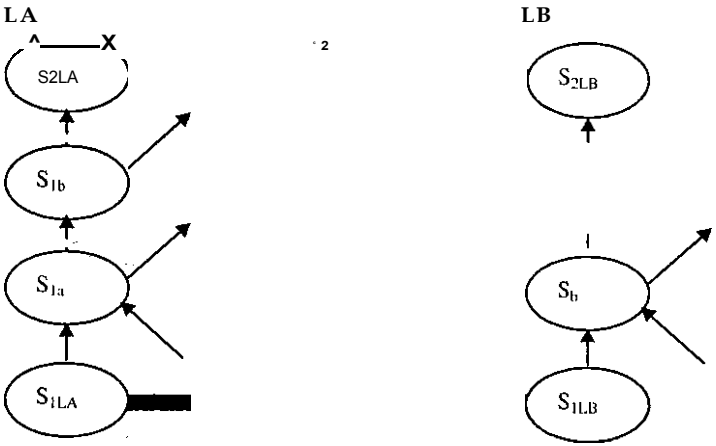


Figura **6:** (S_{1LA} EQ-SYN S_{1LB})& (S_{2LA} EQ-SYN S_{2LB})&(S_{1LA} H^+ S_{2LA}) (S_{1LB} H^+ S_{2LB})

în secțiunea următoare vom arăta pe un caz concret cum poate fi exploatat acest principiu pentru a importa (şi eventual valida/corecta manual) relațiile dintr-un wordnet în care structurile ierarhice au fost stabilite, într-un wordnet pentru care au fost stabilite doar relațiile de echivalentă translațională cu indexul interlingual (ILI).

Ultima etapă a construirii unui grup de sinseturi este transformarea rezultatelor interacțiunii lexicografului cu interfața descrisă anterior într-un format independent de limbă (codificare XML) şi specific editorului multilingual de ontologii lexicale numit VisDic [25]. Odată generat acest format, el poate fi încărcat în VisDic, iar wordnetul pentru limba română poate fi vizualizat în mod sincron cu toate celelalte wordneturi încărcate. In figura de mai jos este ilustrată afişarea în mod sincron a sinsetului românesc *(ființă__1, formă de viațăjt, viețuitoare_1, vietateji)* şi a celui englezesc *(being__1 life formjl, living thingjl, organism_1)* şi a arborilor lor de hiponimi. Cele două sinseturi sunt aliniate via ILI, ambele fiind echivalate independent cu conceptul interlingual cu identificatorul *00002728-n.*

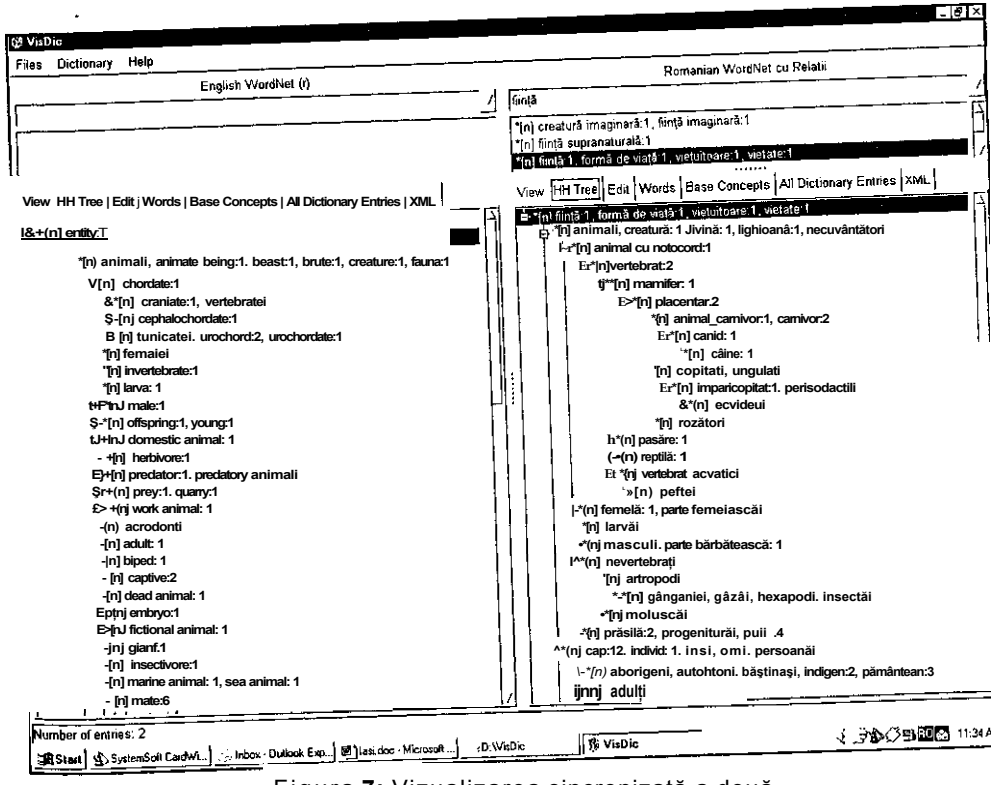


Figura **7:** Vizualizarea sincronizată a două ontologii lexicale cu ajutorul VisDic

Editorul de ontologii multilingve, VisDic, a fost dezvoltat în cadrul proiectului BALKANET pentru a substitui funcționalitatea asigurată în cadrul EuroWordNet de editorul Poiaris, dezvoltat de firma Lernout & Hauspie. Implementat inițial pentru ca rezultatele proiectului BALKANET să poată fi utilizate în regim liber de restricții comerciale (Poiaris poate fi utilizat doar contra cost), VisDic este constant îmbunătățit cu facilități noi a căror necesitare apare pe măsura evoluției proiectului BALKANET, fiind deja unul dintre cele mai puternice instrumente existente pentru gestiunea ontologiilor multilinguale.

5. Principiul conservării trans-linguale a ierarhiei lexicale. Studiu de caz: Condimente, mirodenii, sosuri şi alte ingrediente

Vom considera fragmentele din RO-WordNet şi WordNet 1.5 arătate în figura 8. Săgețile reprezintă relațiile taxonomice (de la hiponime spre hipernime) în cele două wordneturi. Liniile groase reprezintă relațiile de echivalență de traducere (EQ-SYN) dintre sinseturile celor două limbi, aceasta însemnând că sinseturile respective sunt puse în corespondență cu același concept din ILI. Linia groasă întreruptă reprezintă o relație EQ-SYN identificată ca nerespectând principiul conservării trans-linguale a ierarhiilor lexicale din cele două wordneturi. Inconsistența este semnalată deoarece în română relațiile ierarhice (de hiponimie) dintre *mirodenie(RO)* şi *condiment(RO)* ca şi dintre *ketchup{RO)* şi sos(RO) nu sunt verificate de echivalenții lor în limba engleză: sp/ce(EN) este frate cu *condiment{EN)* şi respectiv *ketchup(EN)* este frate cu *sauce(EN).* Dacă structura variantei 1.5 a WordNet este considerată cea corectă, acest exemplu arată că principiul păstrării ierarhiei nu este irefutabil. Pe de altă parte, dacă ar fi rezonabil să considerăm că WN 1.5 este amendabil (de exemplu făcând *mustard(EH)* şi *ketchup{EH)* hiponimii direcți ai lui sauce(en)) ca în figura 9, atunci principiul păstrării ierarhiei ar putea fi o puternică probă a consistenței[1].

în urma restructurărilor ierarhice şi de echivalare translațională, necesare pentru respectarea principiului conservării trans-linguale a ierarhiei lexicale (arătate în figura 9), interesant este faptul că a dispărut relația de echivalență între cuvântul românesc *condiment* şi cuvântul englezesc *condiment.*

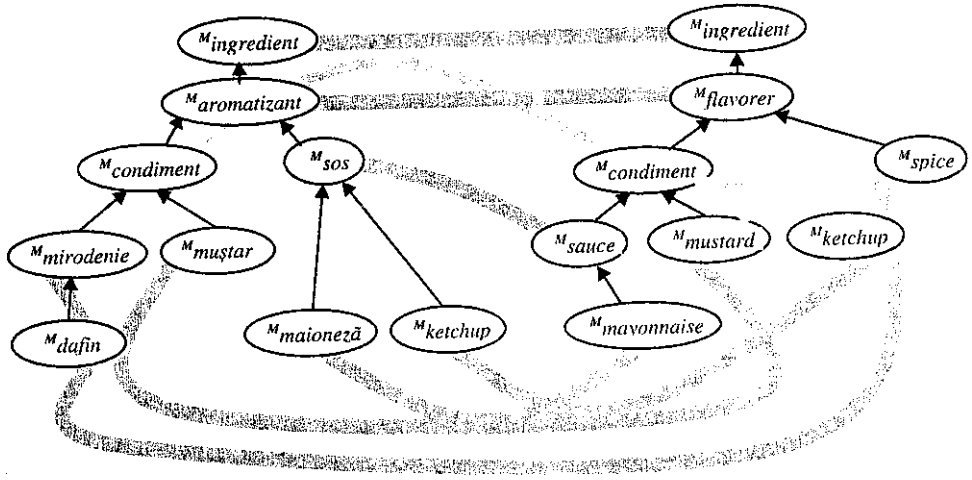


Figura 8: Nerespectarea principiului conservării trans-linguale a ierarhiei lexicale

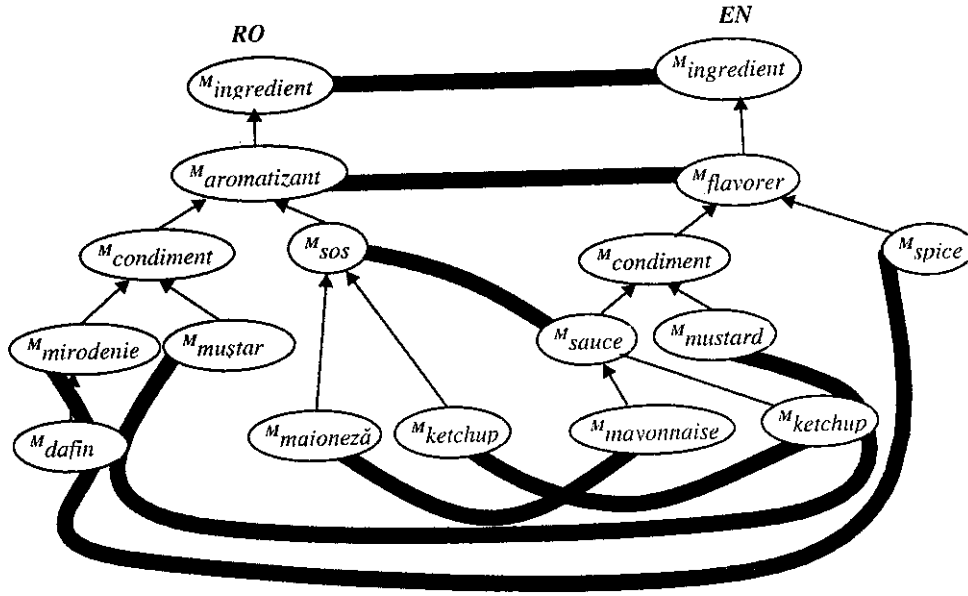


Figura 9: Reprezentare conformă cu principiul conservării trans-linguale a ierarhiei lexicale

Consultată recent asupra acestei probleme, Christiane Felbaum a confirmat existența unei erori în ierarhia WN1.5, probată, de altfel, şi de glosa /u/ketchup (thick spicy sauce made from tomatoesj.

Pentru ca această echivalență să fie posibilă, în condițiile principiului conservării trans-linguale a ierarhiei lexicale, ar trebui ori ca în limba engleză *spice* **să fie** un hiponim al lui *condiment* iar *sauce* **să nu fie** un hiponim al lui *condiment* ci frate, ori în limba română *sos* **să fie** un hiponim al lui *condiment* iar *mirodenie* **să nu fie** un hiponim al lui *condiment* ci frate. Ambele variante au fost respinse de experții consultați, lexicografi şi vorbitori nativi ai limbii engleze şi respectiv române. Singura concluzie posibilă este că în română şi engleză cuvântul *condiment* nu reprezintă exact acelaşi lucru.

6. Concluzii

Realizarea ontologiei lexicale pentru limba română, în contextul multilingual definit de proiecte de tipul EuroWordNet, Balkanet şi GlobalWordnet (www.globalwordnet.org), este esențială pentru procesul de informatizare a limbii române. Experiența internațională arată că un astfel de proiect nu este niciodată închis, reclamând actualizare şi întreținere continuă, apărând mereu noi idei de îmbunătățire a performanțelor şi noi cerințe de exploatare. Specialiştii de la Princeton au anunțat deja versiunea 1.7.1 a Wordnet, mult îmbunătățită. în variantele ce vor urma, pe lângă extensia în continuare a fondului lexical, toate cuvintele non-funcționale apărând în-definiții vor conține referințe spre sinsetul corespunzător contextului de utilizare. Cu alte cuvinte, Wordnet va deveni simultan şi un dicționar şi un corpus adnotat la nivelul sensului. O altă dezvoltare semnificativă o va reprezenta traducerea definițiilor din Wordnet în formule logice, adecvate prelucrărilor inferențiale. Acest proiect, coordonat de Dan Moldovan şi Sanda Harabagiu se află în derulare la Universitatea Texas din Dallas [31], [32].

Astfel de extensii vor trebui considerate în viitor şi în wordnetul pentru limba română aflat deocamdată în fază incipientă. Obiectivul final prevăzut pentru cei trei ani de derulare ai proiectului BALKANET (septembrie 2004) este realizarea unui nucleu de câte 8.000 de sinseturi în fiecare din limbile proiectului.

în acest moment, la mai puțin de un an de la începerea proiectului, wordnetul românesc se află cu mult înaintea graficului prevăzut, având deja create peste 6.000 de sinseturi. Se poate estima că, în condiții normale, în cei peste doi ani care au mai rămas wordnetul românesc va ajunge la peste 20.000 de sinseturi, acoperind pesta 40.000 de literali. Atingerea unui volum lexical similar cu al altor wordneturi necesită însă continuarea proiectului şi după anul 2004, atragerea unor noi colective de specialişti în această întreprindere şi desigur găsirea surselor de finanțare, în principal interne, care să permită dezvoltarea şi întreținerea wordnetului românesc. Operaționalizarea acestui obiectiv poate fi facilitată de contextul organizatoric creat de curând prin înființarea la Academia Română a Comisiei de Informatizare pentru Limba Română (CILR) precum şi a Consorțiului

de Informatizare pentru Limba Română (ConsILR: http://www.consilr.info.uaic.ro/), for executiv al CILR.

A fost construită o platformă software de dezvoltare incrementală a rețelei semantice ce permite implementarea independentă de regiuni ale rețelei şi integrarea ulterioară a acestora. Viabilitatea acestui concept arhitectural şi a demersului de dezvoltare distribuită a wordnetului au fost validate prin implicarea în procesul de construire a 10 specialişti, cărora li s-au adăugat încă 12 studenți masteranzi de la Facultatea de Litere a Universității Bucureşti şi Facultatea de Informatică a Universității "A.I. Cuza" (cele două facultăți ce au programe de Maşter în domeniul prelucrării limbajului natural şi al lingvisticii computaționale). Rezultatele produse în mod independent au fost agregate fără nici o dificultate. Mediul lingware de dezvoltare conține un modul special de verificare a corectitudinii deciziilor lingvistice la crearea sinseturilor româneşti sau la punerea lor în corespondență cu conceptele indexului interlingual. După cum era de aşteptat, procesul de integrare a rezultatelor parțiale furnizate de fiecare membru al' celor două echipe de realizare a evidențiat o serie de inconsistențe cu explicații diverse:

- neatenție în asignarea sensurilor, generată de oboseala expertului decident uman;
- granularitate semantică diferită între sensurile explicitate în XML-LEX şi sensurile conceptelor din ILI;
- absența lexicalizării în limba română a unor concepte existente în ILI şi introducerea unor forme perifrastice cu definiții ad-hoc;
- erori sau incompletitudini existente în sursele lingvistice primare folosite în implementare.

Inconsistentele depistate, atât de natură structurală, dar mai ales cele de natură semantică au fost înregistrate, analizate şi unele dintre ele corectate. Altele, necesită o analiză mai profundă şi rezolvarea lor a fost amânată pentru o etapă ulterioară a proiectului. Aceasta cu atât mai mult cu cât, prin analiza similară pe care am efectuat-o asupra wordneturilor pentru celelalte limbi din proiect, am constatat că există multe similarități ale acestor genuri de inconsistențe. Sunt puse astfel în evidență o serie de concepte din ILI pentru care diferența semantică dintre lexicalizările lor este prea mică pentru a fi sesizată uşor chiar şi de către un vorbitor nativ al limbii respective. Distincții atât de rafinate au, din perspectiva prelucrării automate şi mai ales a traducerii automate, o utilitate limitată iar în context multilingv pot fi chiar surse de eroare. Pericolul micşorării distanței semantice (am putea numi acest fenomen pulverizarea conceptuală) între conceptele din ILI este amplificat de adăugarea unor concepte ce au lexicalizări într-o singura limbă sau într-un număr mic de limbi. O soluție pentru evitarea idiosincrasiilor lexicale într-un context multilingv şi a disparităților de traducere este

gruparea conceptelor foarte apropiate semantic în ceea ce s-ar putea numi *concepte agregat* Lexicalizările înțelesurilor din două sau mai multe limbi, puse în corespondență cu aceleaşi concepte din ILI sau cu concepte membre ale unui agregat, vor putea fi folosite ca echivalenți de traducere în pofida unor diferențieri semantice specifice unei limbi sau alteia *{ciorbă, sarmale, pepper pot, porcupine ball* etc; vezi şi exemplele din secțiunea precedentă). Analiza inconsistentelor interumane în echivalarea înțelesurilor dintr-o limbă cu conceptele interlinguale din ILI, precum şi identificarea conceptelor distincte puse în corespondență cu echivalenți de traducere (extraşi automat din corpusuri paralele sau găsiți într-un dicționar bilingv clasic) pot furniza informații calitative mult mai interesante (cel puțin din perspectiva psiho-lingvisticii) şi mai demne de încredere decât o analiză statistică. Aceasta este o promițătoare direcție de cercetare ce se dezvoltă în paralel cu activitatea principală de construcție a wordnetului pentru limba română.

Referințe bibliografice

[1] Danzin, A. - Towards a European Language Infrastructure" raport al Comisiei Europene, 1992.

[2] Fellbaum, Ch. (ed.) - *WordNet: An Electronic Lexical Database,* MIT Press, 1998, 423 p.

[3] Bloksma, L., Diez-Orzas and Vossen, P. - The User Requirements and Funcțional Specification of the EuroWordNet-project *EWN-deliverable D.001,* LE-4003, 1996.

[4] Vossen, P. (ed.) - "A Multilingual Database with Lexical Semantic Networks", Kluwer Academic Publishers, Dordrecht, 1998.

[5] Stamou, S., Oflazer, K., Pala, K., Christoudoulakis, D., Cristea, D., Tufiş, D., Koeva S., Totkov, G., Dutoit, D., Grigoriadou, M. - BALKANET A Multilingual Semantic Network for the Balkan Languages, in *Proceedings of the International Wordnet Conference,* Mysore, India, 1997.

[6] Tufiş, D. - *"Promovarea Limbii Române în SI-SC",* în "Societatea Informațională - Societatea cunoaşterii: concepte, soluții şi strategii pentru România", Florin Gh. Filip (coord.), Ed. Expert, Bucureşti, ISBN973-8177-42-1, 2001, pp. 131-142.

[7] Erjavec, T., Ide, N., Tufiş, D. - Encoding and Parallel Alignment of Linguistic Corpora in Six Central and Eastern European Languages" in Michael Levison (ed) *Proceedings of the Joint ACH/ALL Conference* Queen's University, Kingston, Ontario (also on http://www.qucis.queensu.ca/ achallc97), 1997.

[8] Dimitrova, L., Erjavec, T., Ide, N., Kaalep, H.J., Petkevic, V., Tufiş, D. - Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. Proceedings of COLING, Montreal, Canada,

[9] Tufiş, D., Bruda, Şt. - Structure Markup in CES and Preliminary Statistics on Romanian Translation of Plato's "Republica", *Proceedings of International Seminar on Encoding,* Ljubliana, February, 1997, also in *TELRI News,* nr. 5.

[10] Tufiş, D., Barbu, A.M., Pătraşcu, V., Rotariu, G., Popescu, C. - Corpora and Corpus-Based Morpho-Lexical Processing, in Tufiş D., P. Andersen (eds.) *Recent Advances in Romanian Language Technology,* Editura Academiei, 1997, pp. 115-128.

[11] Tufiş, D., Rotariu, G., Barbu, A.M. - TEI-Encoding of a Core Explanatory Dictionary of Romanian. In Kiefer, F. and Pajzs J. (eds.) *Papers in Computațional Lexicography,* Hungarian Academy of Sciences, 1999, pp. 219-228.

[12] Ide, N. - *Corpus Encoding Standard: SGML Guidelines for Encoding Linguistic Corpora* First International Language Resources and Evaluation Conference, Granada, Spain, 1998, See also http://www.cs.vassar.edu/CES/.

[13] Tufiş, D. - Tiered Tagging and Combined Classifiers in F. Jelinek, E. Noth (eds) *Text, Speech and Dialogue,* Lecture Notes in Artificial Intelligence 1692, Springer, 1999.

[14] Tufiş, D., Popescu, C, Roşu, R - Automatic classification of documents by random sampling in *Proceeding of the Romanian Academy,* Series A, voi. 1, no. 2, 2000, pp. 18-28.

[15] Tufiş, D. - "A cheap and fast way to build useful translation lexicons" in *Proceedings of the 19th International Conference on Computațional Linguistics, COLING2002,* Taipei, 25-30 August, 2002, pp. 1030-1036.

[16] Ide, N., Erjavec, T., Tufiş, D. - "Sense Discrimination with Parallel Corpora" in Proceedings of the SIGLEX Workshop on Word Sense Disambiguation: Recent Successes and Future Directions. ACL2002, Philadelphia, 2002, pp. 54-60.

[17] DEX - Coteanu, I., Seche, L., Seche, M. (coord.). Dicționarul Explicativ al Limbii Române, ediția a II-a, *Univers Enciclopedic,* 1996, Bucureşti.

[18] Tufiş, D. - Blurring the distinction between machine readable dictionaries and lexical databases. *Research Report, RACAI-RR56, 2000, p. 56.*

[19] Vintilă-Rădulescu, I. - "Resurse lingvistice pentru limba română elaborate la Institutul de Lingvistică «Iorgu Iordan»", în acest volum, 2002.

[20] Erjavec, T., Evans, R., Ide, N., Kilgarriff, A. - The CONCEDE Model for Lexical Databases. *Proceedings of the Second Language Resources and Evaluation Conference* (LREC), Athens, Greece, 2000, pp. 355-362.

[21] Seche, L., Seche, M. - *Dicționarul de sinonime al limbii române.* Univers Enciclopedic, Bucureşti, 1997.

[22] Tufiş, D., Barbu, A.M. - *Computațional Bilingual Lexicography: Automatic Extraction of Translation Dictionaries,* in International Journal on Science and

Technology of Information, Romanian Academy, ISSN 1453-8245, Voi. 4, No. .3-4, 2001, pp. 325-352.

[23]Tufiş, D., Barbu, A.M. - *Extracting multilingual lexicons from parallel corpora,* in Proceedings of the ACH-ALLC conference, New York, 12-17 June, 2001,4 p.

[24]Tufiş, D., Barbu, A.M. - "Lexical token alignment: experiments, results and applications" In Proceedings of LREC2002, Las Palmas, Spain, 2002, pp. 458-465.

[25]Pavelek, T., Pala, K. - *VisDic: A new Tool for WordNet Editing* in Proceedings of the 1st International Wordnet Conference, Mysore, 2002.

[26]Rodriguez, H., Climent, S., Vossen, P., Bloksma, L., Peters, W., Alonge, A., Bertagna, F., Roventini, A. - The Top-Down Strategy for Building EuroWordNet: Vocabulary Coverage, Base Concepts and Top Ontology. In Piek Vossen (ed.) *EuroWordNet: A Multilingual database with lexical semantic networks,* Computers and Humanities, Voi. 32, Nos. 2-3, 1998.

[27]Juilland, A., Edwards, P.M.G, Juilland, I. - The Frequency Dictionary of Rumanian Words. *Mouton & CC,* London-The Hague-Paris, 1965.

[28] Dinu, M. - Personalitatea limbii române, Editura ALL, 1996, 368 p.

[29] Erjavec, T., Ide, N., Tufiş, D. - *Automatic Sense Tagging Using Parallel Corpora,* in Proceedings of the 6th Natural Language Processing Pacific Rim Symposium, Tokyo, Japan, 27-29 November, 2001, pp. 212-219.

[30] Tufiş, D., Cristea, D. - "Methodological issues in building the Romanian Wordnet and consistency checks in Balkanet", In Proceedings of LREC2002, Las Palmas, Spain, May, 2002, pp. 35-41.

[31]Moldovan, D. - "Question Answering Systems in Knowledge Management", IEEE Intelligent Systems, voi. 16, nr. 6, 2001, pp. 90-92.

[32]Harabagiu, S., Miller, G., Moldovan, D. - "WordNet 2 - A Morphologically and Semantically Enhanced Resource", in *Proceedings of SIGLEX-99,* Univ. of Maryland, 1999, pp. 1-8.

Algoritmi de segmentare a textului în unități de tip clauzal

D. GÂLEA,
Institutul de informatică Teoretică, Academia Română, Filiala Iaşi,
dgatea@iit.tuiasi.ro

N. CURTEANU,
Institutul de Informatică Teoretică, Academia Română, Filiala Iaşi,
curteanu@iit.tuiasi.ro

C. LINTEŞ,
Institutul de Informatică Teoretică, Academia Română, Filiala Iaşi

1, Introducere

Scopul lucrării este dublu: (a) Să prezinte şi să compare doi *algoritmi de segmentare* a frazei (româneşti) în unități de tip clauzal. *(b)* Să întregească şi să susțină două componente de bază ale *strategiei lingvistice* SCD (Segmentare-Coeziune-Dependență) [1], [2] de analiză a limbajului natural (LN): procesul de *segmentare* a textului de LN, şi *teoria* FX-bar [3], [23]. Segmentarea textului poate continua sau interfera cu stabilirea arborilor de dependență între unitățile clauzale şi subclauzale (unități sintagmatice) ale textului. Unitățile de tip-clauzal corespund, în general, *relațiilor retorice* dintre unitățile minimale ale discursului, astfel încât algoritmii de segmentare pot fi (şi chiar sunt) utilizați în aplicații ce țin de *teoria* şi *procesarea discursului.* Primul algoritm este o aplicare la limba română a segmentării frazei în unități de tip-clauzal, algoritm dezvoltat de Daniel Marcu în [4], [5] (şi prescurtat în cele se urmează *"algoritm-Marcu",* sau *"algoritm-M').* A doilea algoritm reprezintă o rafinare a segmentării în clauze şi grupuri sintagmatice din cadrul strategiei lingvistice SCD [1], [2], [3], [6] (prescurtat în cefe se urmează prin *"algoritm SCD").* Aceşti algoritmi sunt implementați într-un mediu specializat de procesare (dezvoltat sub C++), şi este realizată o comparație computațională a execuției segmentării de tip-clauzal pe un set consistent de fraze româneşti [7].

Segmentarea textului LN a devenit în ultimii ani un subiect intens cercetat şi cu multiple aplicații. O atenție specială a primit segmentarea textului de LN în unități dediscurs, în particular, segmentarea frazei în *unități minimale* de *discurs* de multe ori şi pe bună dreptate asociate cu unități de tip clauzal în numeroase

teorii sintactice, semantice, şi de discurs. Unităţile textuale de tip-clauzal obţinute (sau proiectate) prin mijloace orientate sintactic, către analiza de suprafaţă [Eng: *shallow],* s-au dovedit a fi esenţiale în numeroase tipuri de procesare a LN: parsare, traducere automată, generare de LN, interpretare de discurs, extragere de date lingvistice, regăsirea informaţiei, rezumare automată, rezoluţia anaforei etc. Un caz special de segmentare a textului este ceea ce numim *'chunking',* un proces dedicat obţinerii unor tipuri de *"segmente"* [Eng: *chunks]* dominate de anumite categorii (verb, substantiv, adjectiv-adverb, clauză). în continuare, vom folosi doar termenul de *segmentare* a textului de LN, considerând *chunking-u*\ drept un caz particular al procesului de segmentare a LN.

în analiza şi implementarea celor doi algoritmi de segmentare, *algoritmul-*M şi *algoritmul-SCD,* cel puţin două aspecte le considerăm a fi importante: **(a)** Se demonstrează că *algoritmul-M* de segmentare este scufundat în *algoritmul-SCD,* ceea ce înseamnă că primul dintre cei doi algoritmi poate fi obţinut ca un caz particular al claselor de marcheri, ierarhiei acestor clase, şi a segmentării (dependenţelor) obţinute de cel de-al doilea algoritm, **(b)** *Algoritmul-SCD* de segmentare poate fi conceput ca un bun punct de start în *proiectarea unui cadru general* pentru *algoritmii de segmentare* a textului de LN. Un asemenea cadru ar fi compus din: (b1) mai multe *sisteme de transformare* aplicate în cascadă, fiecare sistem component fiind format din *seturi şi subseturi specifice de etichete,* (b2) o *ierarhie* stabilită între câteva dintre cele mai importante *clase* ale acestor *etichete,* şi (b3) o *gramatică formală* (sau un automat finit) pentru recunoaşterea (sub)secvenţelor şi arborilor de etichete (în concordanţă cu ierarhia claselor de etichete). în abordarea prezentă, aceste aspecte sunt exemplificate de către o implementare C++ a celor doi algoritmi într-un mediu specializat, o bază de date a marcherilor (de discurs) româneşti, şi o ierarhie specifică a claselor de marcheri lingvistici. Cei doi algoritmi de segmentare consideraţi sunt executaţi şi comparaţi pentru un set consistent de fraze româneşti. Posibile dezvoltări şi aplicaţii sunt menţionate în [21], [22].

Importanţa segmentării de tip clauzal a frazei în procesul de parsare a textului a fost scoasă în evidenţă încă de la începutul anilor '80, iar studii teoretice datează mult mai devreme. în România, primele lucrări ştiinţifice şi contracte de lingvistică computaţională au conţinut, printre alte realizări meritorii, şi primele încercări de realizare a segmentării automate a frazei în clauze finite (şi non-finite) [8], [9], [10], [11]. în pofida unor mijloace formale neadecvate (gramatici formale) şi de programare (reţele ATN) disponibile la acel timp, ideile principale pe care se bazau abordările menţionate nu numai că au reprezentat premiere pentru acele timpuri, dar multe din ideile de atunci îşi păstrează încă o surprinzătoare actualitate, aceste fenomene de *come-back* ciclic fiind frecvente (şi perfect explicabile de evoluţia tehnologică) în momentul de faţă. Trebuie menţionate aici folosirea intensivă *a marcherilor de discurs (cue phrases, connectives),* întâlnită şi în [4], [5], [12], [13], a *predicativităţii* (apariţia categoriilor *deverbale)* [14], [15], a utilizării automatelor finite în analiza LN etc.

De fapt, o versiune a gramaticii formale preluată din [8] este folosită în *Pasul 6* al *algoritmului SCD-2002* de segmentare, în concatenarea marcherilor de nivel M3 şi M2 (vezi Secţiunea 4), în timp ce rudimente ale unor reguli similare din aceeaşi gramatică se regăsesc în algoritmul-M de segmentare, la compunerea acţiunilor care lucrează cu apariţia multiplă a marcherilor (de discurs) [4], [5] (vezi Secţiunea 3).

2. Segmentarea de tip clauzal cu algoritmul M-1997

Prescurtat în continuare ca *"algoritmul de segmentare* M-1997", sau simplu *"algoritmul* M-1997", algoritmul de *segmentare-Marcu* a frazei în unităţi de tip-clauzal [4], [5] funcţionează ca un automat finit, sau ca o reţea de tranziţie, bazat pe un set de *stări* şi *acţiuni.* în [4] se face o analiză de corpus a potenţialilor *marcheri de discurs,* numiţi şi *"sintagme indicatoare"* [Eng: cive *phrases]* şi *"conective",* cu scopul de a evalua contribuţia potenţială a diferiţilor marcheri la determinarea (delimitarea) unităţilor textuale elementare pe care sunt definite *relaţiile retorice,* în cadrul unităţii textuale standard care este fraza. în încercarea de a stabili principalele tipuri de funcţii ale marcherilor, şi anume de tip clauzal, frazai, de discurs, sau pragmatic, algoritmul de segmentare M-1997 consideră mai întâi următoarele *trei clase* de marcheri:

(Mari) în *prima clasă* sunt cuprinşi marcherii (sintagmele indicatoare) care joacă un rol în cadrul discursului pentru majoritatea fragmentelor de text ale corpusului analizat. Elementele din (Mari) vor fi numite în cele ce urmează *"marcheri de discurs",* iar specifici acestei prime clase sunt marcheri ca *"deşi'* [Eng: *although], "pe lângă"* [Eng: *besides], "dacă"* [Eng: *if\, "atunci'* [Eng: *then]* etc.

(Mar2) Marcherii din a *doua clasă,* numiţi *"marcherii de frază/clauză",* joacă în discurs, pentru majoritatea fragmentelor de text în care apar, rolul de *adiacenţi* la alţi marcheri de discurs sau clauzali. Un membru specific al clasei (Mar2) este considerat a fi *"şi'* [Eng: *and],* deoarece are rol clauzal de fiecare dată când apare înaintea altui marcher de discurs sau clauzal, cu toate că poate avea atât rol de discurs cât şi clauzal atunci când apare izolat.

(Mar3) *A treia clasă* conţine marcheri care s-au dovedit că joacă un rol de *delimitare a clauzelor* în majoritatea fragmentelor de text investigate în [4]; ei vor fi referiţi, simplu, ca *"marcheri clauzali'* (sau "de *clauză").* (Mar3) include, de asemenea, acei marcheri pentru care analiza de corpus nu a putut distinge între funcţia lor de discurs şi cea clauzală. *"După"* [Eng: *after]* este un astfel de element reprezentativ al (Mar3).

Marcu [4] a selectat mai mult de 450 de *marcheri* (pentru engleză) în cadrul analizei sale de corpus pentru marcherii de discurs şi de frază/clauză. Marcherii sunt stocaţi şi procesaţi într-o *bază de date* ale cărei înregistrări conţin următoarele câmpuri:

a. Câmpul denumit *Example* conţine un *fragment de text* din care a fost extras marcherul.

b. Câmpul *Marker* codifică *marcherul* însuşi, împreună cu marcherii de punctuaţie contextuali şi, atunci când este necesar, ceilalţi *marcheri adiacenţi.*

c. Câmpul *Usage* furnizează unul sau mai multe dintre *rolurile funcţionale* ale marcherului:

   (d) *Frazallclauzal* **(S),** atunci când marcherul nu îndeplineşte *nici o funcţie* în structurarea discursului;
   (c2) *De discurs* **(D),** când marcherul evidenţiază o *relaţie de discurs* între două unităţi textuale;
   (c3) *Pragmatic* **(P),** dacă există o *relaţie* între o construcţie lingvistică (sau non-lingvistică) care conţine marcherul, şi convingerile, planurile, intenţiile şi/sau scopurile de comunicare ale vorbitorului.

d. Câmpul *Break_action* (acţiune de oprire) conţine un nume de *acţiune* din mulţimea acţiunilor ce vor fi executate în cadrul procesului de segmentare. Acest proces este controlat de către un set de semnalizatori *(flaguri).* Execuţia unei acţiuni din mulţimea {NOTHING, NORMAL, COMMA, NORMALJ"HEN_c6mMA, END,'mATCH_PAREN, COMMA_PAREN, MATCH_DASH, SETÂND, SET_OR, DUAL} are unul dintre următoarele efecte:

   (d1) creează o *margine* pentru unitatea textuală elementară în *string-u*\ de intrare;
   (d2) setează un semnalizator *(flag).*

e. Câmpul *Position* specifică poziţia marcherului de discurs în cadrul unităţii textuale căreia îi aparţine. Valorile acestui câmp sunt B, **M** şi *E,* după cum marcherul este situat *la început* (B), *în mijlocul (M)* sau, respectiv, *la sfârşitul* (£) unităţii textuale.

3. Algoritmul de segmentare *M-1997*

Algoritmul M-1997 primeşte în intrare o frază **S** şi masivul *markers[n]* al marcherilor potenţiali de discurs şi clauzali din fraza **S.** Masivul *markers[n]* conţine marcherii recunoscuţi în **S.** Fiecare element al acestui masiv este caracterizat de către următoarea *structură de trăsături:*

- *Acţiunea* asociată acelui marcher;
- *Poziţia* marcherului în cadrul unităţii textuale elementare (B, **M** sau £);
- *Semnalizatorul hasjdiscoursejunction* care iniţial este setat la valoarea *"no".*

Câteva dintre variabilele importante cu care lucrează algoritmul **M**-1997 sunt: *"status", "parentheticaC* şi *"clauses".*

Algoritmul **M**-1997 pentru identificarea unităţilor de tip-clauzal din cadrul unei fraze are *două* părţi principale:

(1) Când variabila *"status"* este NIL, algoritmul **M**-1997 execută acţiuni care pot introduce margini ale unităţii textuale sau pot modifica variabila, influenţând procesarea marcherilor ulteriori. Pentru partea (1) a algoritmului **M**-1997, atunci când variabila *"status"* \a valoarea NIL, sunt considerate următoarele situaţii:

(1a) Dacă tipul de marcher este DUAL, determinarea marginilor unităţii textuale depinde de marcherul adiacent care precede marcherul curent analizat. în această situaţie, algoritmul **M**-1997 setează variabila *"status"* la aceeaşi valoare ca şi în cazul unui marcher de tip COMMA.

(1b) Dacă marcherul analizat curent nu este adiacent cu marcherul imediat precedent, atunci este identificată o margine a unităţii textuale.

(1c) Cel mai frecvent tip de marcher (şi de acţiune) este NORMAL, marcher care identifică o nouă unitate de tip clauzal a cărei margine-dreapta este dată de marcherul curent analizat.

(1d) Când marcherul de tip COMMA este precedat de un marcher de discurs, *sau*

(1e) Tipul marcherului este NORMAL__THEN_COMMA, *atunci* algoritmul **M**-1997 identifică o nouă unitate de tip-clauzal ca şi în cazul marcherului de tip NORMAL.

în oricare dintre cazurile (1c), (1d), (1e), variabila *"status"* este actualizată astfel încât o margine a unităţii textuale să fie identificată la prima apariţie a unei virgule (COMMA).

(1f) Pentru marcherul de tip NOTHING, singura acţiune constă în a atribui marcherului o utilizare specifică discursului.

(1g) Marcherii care introduc posibile apariţii de unităţi textuale parantetice (texte între paranteze) au doar efectul de a actualiza variabila *"status",* ca şi în cazul apariţiei marcherilor "sf şi *"sau".*

(2) Atunci când variabila *"status"* nu este NIL, algoritmul **M**-1997 execută acţiuni specifice pentru a realiza:

(2a) Tratarea informaţiei din paranteze. O dată identificată o paranteză deschisă, o linie-de-despărţire [Eng: *dash]* (între două asemenea liniuţe se introduce de obicei o apoziţie sau un text explicativ), sau un

marcher de discurs a cărui acțiune asociată este COMMA_PAREN, algoritmul M-1997 caută prima paranteză închisă, linie-de-despărțire, sau virgulă, ignorând toți ceilalți marcheri întâlniți pe parcurs. Acest tratament atrage după sine faptul că informației parantetizate *nu* îi este atribuită nici o stare pentru unitățile textuale elementare. Totuşi, algoritmul M-1997 evită stabilirea de margini parantetizate în cazurile în care prima virgulă care urmează după un marcher COMMA_PAREN este imediat urmată de un marcher *"şi"* ori *"sau".* De menționat este, de asemenea, că tratamentul aplicat informației dintre paranteze în algoritmul M-1997 poate conduce la rezultate eronate, ca în exemplul *"l-am dat lui Ion o rachetă de tenis, care i-a plăcut şi o minge de plastic, care nu i-a plăcut".* Acest tip de erori poate fi evitat printr-o tratare mult mai adecvată în cadrul algoritmului de segmentare *SCD.*

(2b) Dacă variabila *"status"* conține acțiunea COMMA, apariția primei virgule care nu este adiacentă unui marcher *"şi'* ori *"sau"* determină identificarea unei noi unități elementare de discurs. Algoritmul M-1997 nu este, capabil, în general, să distingă suficient de precis între rolurile de discurs şi frazale/clauzale ale marcherilor *"şi"* şi *"sau".* Anumite situații sunt totuşi recunoscute ca introducând funcții de discurs, ca de exemplu apariția unui marcher de discurs imediat după un *"şi"* ori *"of* caz în care valoarea semnalizatorului *has_discourse_ function* este stabilită la *"yes".*

Forma originală a algoritmului M-1997 [4], [5] este extinsă şi îmbunătățită în implementarea noastră pentru limba română (subsecțiunea 5.3) cu o analiză mai detaliată la nivelul ei superior, pentru apariții multiple şi corelate ale marcherilor de discurs/clauză [7].

4. Algoritmul de segmentare SCD-2002

Această secțiune prezintă partea de segmentare şi dependență, în principal la nivel de clauză, desprinsă din strategia lingvistică *SCD (Segmentare-Coeziune-Dependență)* [1], [2], [3], [6], [23]. Forma actuală a algoritmului, referită în restul articolului prin prescurtarea SCD-1994, este foarte apropiată de versiunea publicată în [1], [2]. Noutatea principală a algoritmului SCD-2002 față de SCD-1994 constă într-o rafinare a claselor de marcheri, o nouă ierarhie a acestora, şi în noul algoritm de stabilire a segmentării şi dependenței (structurării) clauzelor şi grupurilor sintagmatice. Vom pune în evidență relația dintre algoritmul M-1997 şi algoritmii SCD-1994 şi SCD-2002, arătând că primul este scufundat în ceilalți doi [7].

Rezultatele obținute prin execuția algoritmilor de segmentare M-1997 şi SCD-2002 pe aceleaşi fraze conduc la aceeaşi concluzie: SCD-2002 are o granularitate (mult) mai fină a claselor de marcheri în comparație cu cea a claselor algoritmului M-1997, iar rafinarea acțiunilor implicate în SCD-2002 conduce la delimitarea de unități textuale de tip-clauzal mai precise (de fapt mai corecte şi mai adecvate) decât cele obținute de către algoritmul M-1997, prețul computațional ce trebuie plătit pentru acest fapt rămânând să fie analizat.

Este de menționat că segmentarea clauzală practicată de SCD-2002 este doar un aspect particular al segmentării textului, deoarece se obțin şi alte *"bucăți' mal* mici de text dominate de nuclee semantice de tip N (Substantiv), V (Verb), A (Adjectiv-Adverb). Segmentarea rezultată din clasele de marcheri SCD-2002 se află într-o strânsă relație cu noua teorie *X-bar funcțională* (FX-bar) [3], [23], o altă componentă importantă a strategiei lingvistice generale *SCD.*

Din schema generală FX-bar propusă în [3] se detaşează următoarele nivele de proiecție la nivel lexical şi gramatical:

Tabelul 4.1.

Nivele de proiecție ale schemei FX-bar (vezi [31, [231)

Marcheri	Nivelul de Proiecție	Structura gramaticală	Exemple
trăsătura PRED sau EXIST (OBJECT)	nivel de lexicon; prin convenție, (BAR = -1)	[forma de dicționar a cuvântului; X = N,V, A, Pron,	*\a ploua ^conducere* (trăsătura ; PRED) *{clădire* (trăsătura EXIST, înțelesul obiectual) *clădire* (PRED, pentru înțelesul acționai) *creion* (EXIST) *filoună*
MO-marcher reprezintă aplicarea inflexiunii	XO (BAR = 0)	forma lexicală (de text) a cuvântului; X=N,V, A, ...	
M1-marcher se aplică nucleului XO M1(X0)=X1	X1 =CL0;(BAR=1) poate fi identificat şi cu nivelul 0 de proiecție a clauzei, BAR-CL = 0	sintagme XG (X=N, V, A), i.e. grupuri nominale, verbale, adjectivale ladverbiale	*orice steag alb ploua aleargă repede \nu aleargă deloc foarte bine studiat Măria i-a dat un măr fiicei sale.*
M2-marcher se aplică .proiecției X1 M2(X1)=X2=CL1 M2 se aplică unei singure clauze	proiecția X2 = C BAR = 2 şi BAR-CL = 1		*O femeie dăruind un măr unui bărbat conține o [clauză infinită.*
M3(CL1,CL1)=CL2 marcheri de discurs; M3 se aplică la două sau mai multe clauze	nivelul de proiecție X3 = CL2; BAR = 3 si BAR-CL = 2	relații de discurs între clauze finite	*Dacă plouă atunci plec mai devreme şi îmi iau şi \umbrela.*

4.1. Clasele de marcheri pentru algoritmul SCD-2002

Pentru algoritmul de segmentare SCD-2002 propunem o anumită rafinare a claselor de marcheri şi a ierarhiilor acestor clase din [1], [2], [7], schimbări ce constau în următorul set de marcheri, în concordanță cu *Tabelul* 4.1. de mai sus:

M3 = { *marcheri (de discurs) inter-clauzali}.*

Clasa de marcheri M3 este formată din *funcții* sau *relații* (atunci când marcherii sunt corelați), având ca argumente două sau mai multe clauze finite (unele dintre ele pot fi infinite). Aceşti marcheri sunt ceea ce [4], [5], precum şi alte abordări numesc *"marcheri de discurs",* şi se aplică proiecțiilor sintactice de nucleu X2 = CL1 (şi de nivel X3), de tip clauzal în teoria FX-bar (vezi Tabelul 4.1.).

M3 poate fi partiționată în următoarele subclase (în ordinea *descrescătoare* a *priorității* de definire a relațiilor de dependență - vezi Fig. 4.1.1.):

M33 = {marcheri (de discurs) inter-clauzali care introduc o dependență (neambiguă) de *supra-ordonare strictă}. Supra-ordonarea strictă* înseamnă *ridicarea* efectivă a *(cel puțin) unui nivel* de dependență clauzală, şi este reprezentată de marcheri precum *"atuncf, "altfef* etc.

M32 = {marcheri (de discurs) inter-clauzali care introduc dependență de *supra-ordonare,* incluzând *semnele de punctuație* precum două puncte, punct-şi-virgulă, paranteză închisă, linie-de-despărțire etc.}. *Supra-ordonarea* presupune *ridicarea* unuia sau mai multor nivele de dependență clauzală, sau rămânerea pe acelaşi nivel de dependență în cadrul unei dependențe de tip-coordonare. Exemple tipice de marcheri din clasa M32 sunt *"daf\ "aşadaf, "chiar", "lajeljde)", "în__comparațieJcu)"* etc.

M31 = {marcheri (de discurs) inter-clauzali care introduc unul sau mai multe nivele de dependență de *sub-ordonare,* incluzând *semne de punctuație* ca paranteza deschisă, linia-de-despărțire etc.} Aceasta este o clasă largă de marcheri de discurs formată din numeroase tipuri de relații între clauze: logice, sintactice, semantice, pragmatice etc.

Aşa cum a fost menționat mai sus, fiecare dintre clasele M33, M32 şi M31 poate, la rândul ei, să fie partiționată în subclase care conțin marcheri de tip relațional (exprimați prin corelație), ce stabilesc relații între clauze, sau ca funcții de clauze (cu cel puțin două argumente).

M2 = { *marcheri care introduc o clauză (finită sau infinită), sau un grup sintagmatic al cărui nucleu semantic este una din categoriile sintactice predicaționale N, V, A* }. Compusul sintactic (sau *grupul sintagmatic* în termenii [3]) XG, X = N, V, A, poate fi asimilat unei clauze degenerate, infinite (vezi Tabelul 4.1) în cazul X = N, A.

M2 este divizată în următoarele subclase (în ordinea descrescătoare a priorității de introducere a relațiilor de dependență):

M25 = {marcheri care introduc *clauza relativă}.*

Explicația constă în faptul că o clauză relativa reprezintă cea mai complexă unitate sintagmatică ce joacă rol de modificator, şi care se aplică nucleului NG al clauzei relative:

M24 = {apariția unui *grup verbal finit* (FVG) sau, echivalent, apariția valorii FINITE pentru trăsătura TENS atribuită unui verb, introducând deci o *clauză finită}.*

întregul grup verbal poate moşteni valoarea trăsăturii FINITE dacă nucleul său V sau altă componentă importantă din VG poartă această valoare a trăsăturii TENS (de exemplu, auxiliarul din VG).

M23 = {apariția unei sintagme *predicaționale* XG (sau X1), X=V, N, A, al cărei nucleu semantic este o *categorie predicațională,* purtând valoarea PRED = ACT (posibil încă la nivel de lexicon), şi introducând astfel o *clauză infinită}.*

Clasele de marcheri M24 şi M23 introduc structuri de nivel-X2, şi anume clauze finite sau infinite, formate dintr-o sintagmă X1 (sau grup XG, X = N, V, A) care reprezintă *nucleul semantic, finit* (TENS = FINITE) sau *predicațional* (PRED = ACTional), al structurii de nivel-X2, urmată de sateliți (argumente şi/sau adjuncți) corespunzători de tip NG (inclusiv NG-uri prefixate de o prepoziție, deci clasica sintagmă PP). Unele dintre argumente, cum este cazul *subiectului gramatical,* pot preceda nucleul semantic de tip X1 al clauzei căreia îi aparțin [3]. Să mai precizăm că există o *ordine sistemică (canonică)* [18], [19], a sateliților, sau *"actanților"* (argumente şi adjuncți) dintr-o clauză (finită sau infinită): ACT(or), PAT(ient), ADDR(essee), ORIG(ine), LOC(ation) etc. Ordinea canonică este specifică fiecărui LN, si se poate obține în urma unei cercetări statistice si lingvistice foarte atente [18].'

Putem găsi recent un *principiu de predicativitate* similar cu cel folosit în *strategia lingvistică* SCD, şi aplicat la sintagmele nominale din limba italiană [14] sau la adjectivele *"deverbale"* [14], [16]. în timp ce predicativitatea verbelor este frecventă şi naturală, trăsătura de *nepredicativitate* [17, p. 22] (de fapt nepredicaționalitate) a verbelor de tip *existențial* este şi ea la fel de frecventă (formele lui "a *fi'),* valoarea lor FINITE, dublată sau nu de valoarea trăsăturii PRED = ACT, anunțând totuşi apariția unei clauze finite.

M22 = {marcheri care introduc relații de tip-JOIN, adică *conjuncții* de tipul V. *"sau", "lajel_ca_(şi)", "împreunăjcu)"}.*

M21 ={ COMMA (sau VIRGULA) }.

Clasele M22 şi M21 cuprind marcheri cu un grad important de ambiguitate deoarece pot introduce orice structură de tip X1 (grupuri XG, X = N, V, A) sau X (clauze finite sau infinite).

M1 = { *marcheri care delimitează (introduc) structuri XG }.*

Conform strategiei SCD şi teoriei FX-bar [3], [23], clasa de marcheri M1 constă în *marcheri de* r?/Ve/-X1, X = N, V, A, adică marcheri care se aplică construcţiilor sintactice de *nivel-X'*\ (denotat şi XG, şi numit X-grup). Aceste sintagme constau, de fapt, dintr-un *nucleu semantic* înconjurat de *modificatori* (adjective sau adverbe) şi/sau *specificatori* (sau *cuantificatori,* unii generalizaţi, printre cuantificatori incluzându-se determinatorii, negaţia etc).

Aşa cum există o ordine sistemică a sateliţilor unui nucleu semantic într-o clauză (sintagmă de nivel-X2), în mod similar există o *"ordine structurală",* dată de *"distanţa"* modificatorilor, cuantificatorilor, prepoziţiilor etc. faţă de nucleul XO, pentru constituenţii unei sintagme de nivel-X1. Astfel, în limba română (franceză, engleză), cel mai "apropiat" faţă de nucleul XO trebuie să fie *modificatorul* (adjectivul sau adverbul), urmează apoi *cuantificatorul* (care ocupă locul modificatorului dacă acesta lipseşte), apoi prepoziţia (ad-poziţia, în general) etc. De exemplu, nu este sintactic corectă sintagma *frumos orice copil*\ sau *"orice frumos pe copif.* Nucleul XO înconjurat de modificatori şi/sau specificatori (cuantificatori) poate fi marcat funcţional prin *pre-poziţii* (în cazul grupului nominal NG din română, engleză, franceză), dar şi prin *post-poziţii* (în cazul NG sau VG din engleză sau germană). Marcarea clitic-funcţională (prin particule pre- sau *post-poziţionale)* poate exprima *cazul* (pentru NG), sau *timpul, semantica* (pentru VG) etc. Principalele elemente componente ale unei structuri XG corespund şi subclaselor de marcheri ai clasei M1.

M1 poate fi divizată în subclase de marcheri, subclase utile în delimitarea substructurilor XG (X1), X = N, V, A, în conformitate cu un criteriu cum este *distanţa* dintre nucleul semantic XO şi elementele funcţionale care îl *"înconjoară"',* un asemenea nucleu este, în ultimă instanţă, un substantiv comun obiectual (numit şi *autosemanticîn* [19]), un nume propriu, sau un substantiv personalizat (dar fără nume propriu, denominalizat).

M14 = {apariţia unui substantiv comun *obiectual* (nepredicaţional, autosemantic), a unui nume propriu, sau a unui substantiv personalizat denominalizat}

M13 = {apariţia unui *modificator* (adjectiv, adverb, adjectiv pronominal)}

M12 = {apariţia unui *cuantificator* (generalizat)}

M11 = { pre-poziţii sau *post-poziţii* exprimând *cazul* (pentru N), timpul sau *semantismul* (pentru V) etc }

Ultima clasă de marcheri, notată MO (sau M00 pentru uniformitate), şi ai cărei marcheri se aplică *formei de dicţionar a* cuvântului, este reprezentată de *rolul funcţional* al *flexionării.*

Recapitulând, clasele de marcheri considerate de strategia lingvistică SCD, în particular de *algoritmul de segmentare* SCD-2002, pot fi reprezentate grafic de următoarea ierarhie [7]:

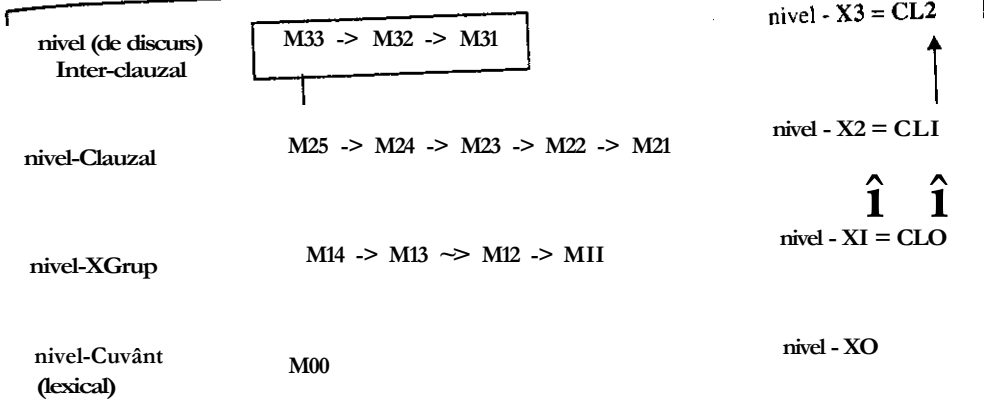


Figura 4.1.1. Clasele de marcheri SCD-2002 şi ierarhia lor

Orientarea arcelor din Fig. 4.1.1., stabilite între clasele şi subclasele de marcheri, provine dintr-o *ordine* de *prioritate descrescătoare* între marcherii consideraţi, şi este reprezentată mai jos prin relaţia ">" dintre clasele şi subclasele de marcheri. Această ierarhie este o *ipoteză de bază* impusă în *strategia lingvistică* SCD şi, prin consecinţă, şi în *algoritmul de segmentare SCD [7].*

(4.1.2) VQ = 1+4) M(k+1)(i+1) * M(k+1)j (k = 0+2);

(4.1.3) ᵥ(k = 0+2) M(k+1)i > Mkj 0 = 1+5), Q = 0+5).

Aceste inegalităţi ne spun că marcherii din subclasa M(k+1)(j+1) sunt de *prioritate mai mare* în comparaţie cu marcherii din subclasa M(k+1)j, (k = 0+2), (j = 1+4), în cadrul aceleiaşi clase M(k+1) de marcheri aflată pe *acelaşi nivel* de proiecţie lingvistică, iar marcherii din aceeaşi clasă M(k+1) au o *prioritate mai mare* faţă de marcherii din clasa Mk de pe nivelul *inferior de* proiecţie lingvistică.

Această ierarhie a marcherilor şi claselor de marcheri este considerată de noi ca fiind validă pentru *limba română.* Probabil că anumite modificări vor fi necesare când se trece de la un LN la altul. Dacă ne situăm în domeniul mai restrâns al limbajelor indo-europene (cum sunt franceza, engleza, germana, italiana, spaniola, posibil rusa), se poate aprecia că structurile şi clasele de marcheri propuse în Tabelul 4.1. şi Fig. 4.1.1. rămân aceleaşi sau foarte asemănătoare, cu anumite modificări parametrizate în funcţie de limbaj.

4.2. Algoritmul SCD-2002 de segmentare şi stabilire a dependenţelor

Urmând algoritmii de segmentare şi dependenţă (numiţi şi *meta-algoritmi* SCD) propuşi în [1] şi [2] (denotaţi în continuare SCD-1994), rafinaţi cu clasele de marcheri considerate în subsecţia precedentă, se obţine forma prezentă a *algoritmului de segmentare SCD* (denotată SCD-2002), conform [7]. Dezvoltăm aici forma *secvenţial-liniară* a acestui algoritm, însă în [1] sunt expuse şi o formă *secvenţial-recursivă,* ca şi o versiune *paralelă* a algoritmului. O formă *"inversată"* (pentru care în intrare avem un arbore de derivare sau o formulă logică, iar în ieşire - ca şi în intrarea în algoritmul standard - avem o frază) poate fi folosită pentru a ghida procesul de generare a unei fraze de LN [2], schimbând operaţia de recunoaştere a marcherilor cu cea de generare a lor, şi analiza (parsarea) compuşilor sintactici cu generarea lor.

în *descrierea algoritmului* de segmentare SCD-2002 sunt folosite câteva *operaţii* al căror înţeles este bine să fie precizat de la început.

(4.**2**.a) *Recunoaşterea marcherilor* înseamnă inserarea în text a unor etichete adecvate, ce corespund marcherilor care realizează delimitarea unităţilor textuale sintactice, semantice, şi de discurs.

(4.**2**.b) *Verificarea marcherului* înseamnă preluarea, din baza de date a marcherilor, a celor mai importante valori din structura de trăsături a acelui marcher.

(4.**2**.c) *Segmentarea* implică o analiză liniară (parsare) a secvenţei de etichete de marcheri, şi *recunoaşterea* unei subsecvente (eventual discontinuă) care face parte din secvenţa originală de etichete de marcheri.

(4.**2**.d) *Recunoaşterea structurii sintactice* înseamnă *segmentarea* şi *recunoaşterea* structurilor sintactice elementare cum sunt NG, VG, AG, clauza infinită, şi clauza finită.

(4.**2.e)** *Compunerea structurilor (de dependenţă)* constă în stabilirea dependenţelor (sub-ordonare, co-ordonare, supra-ordonare). succesive dintre structurile sintactice recunoscute, pe baza rolului funcţional specific al marcherilor care delimitează aceste structuri, şi utilizând ierarhia corespunzătoare dintre clasele cărora le aparţin aceşti marcheri (vezi Fig. 4.1.1. şi relaţiile 4.1.2.-4.1.3.).

Algoritmul de segmentare SCD-2002

StepOI. Recunoaşterea pe text a marcherilor din clasa M3;

StepOI. Recunoaşterea pe text a marcherilor din clasa M2;

Step03. Verificarea contextuală şi recunoaşterea apariţiei corelate a marcherilor de tip M3 şi M2[1];

Step04. Segmentarea frazei în clauze finite;

Step05. Segmentarea (chunking), dacă este necesar, a clauzelor finite în clauze infinite;

[Stop: Dacă scopul procesării este de a obţine o structură liniară a clauzelor finite şi/sau infinite din frază].

Step06. Verificarea marcherilor M3 şi stabilirea relaţiilor de dependenţă inter-clauzală[2];

[Stop: Dacă scopul procesării este doar de a obţine arborele de dependenţă a clauzelor finite (şi infinite) din frază].

Step07. Recunoaşterea pe text a marcherilor din clasa M1;

Step08. Verificarea contextuală şi recunoaşterea (eventualei apariţii corelate) a marcherilor M1[3];

Step09. Recunoaşterea structurilor XG (X = N, V, A)[4];

Step 10. Verificarea marcherilor M24 şi M23, şi stabilirea relaţiilor de dependenţă dintre structurile infinite, intra-clauzale de tip XG[5];

[Stop].

Indicii superiori **(n)** care apar în algoritmul de mai sus corespund următoarelor *remarci:*

(1) Marcherii corelaţi pot fi reprezentaţi ca upluri ordonate (liste) de marcheri.

(2) Relaţiile de dependenţă clauzală pot fi stabilite (ca în [8, Anexa 9, p. 108], de exemplu) prin utilizarea unei *gramatici formale* (ambigue) definită pe secvenţe de marcheri din (sub)clasele M3, M25, M22, şi M21.

(3) Marcherii complecşi pot fi sintagme sau expresii de tipul gradelor de comparaţie a adjectivelor, diferiţi cuantificatori generalizaţi etc.

(4) în execuţia acestui pas se realizează parsarea sintagmelor XG dintr-o clauză finită şi infinită.

(5) Dependenţele dintre structurile de tip XG sunt stabilite în principal prin utilizarea trăsăturilor şi valorilor de trăsături TENS = FINITE sau INFINITE, şi PRED = AC sau EXIST, pe care le posedă nucleele semantice ale sintagmelor XG, X = N, V, A (a se vedea [3], [23]). Aceste valori pot fi *moştenite* din reprezentarea din lexicon a cuvintelelor care poartă aceste trăsături şi care formează XG, sau pot fi *dobândite* de către nucleul semantic al XG în procesul de recunoaştere (parsare) a structurii.

5. Compararea algoritmilor de segmentare

5.1. Algoritmii de segmentare SCD-1994 şi SCD-2002

Algoritmii SCD-1994 expuşi în [1], [2] se bazează pe *patru* (sub)clase principale de marcheri, denotate acolo prin (clasele de) **"1**-marcheri" până la **"4-**marcheri". Aceste subclase de marcheri din SCD-1994 corespund următoarelor (sub)clase de marcheri din prezentul algoritm SCD-2002 [7]:

(5.1.1) **1**-marcheri = M3 u M25 u M22;

2-marcheri = M24;

3-marcheri = M23;

4-marcheri = M21 u M1

Prezentăm în continuare algoritmul de segmentare SCD-1994 (în forma secvențial-recursivă), aşa cum a fost expusă în [1, p.68-69], având ca scop parsarea LN. Algoritmul SCD-1994 (în forma secvențial-liniară) şi destinat sarcinii de *generare* a LN este prezentat în [2, p. 172-173].

Algoritmul de segmentare SCD-1994 în formă secvențial-recursivă (SR)

StepOI. Recunoaşterea marcherilor de clauză.

Step02. Recunoaşterea sintagmelor VG (grupuri verbale) finite şi infinite.

Step03. Verificarea contextuală a marcherilor.

Step04. Segmentarea clauzală.

Step05. Segmentarea sub-clauzală.

Step06. Recunoaşterea **1**-marcherului;

Recunoaşterea **1**-structurii:

Wait-until **1**-structura este completă.

Step07. Recunoaşterea **2**-marcherului;

Recunoaşterea **2**-structurii:

Wait-until structura de nivel-X2 este completă*.

Step08. Recunoaşterea **3**-marcherului;

Recunoaşterea **3**-structurii.

Step09. Recunoaşterea **4**-marcherului;

Procesarea **4**-structurii.

StepIO. **3**-structură completă?

Nu: *Go-to* Step08.

Da: Compune 3-structuri; Go-fo Stepl 1.

Step11. **2**-structură completă ?

Nu: *Go-to* Step07.

Da: Compune **2**-structuri; *Go-to* Stepl 2.

Stepl 2.1-structură completă ?

Nu: *Go-to* Step06.

Da: Compune **1**-structuri; *Go-to* Stop.

Stop.

* *Structuri* AX-bar (în original, în [1]), înțelegând structuri sintactic derivate din *schemele* X-bar *augmentate,* definite în [20] şi extinse în [3 Scopul acestui pas al algoritmului este de a completa clauza finit introdusă printr-un grup verbal finit.

Principala problemă cu algoritmul de segmentare şi dependență SCD 1994 (forma SR) este că sunt necesare *"multiple nivele de recursie pentr a completa şi compune structurile"* [1, p.69].

5.2. Algoritmii de segmentare M-1997 şi SCD-2002

în această subsecțiune vom arăta că algoritmul de segmentare M-199 este *scufundat* în algoritmul SCD-2002 (de fapt, şi în SCD-1994) [7].

M-1997 este un algoritm de *"suprafață"* destinat segmentării discursului î unități textuale de tip-clauzal. în timp ce, pentru acest scop, M-1997 foloseşt numai *marcheri de discurs {"cue phrases"* sau *conective),* algoritmul SCD-200 utilizează un set de clase de marcheri mai larg şi în acelaşi timp mai rafinat, s care include clasele de marcheri din M-1997 ca un caz particular. Mai precis relațiile dintre clasele de marcheri Mari, Mar2, şi Mar3 (vezi Secțiunea 2) utilizat pentru M-1997, şi clasele de marcheri Mkj ale algoritmului SCD-2002 su următoarele:

(5.2.1) Mari u Mar2 u Mar3 c M3 u M25 u M22 u M21

sau, posibil, mai precis:

(5.2.2) Mari u Mar3 c M3 u M25 *şi* Mar2 c M22 u M21

Diferența dintre algoritmii M-1997 şi SCD-2002 nu constă doar în faptul c al doilea algoritm are un număr mai mare de clase, care sunt mai fine (m precise), ci, mai important este faptul că aceste clase formează un *sistem ierarh* (expus în Fig. 4.1.1.) ce este utilizat în procesele de *segmentare* şi de *stabilire dependențelor.* SCD-2002 furnizează noi clase de marcheri, cum sunt M23 şi M2 (apariția categoriilor predicaționale şi/sau având un timp finit), precum şi clasa M

cu subclasele sale (apariția unor componente ale sintagmei XG, X = N, V, A). Acesta este un *prim argument* din care rezultă că M-1997 este *scufundat* în SCD-2002. *"Scufundarea"* este un termen care reflectă, de fapt, un proces de rafinare şi de creştere a preciziei în calculul marginilor (limitelor) unităților textuale şi a dependențelor dintre ele, pentru SCD-2002 în comparație cu M-1997.

Al doilea argument important care susține validitatea relației afirmate între cei doi algoritmi este următorul: fiecare *acțiune* din M-1997 are un corespondent într-o *operație* (sau o mulțime de *operații)* din algoritmul SCD-2002 (subsectiunea 4.2).

Pentru segmentare, algoritmul M-1997 asociază fiecărui marcher, în baza de date a marcherilor, o anumită *acțiune* ce este statistic determinată de către analiza de corpus efectuată în [4], Corespondența dintre operațiile algoritmul-SCD, şi o *acțiune* din algoritmul-M, se face în felul următor:

(5.2.a) *Acțiunea* (şi marcherii) NORMAL din algoritmul-M are acelaşi efect cu operațiile de procesare a marcherilor de discurs din clasa M3 a algoritmului-SCD. Când este întâlnit un asemenea marcher, aceasta înseamnă că o clauză (în SCD-2002) sau o unitate de tip-clauzal (în M-1997) este pe cale de a se încheia şi o altă clauză, respectiv unitate de tip-clauzal, este probabil că va începe.

(5.2.b) *Acțiunile* COMMA, SET_AND, şi SET_OR din algoritmul-M sunt folosite pentru a dezambiguiza rolul unor marcheri din M3 pentru care nu se poate aplica întotdeauna regula generală *(acțiunea* NORMAL). Aceşti marcheri sunt următorii pentru limba română:"," [Eng: *comma], "şi*\ şi *"sau".* Rolul acestor marcheri este ambiguu deoarece comportamentul lor nu este uniform în cadrul delimitării unităților textuale. SCD-2002 rezolvă aceste cazuri cu ajutorul utilizării unei gramatici formale de marcheri care descrie principalele reguli de delimitare şi dependență a clauzelor (în limba română). Această gramatică (vezi indicele superior **(2)** din SCD-2002 şi remarca corespunzătoare) are ca scop să recunoască secvențele cele mai frecvente de marcheri din clasele M3 şi M2. Numai *câteva* dintre aceste reguli sunt incorporate în mod explicit în algoritmul M-1997 original.

(5.2.c) O unitate de tip-clauzal din M-1997 nu este în mod necesar o clauză finită în sensul gramatical al noțiunii, aşa cum este folosit în algoritmul-SCD. O asemenea unitate de tip-clauzal, în sens M-1997, poate fi o întreagă frază, formată din mai multe clauze finite. M-1997 foloseşte, de fapt, pentru segmentarea liniară a frazei în unități de tip-clauzal numai *trei reguli* din cele folosite de SCD-2002, iar aceste reguli sunt sintetizate de către *acțiunile* COMMA, SET_AND, SET_OR.

(5.2.d) *Acțiunile* MATCH_PAREN, MATCH_DASH, COMMA_PAREN sun utilizate de către M-1997 pentru a delimita acele întinderi de tex cate pot fi omise atunci când fraza este segmentată în unități de tip-clauzal. Aceste părți "explicative" din text, considerate a nu f importante, sunt, în text, puse între *paranteze,* (perechi de) *liniuțe de-despărțire,* sau (perechi de) *virgule.* Algoritmul M-1997 nu tratează aceste întinderi "parantetizate" de text ca fiind unități de tip-clauzal propriu-zise, ci le consideră doar ca fiind scufundate în unitatea de tip-clauzal de care aparțin. Pentru SCD-2002, aceste *acțiuni* M-1997 nu au un corespondent specific deoarece *paranteza* (închisă şi deschisă), *virgula,* şi *liniuța-de-despărțire* sunt tratate ca marcheri de discurs (M3), şi fac parte din *gramatica de marcheri compuşi* (concatenați) care este asociată cu algoritmul SCD-2002 de segmentare şi dependență a clauzelor dintr-o frază.

(5.2.e) Din acelaşi motiv ca cel menționat mai sus, în (5.2.d), *acțiunil* DUAL, NORMAL_THEN_COMMÂ din M-1997 nu au, nici ele, u corespondent explicit în SCD-2002; aceste două acțiuni sunt d asemenea înglobate în *gramatica formală* de *secvențe de marche de discurs,* care se dovedeşte a fi, în mod clar, mai generală, uşo de extins (sau de restrâns), este dependentă de LN specifi analizat, şi modelează comportamentul marcherilor simpli compuşi (concatenați) de tip M3 şi M2.

Relațiile (5.2.1-2) şi observațiile (5.2.a-e) demonstrează că algoritmul d segmentare M-1997 este (chiar strict) scufundat în algoritmii-SCD (atât SCD-200 cât şi SCD-1994) [7]. Acest fapt, stabilit teoretic aici, este confirmat de cătr rezultatele empirice ale implementărilor, prezentate în subsectiunea care urmează

5.3. Execuția segmentării pentru algoritmii M-1997 şi SCD-2002

Actuala etapă de implementare a algoritmilor de segmentare [7] es prezentată în exemplele care urmează. *Step06* din SCD-2002, şi *Step12* din SCD 1994 stabilesc relațiile de *dependență inter-clauzală,* folosind o gramatică forma pentru marcherii de discurs, simpli şi compuşi (concatenați), din clasele M3 şi M Această fază a algoritmului nu este încă implementată, în prezent. Să menționă că stabilirea *dependențelor intra-clauzale* este (parțial) implementată pr utilizarea, pentru moment, (numai) a subclaselor M2 şi M1 de marcheri. Margini inter-clauzale din text sunt reprezentate prin *paranteze pătrate,* în timp ce pent marginile şi dependențele intra-clauzale sunt folosite *parantezele rotun* (obişnuite). Indicii inferiori ai parantezelor pătrate arată numărul curent al unității textuale de tip-clauzal din algoritmul M-1997, respectiv numărul curent al clauz obținute din algoritmul SCD-2002.

Exemplul 5.3.1.

Ex.5.3.1.Tag. (Etichetarea morfologică realizată cu mediul *TexTag* - vezi Fig. 5.4.1. şi Fig. 5.4.2.)

<NSRY,23,0>Câmpul</NSRY,23,0> <V3,24,0>era verde</V3,24,0> <CR,25,0>şi</CR,25,0> <NSRY,26,0>vița</NSRY,26,0> <S,27,0>de</S,27,0> <NSRN,28,0>vie</NSRN,28,0> <PXA,29,0>se</PXA,29,0> <V3,30,0>acoperise<A/3,30,0> <S,31,0>cu</S,31,0> <NPN,32,0>lăstari</NPN,32,0> <APN,33,0>verzi</APN,33,0><COMMA,34,0>,</COMMA,34,0> <NPRY,35,0>copacii</NPRY,35,0> <S,36,0>de pe</S,36,0> <NSRY,37,0>marginea</NSRY,37,0> <NSOY,38,0>şoselei</NSOY,38,0> <V3,39,0>înfrunziseră</V3,39,0> <CR,40,0>şi</CR,40,0> <NSRY,41,0>briza</NSRY,41,0> <V3,42,0>sufla</V3,42,0> <S,43,0>dinspre</S,43,0> <NSRN,44,0>mare</NSRN,44,0><POINT,45,0>.</POINT,45,0>

Ex.5.3.1.Mar. (Rezultatul segmentării (fără dependențe), obținut prin aplicarea algoritmului M-1997 în cadrul mediului *ClauSEGM* - vezi Fig. 5.4.3.)

[Câmpul era verde şi vița de vie se acoperise cu lăstari verzi, copacii de pe marginea şoselei înfrunziseră şi briza sufla dinspre mare.]i

Ex.5.3.1.SCD. (Rezultatul segmentării (fără dependențe), obținut prin aplicarea algoritmului SCD-2002 în cadrul mediului *ClauSEGM* - vezi Fig. 5.4.4.)

[(Câmpul) era verde ^ şi[(vița) (de (vie)) se acoperise (cu (lăstari) (verzi)) *h ,[* (copacii) (de pe (marginea (şoselei))) înfrunziseră]₃ şi[(briza) sufla (dinspre (mare)).]₄

Exemplul 5.3.2.

Ex.5.3.2.Tag.

<S,1,0>în</S,1,0> <NSN,2,0>întuneric</NSN,2,0> <V2,3,0> ai fi zis</V2,3,0> <C,4,0>că</C,4,0> <V3,5,0>fulgeră <A/3,5,0> <R,6,0>ca</R,6,0> <NSRY,7,0>vara</NSRY,7,0> <COMMA,8,0>,</COMMA,8,0> <C,9,0>dar</C,9,0> <NPRY, 10,0>nopțile</NPRY,10,0> <V3,11,0>erau reci<A/3,11,0> <CR,12,0>şi</CR,12',0> <QZ,13,0>nu</QZ,13,0> <PPSD, 14,0>ți</PPSD,14,0> <PXA,15,0>se</PXA,15,0> <V3,16,0> părea</V3,16,0> <R,17,0>deloc</R,17,0> <C,18,0>că</C, 18,0> <PXA,19,0>se</PXA,19,0> <V3,20,0>apropie</V3,20,0><NSRY,21,0>furtuna</NSRY,21,0><POINT,22,0>.</POINT,22,0>

Ex.5.3.2.Mar.

[în întuneric ai fi zis *Ji* [că fulgeră ca vara,]₂ [dar nopțile erau părea deloc]₃ [că se apropie furtuna.]₄

EX.5.3.2.SCD.

[(în (întuneric)) ai fi zis]i [că fulgeră (ca (vara))]₂, [dar (nop]₃ şi [nu (ți) se părea (deloc)]₄ [că se apropie (furtuna).]₅

Exemplul 5.3.3.

Ex.5.3.3.Tag.

<NSRY,46,0>Poarta</NSRY,46,0> <V3,47,0>era deschis <COMMA,48,0>,</COMMA,48,0> <TSR,49,0>un</TSR,49,0> <NS </NSN,50,0> <V3,51,0>şedea<A/3,51,0> <S,52,0>la</S,52,0> <NSN /NSN,53,0> <S,54,0>pe</S,54,0> <TSR,55,0> o</TSR,55,0> bancă</NSRN,56,0><COMMA, 57,0>, </COMMA,57,0> <TSR,58,0 <NSRN, 59,0>ambulanță</NSRN,59,0> <V3,60,0>aştepta<A/3,60, </S,61,0> <NSRY, 62,0>uşa</NSRY,62,0> <S, 63,0>de</S,63,0> serviciu</NSN,64,0> <CR, 65,0>şi</CR,65,0> <VG,66,0>intr <V1,67,0> am simțit</V1,67,0> <NSRY,68,0>mirosuK/NSRY,68,0> pardoselii</NSOY,69,0> <S,70,0>de</S,70,0> <NSRN,71,0>marmură <S,72,0>şi</S,72,0> <S,73,0>de</S,73,0> <NSN,74,0>spital</NSN 75,0>.</POINT,75,0>

Ex.5.3.3.Mar. (întindere de text între paranteze acolade {...}

[Poarta era deschisă, {un soldat şedea la soare pe o bancă aştepta la uşa de serviciu şi intrând am simțit mirosul pardoselii de spital.

Ex.5.3.3.SCD.

[(Poarta) era deschisă]\ ,[(un (soldat)) şedea (la (soare) (p k ,[(o (ambulanță)) aştepta (la (uşa) (de (serviciu)))]₃ şi[intrând am (pardoselii)) (de (marmură) (şi (de (spital)))).]₄

Exemplul 5.3.4.

Ex.5.3.4.Tag.

<NPRY,1,0>Trupele</NPRY,1,0> <V3,2,0>treceau<A/3,2, lângă</S,3,0><NSRN,4,0>casă</NSRN,4,0><COIv1MA,5,0>,</COM 6,0>pe</S,6,0> <NSRN,7,0>şosea</NSRN,7,0><COMMA,8,0>, <CR,9,0>şi</CR,9,0> <NSRY,10,0>praful</NSRY,10,0> <RELO,1 R E L O . H . O x Z . ^ . O ^ / Z . ^ . O x P P S A . I S . O H ^ P P S A . I S ^ <V V3,14,0> <PXA,15,0>se</PXA,15,0> <V3,16,0>astemea<A/3,16,

</S,17,0> <NPRY,18,0>frunzele</NPRY,18,0> <NPOY,19,0>copacilor</NPOY,19,0><POINT,20,0>.</POINT,20,0>

Ex.5.3.4.Mar.

[Trupele treceau pe lângă casă, pe şosea, şi praful^ [pe care-l ridicau se aşternea pe frunzele copacilor.]$_2$

Ex.5.3.4.SCD. (clauză relativă - atributivă)

[(Trupele) treceau (pe lângă (casă)) , (pe (şosea)) *]t* ,[şi (praful) [pe care-(l) ridicau se aşternea (pe (frunzele (copacilor)))]$_2$.]$_3$

Exemplul 5.3.5.

Ex.5.3.5.Tag.

<QZ,76,0>Nu</QZ,76,0> <PPSA,77,0>m</PPSA,77,0><Z,78,0>-</Z,78,0> <V3,79,0>a văzut<A/3,79,0> <CR,80,0>şi</CR,80,0> <QZ,81,0>n</QZ,81,0><Z,82,0>T,</Z,82,0><V1,83,0>am ştiut</V1,83,0> <C,84,0>dacă</C,84,0> <V3,85,0>e</V3,85,0> <NSRY,86,0>cazuK/NSRY,86,0> <C,87,0>să</C,87,0> <PPSA,88,0>mă</PPSA,88,0> <V1,89,0>duc<A/1,89,0> <S,90,0>la</S,90,0> <PPS,91,0>el</PPS,91,0> <C,92,0>să</C,92,0><Z,93,0>-</Z,93,0><PPSA,94,0>i</PPSA,94,0> <V1,95,0>raportez<A/1,95,0> <C,96,0>că</C,96,0> <V1,97,0>am sosit</V1,97,0> <C,98,0>sau dacă</C,98,0> <QZ,99,0>nu</QZ,99,0> <V3,100,0>e mai bine</V3,100,0> <C,101,0>să</C,101,0> <PPSA,102,0>mă</PPSA,102,0> <V1,103,0>duc<A/1,103,0> <C,104,0>să</C,104,0> <PPSA,105,0>mă</PPSA,105,0> <V1,106,0>aranjez<A/1,107,0>putin</R,107,0><POINT,108,0>.</POINT,108,0>

Ex.5.3.5.Marc.

puţin.]$_8$ J 1— "u o . . w l u i n e Je isa ma duc]$_7$ [să mă aranjez

Ex.5.3.5.SCD.

[Nu (m)-a văzut *]i* şi[n-am ştiut]$_2$ [dacă e (cazul)]$_3$ [să (mă) duc (la (el))]$_4$ [să-(i) raportez]$_5$ [că am sosit]$_6$ [sau dacă nu e mai bine]$_7$ [să (mă) duc]$_8$ [să (mă) aranjez (puţin),]$_9$

Exemplul 5.3.6.

Ex.5.3.6.Tag.

<NSRY,109,0>Fereastra</NSRY,109,0> <V3,110,0>era deschisă<A/3,110,0><COMMA,111,0>,</COMMA,111,0> <NSRY,112,0>patuK/NSRY,112,0> <PSS,113,0>meu</PSS,113,0> <V3,114,0>era acoperit<A/3,114,0> <S,115,0>cu</S,115,0> <NSRY,116,0>pătura</NSRY,116,0><COMMA,117,0>,</COMMA,117,0> <NSRY,118,0>masca</NSRY,118,0> <S,119,0>de</S,119,0> <NPRY,120,0>gaze</NPN,120,0> <S,121,0>cu</S,121,0> <NSRY,122,0>cutia</NSRY,122,0> <PSS,123,0>ei</PSS,123,0> <ASN,124,0>lunguiaţă</ASN,124,0> <S,1[illegible]0>de</S,125,0> <NSRN,126,0>tinichea</NSRN,126,0> <CR,127,0>şi</CR,1[illegible]0> <NSRY,128,0>casca</NSRY,128,0> <S,129,0>de</S,129,0> <NSN,130,0>oţel</NSN,130,0> <V3,131,0>erau agăţate<A/3,131,0> <S,132,0>pe</S,132,0> <DMSR,133,0>acelaşi</DMSR,133,0> <NSN,134,0>cuier</NSN,134,0><POINT,135,0>.</POINT,135,0>

Ex.5.3.6.Mar. (întindere de text între paranteze acolade {...})

[Fereastra era deschisă, {patul meu era acoperit cu pătura}, masca gaze cu cutia ei lunguiaţă de tinichea şi casca de oţel erau agăţate pe ace cuier.]i

EX.5.3.6.SCD.

[(Fereastra) era deschisă]i ,[(patul) (meu) era acoperit (cu (pătura)) (masca) (de (gaze) (cu (cutia) (ei (lunguiaţă)) (de (tinichea)))) şi (casca) (de (oţ erau agăţate (pe (acelaşi (cuier))).]$_3$

Exemplul 5.3.7.

Ex.5.3.7.**Tag.**

<V1,1,0>Aş vrea<A/1,1,0> <C,2,0>să</C,2,0><Z,3,0>-</Z,3,0><PPSD,4,0>ţi</PPSD,4,0> <V1,5,0>spun<A/1,5,0> <C,6,0>că</C,6,0> <CR,7,0>şi</CR,7,0> <R,8,0>mai</R,8,0> <R,9,0>târziu</R,9,0><COMMA,10,0>,</COMMA,10,0> <CR,11,0>şi</CR,11,0> <S,12,0>într</S,12,0><Z,13,0>-</Z,13,0><ASN,14,0></ASN,14,0> <NSRN,15,0>parte</NSRN,15,0><COMMA,16,0>,</COMMA,16,0> <V1,17,0>am văzut<A/1,17,0> <C,18,0>că</C,18,0> <NPRY,19,0>lucruri</NPRY,19,0> <PXA,20,0>se</PXA,20,0> <V3,21,0>întâmplă<A/3,21,0> <R,22,0>tot aşa</R,22,0><COMMA,23,0>,</COMMA,23,0> <C,24,0>dar</C,24,0> <V3,25,0>ar fi nevoie</V3,25,0> <S,26,0>de</S,26,0> <PI,27,0>oarecari</PI,28,0> <NPN,28,0>precizări</NPN,28,0> <CR,29,0>şi</CR,29,0> <V1,30,0>sim V1,30,0> <C,31,0>că</C,31,0> <QZ,32,0>nu</QZ,32,0> <PPSD,33,0>mi</PPSD,33,0><Z,34,0>-</Z,34,0><V3,35,0>ar ajunge<A/3,35,0> <NSRY,36,0>respiraţia/NSRY,36,0><COMMA,37,0>,</COMMA,37,0> <C,38,0>că</C,38,0> <V1,39,0>ocoli<A/1,39,0> <R,40,0>prea</R,40,0> <R,41,0>mult</R,41,0><POINT,42,0>.</POINT,42,0>

Ex.5.3.7.**Mar.** (întindere de text între paranteze acolade {...})

[Aş vrea *]t* [să-ţi spun]$_2$ [că şi mai târziu, {şi într-altă parte,} am văzut]$_3$ lucrurile se întâmplă tot aşa,]$_4$ [dar ar fi nevoie de oarecari precizări şi simt j$_5$ nu mi-ar ajunge respiraţia,]$_6$ [că aş ocoli prea mult.]$_7$

Ex.5.3.7.SCD.

[Aş vrea h [să-(ți) spun]₂ [că şi (mai (târziu)) , şi (într-(altă (parte))) , am văzut]₃ [că (lucrurile) se întâmplă (tot aşa)]₄ ,[dar ar fi nevoie (de (oarecari (precizări)))]₅ şi[simt]₆ [că nu (mi)-ar ajunge (respirația)]₇ ,[că as ocoli (prea (mult))].]₈

5.4. Programele *TexTag* şi *ClauSEGM*

în cele ce urmează sunt prezentate câteva imagini de execuție în cadrul programelor *TexTag* şi *ClauSEGM,* scrise în Visual C++ 5.0, şi utilizate pentru a eticheta şi segmenta texte de LN (limba română) [7]. Figurile 5.4.1. şi 5.4.2. se referă la *TexTag,* Figura 5.4.3. conține execuția algoritmului de *segmentare M-*1997 în cadrul *ClauSEGM,* iar Figura 5.4.4. conține o execuție a algoritmului de *segmentare* SCD-2002 sub mediul *ClauSEGM.* Stabilirea relațiilor de *dependență inter-* şi *intra-clauzale,* pentru aceleaşi două tipuri de algoritmi, urmează să fie implementată în cadrul aceluiaşi mediu *ClauSEGM.*

Q e«08 > r«ftg*r Edfe*

~3 P"

Figura 5.4.1. Rezultatul etichetării morfologice sub *TexTag*

Figura 5.4,2. Lista de etichete selectată cu un meniu din *TexTag*

Figura 5.4.3. O execuție a algoritmului de segmentare *M*-1997 cu programul *ClauSEGM*

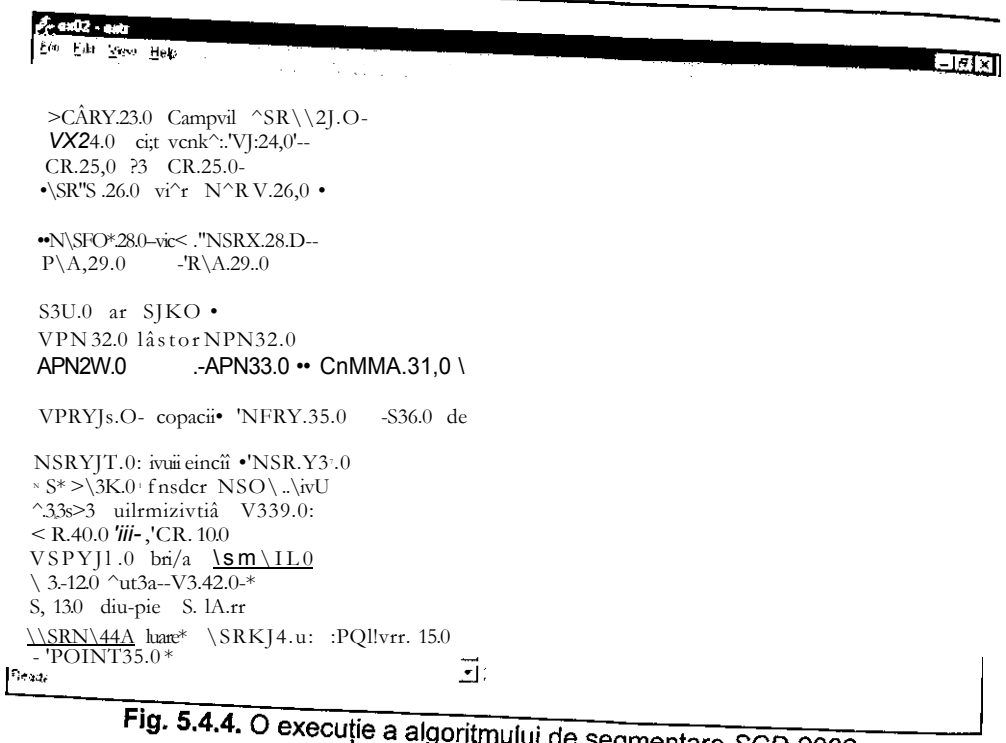


Fig. 5.4.4. O execuție a algoritmului de segmentare *SCD-2002* sub mediul *ClauSEGM*

6. Concluzii

Rezultatele obținute în această lucrare nu se referă strict la compararea şi implementarea celor doi algoritmi de segmentare. Avem, de fapt, două tipuri de algoritmi de segmentare (şi dependență), şi fiecare din cele două tipuri reprezintă linii specifice de cercetare, cu importante consecințe asupra domeniilor de procesare a LN cărora se adresează: algoritmul M-1997 [4] este destinat (teoriei şi) aplicațiilor de procesare a discursului, generare automată a LN, şi rezumării automate, în timp ce algoritmul SCD-2002 [7] .se încadrează mai curând în teorii sintactice ale LN, cum sunt teoria FX-bar [3], parsarea bazată pe teorii (principii) sintactico-semartice ale LN, dar şi punerea în evidență a structurilor (segmentelor) şi relațiilor de discurs [6].

Demonstrarea relației (de scufundare) dintre cele două tipuri de algoritmi de segmentare [7], schițarea (în secțiunea 1) a unui *cadru formal general* pentru *algoritmii de segmentare* a LN, în particular a segmentării de tip *chunking,* propunerea (în cadrul algoritmilor-SCD) unei metode generale de segmentare în unități textuale a LN şi de stabilire a dependențelor între ele, toate acestea constituie posibile noi perspective pentru abordările teoretice şi aplic în procesarea automată a LN, inclusiv, şi mai ales, pentru limba romî

Revenind la aspectele concrete expuse în acest artico algoritmilor către analiza complexă a structurilor semantico-discursive clasele de marcheri, şi perfecționarea actualelor implementări rămî direcții de continuare a prezentei abordări.

Referințe bibliografice

[I] Neculai Curteanu (1994). *From Morphology to Discourse Th Structures in the SCD Parsing Strategy,* Language and Akademia Libroservo, Prague, p. 61-73.

[2] N. Curteanu, G. Holban (1996). *Strategia lingvistică SCD aplicat generarea limbii române,* Limbaj şi Tehnologie (D. Tufiş, Academiei Române, p. 169-176.

[3] Neculai Curteanu (2000). *"Towards a Funcțional X-bar Theor* Report, Institute of Theoretical Informatics, Romanian Academy 32 p.

[4] Daniel Marcu (1997). *The Rhetorical Parsing, Summarization, a of Natural Language Texts,* Ph.D. Thesis, Univ. of Toronto, Cana

[5] Daniel Marcu (2000). *The Theory and Practice of Discourse Summarization,* The MIT Press, Cambridge.

[6] O. Popârda, N. Curteanu (2002). *L'evolution du discours jur analyse par la strategie linguistique SCD,* LINCOM Studies i Linguistics, LINCOM Europa, Munchen.

[7] N. Curteanu, C. Linteş (2002). *Segmentation Algorithms fo Textual Units,* Research Report, Institute of Theoretical Informati Academy.

[8] N. Curteanu, D. Cristea, P. Mihaescu (1982). *Cercetări în dome om-calculator prin intermediul limbajului natural.* Contract de 4774/1982, Universitatea Iaşi - ICI Bucureşti.

[9] Neculai Curteanu (1983). *Algoritmi de analiză sintactică a fraze românești.* Lucrările Conferinței INFO-IAŞI'83, p. 533-549.

[10] D. Cristea, N. Curteanu, P. Mihaescu (1983). *Implementar morfologic şi definitivarea proiectului de analiză sintactică.* cercetare nr. 1906/1983, Universitatea Iaşi - ICI Bucureşti.

[II] N. Curteanu (1984). *Aspecte ale analizei logice a limbajului nat* de cercetare nr. 4709/1984, Universitatea Iaşi - ICI Bucureşti.

[12] Rebecca Passonneau, Diane Litman (1997). *Intention-based segmentation: human reliability and correlation with linguistic cues,* in Proc. 31 th Annual Meeting of ACL, Ohio, p. 148-155.

[13] Lance Ramshaw, Michel P. Marcus (1999). *Text Chunking Using Transformation-based Learning,* in (S. Armstrong *et al.,* eds.) "Natural Language Processing Using Very Large Corpora", Kluwer Acad. Publ., p. 157-176.

[14] Victor Raskin, S. Nirenburg (1999). *"Lexical Rules for Deverbal Adjectives ,* in E. Viegas (Ed.) *Breadth and Depth of Semantic Lexicons,* Kluwer Acad. Publ., p. 99-119.

[15] M. Johnson, Federica Busa (1999). *"Qualia Structure and Compositional Interpretation of Compounds",* in E. Viegas (ed.) *Breadth and Depth of Semantic Lexicons,* Kluwer Acad. Publ., p. 167-186.

[16] Denis Bouchard (2001). La *source semantique des facteurs heterogenes qui regissent la distribution des adjectifs,* Conferința Internațională "Representations du Sens Linguistique", Bucureşti.

[17] Dumitru Irimia (1997). *Morfo-sintaxa verbului românesc.* Editura Universității "Al. I. Cuza", Iaşi.

[18] Eva Hajicova, H. Skoumalova, P. Sgall (1995). *An Automatic Procedure for Topic-Focus Identification.* Computațional Linguistics, 21(1): 81-94.

[19] P. Sgall, E. Hajicova, J. Panevova (1986). *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects.* Kluwer Academic Publishers, Dordrecht.

[20] Neculai Curteanu (1988). *Augmented X-bar Schemes.* COLING'88 Proceedings, Budapest, p. 130-132.

[21] Dan Tufiş (2000). *Using a Large Set of EAGLES-compliant Morpho-Syntactic Descriptors as a Tagset for Probabilistic Tagging,* in Proceedings of the LREC'2000 International Conference, Athens.

[22] Dan Tufiş, A.M. Barbu (2001). *Computațional bilingual lexicography: automatic extraction of translation dictionaries,* In Romanian Journal on Information Science and Technology, voi. 4, no. 3.

[23] Neculai Curteanu (2002). *Către o teorie X-bar funcțională* (în prezentul volum).

O metodă automată pentru inserarea diacriticelor în texte în limba română

Rada F. MIHALCEA
University of Texas at Dallas, Richardson, Texas, U.S.A.
rada(a>utdallas.edu

Vivi A. NĂSTASE
University of Ottawa, Ottawa, Canada
vnastase(5>site.uottawa.ca

1. Introducere

Problema restaurării diacriticelor constă în inserarea diacriticelor într-u text în care lipsesc. Creşterea continuă a numărului de texte disponibile pr Internet face ca metodele automate de inserare a diacriticelor să devină componentă esențială în multe aplicații importante, cum ar fi extragerea d informații, traducerea automată, colecționarea de texte, construirea dicționarel electronice şi multe altele. Corectarea erorilor ortografice poate să aibă un impa major asupra calității rezultatelor obținute în aceste aplicații. De exemplu, î absența unei metode de restaurare a diacriticelor, unele cuvinte devin ambigu cum este cazul cuvintelor din limba română *peşte, peste* sau *păturí, paturi.* căutare bazată pe astfel cuvinte poate returna multe texte irelevante (de exempl o căutare pentru *peşte* ar returna şi documente conținând *peste).* De asemene traducerea unor astfel de cuvinte într-o limbă străină poate fi eronată (de exempl traducerea corectă a cuvântului *păturí* în limba engleză este *blankets,* dar absența diacriticului este tradus greşit ca şi *beds).*

Metodele dezvoltate până în prezent pentru rezolvarea acestei problem se bazează în general pe dicționare şi pe diverse procesoare lexicale şi/sa sintactice. Multe dintre limbile care se confruntă cu problema restaură diacriticelor nu beneficiază însă de astfel de resurse, şi ca urmare aplicabilitate acestor metode este limitată la limbi bine studiate care dispun de suficien resurse. Lucrarea de față prezintă o metodă automată de reinserare a diacritice în text care necesită doar o colecție de texte de dimensiuni modeste. Sp deosebire de alte metode dezvoltate anterior, metoda introdusă în această lucra nu necesită nici un fel de dicționare sau procesoare morfologice şi/sau sintactic

şi prin urmare poate fi folosită pentru prelucrarea de texte în orice limbă care dispune de un număr minim de texte cu diacritice. Datorită lipsei de restricţii, metoda propusă este foarte generală şi uşor aplicabilă pentru orice limbă. Pentru a demonstra aceasta afirmaţie, după ce vom prezenta experimentele pentru texte în limba română, vom arăta câteva rezultate obţinute pentru limbile cehă, poloneză şi maghiară.

2. Experimente anterioare

Restaurarea diacriticelor nu este în sine o problemă dificilă. Experimentele efectuate până în prezent au demonstrat că folosirea de dicţionare electronice poate duce la o acurateţe de peste 90% în restaurarea accentelor pentru limbile franceză şi spaniolă [9],[11],[5]. Metoda descrisă de Michael Simard în [9] este o îmbunătăţire adusă unei metode propusă anterior de El-Beze [4]. Această metodă se bazează pe Hidden Markov Models şi învaţă folosind cuvintele învecinate. Precizia raportată este de 99%. Tufiş şi Chiţu [10] propun o metodă similară pentru inserarea diacriticelor în texte în limba română. Yarowsky prezintă în [11] un set de metode folosite pentru restaurarea accentelor în limbile franceză şi spaniolă. Majoritatea algoritmilor pe care îi prezintă se bazează pe dicţionare şi cuvinte învecinate pentru a decide asupra ortografiei potrivite pentru fiecare cuvânt ambiguu. Yarowsky compară N-gram taggers, clasificatoare Bayesiene şi liste de decizii cu metoda de bază care constă în folosirea unui dicţionar. Pentru cele două limbi considerate în experimentele raportate, listele de decizii duc la performanţele cele mai ridicate. Toate aceste tehnici se bazează însă pe context, dicţionare şi în unele cazuri pe informaţii adiţionale de natură morfologică şi sintactică. Nagy et al. prezintă în [7] o abordare diferită a problemei, în care şiruri de litere sunt extrase din fiecare cuvânt şi folosite pentru a obţine statistici. Folosind metoda propusă, s-a observat o precizie foarte bună obţinută pe texte în limba franceză. Experimentele prezentate in [7] sunt asemănătoare cu cele raportate în [1], unde măsuri de similaritate între trigrame sunt folosite pentru a automatiza corectarea greşelilor de ortografie.

Majoritatea studiilor efectuate până în acest moment pe această temă, s-au ocupat de limbi bine cunoscute şi răspândite, cum ar fi franceza şi spaniola. Foarte puţine studii s-au concentrat pe limbi mai puţin mediatizate cum ar fi ceha, slovena, turca sau alte limbi care folosesc diacritice. Tabelul 1[1] prezintă diacriticele folosite în limbile europene cu alfabet latin. După cum rezultă din această listă, numeroase limbi se confruntă cu problema restaurării diacriticelor. Din setul de 36 de limbi cuprinse în tabel, engleza pare să fie singura limbă pentru care diacriticele

Tabelul cuprinde numai litere mici. Fiecărei litere mici îi corespunde o literă mare. Informaţia din acest tabel a fost agregată din liste de diacritice în limbi europene, disponibile la adresa www.tiro.com/di intro.html

nu constituie o problemă. Cuvintele din engleză care conţin diacritice au fost împrumutate din alte limbi, şi varianta acestora fără diacritice^ nu are un corespondent care să ducă la ambiguitate. Diacriticele par insa sa aiba un rol important în diferenţierea cuvintelor. Engleza, care după cum spuneam nu are diacritice per se, are în schimb o ambiguitatea semantică mai ridicata .

Tabel 1

Diacritice din limbile europene cu alfabet latin

Limba	Diacritice	Limba	Diacritice
Albaneză	şe	Malteză	Cgh2
Bască	nu	Norvegiană	âaso
Bretonă	âenuu'	Olandeză	ee
Catalană	âşeeiî166uii	Poloneză	^cşln6§z£
Cehă	âetfeift6F§fuuy2	Portugheză	ââşeîiooo
Daneză	â&0	Română	ââîşţ
Engleză	None	Sami (Laponă)	âi'Cdnr) Stz
Estoniană	âC86§uz	Serbo-croată	c£dSz
Faroeză	âaedi 6ouy	Slovacă	ââed'eiirnâof&f*ix</l*
Finlandeză	ââosz	Slovena	*Uz*
Franceză	ââa^eeeeîîooeuuy	Spaniolă	eun
Galiţiană	âeiou	Suedeză	ăâo
Germană	âouB	Turcă	ţgiioşu
Islandeză	âasdei66uyţ>	Sorbiană(l)	
Italiană	âeeiiîoouu	Sorbiană (2)	c5zS1AfSszz
Maghiară	âei656uUii	Welsh	âeîouwy

Aplicabilitatea metodelor menţionate anterior este limitată în următoarele cazuri:

Studii efectuate pe corpusuri bilingve paralele, ar arătat ca vocabularul construit dmr-u text în limba engleză este aproximativ jumătate din vocabularul construit pe baza aceluiaş text într-o altă limbă. Competiţia SENSEVAL [6] raportează de asemenea precizii mul ma mici pentru engleză comparativ cu alte limbi în rezolvarea ambiguităţii semantice. Lips diacriticelor în limba engleză ar putea constitui o explicaţie a acestui fenomen.

1. Dicționarele electronice nu sunt disponibile sau doar dicționare de dimensiuni relativ mici sunt făcute publice. Mai mult decât atât, în cazul în care dicționarul însuşi nu are diacritice, metodele care se bazează pe această resursă pentru restaurarea diacriticelor devin inaplicabile.
2. Procesoarele folosite pentru analiză morfologică şi/sau sintactică, considerate folositoare pentru problema restaurării diacriticelor, nu există sau nu sunt public disponibile.
3. Numărul de texte disponibile conținând diacritice este relativ mic. Mărimea corpusurilor publice sau disponibile prin Internet influențează mărimea vocabularului care poate fi construit ad-hoc pe baza acestor texte. în plus, majoritatea siturilor care publică texte pe Internet preferă în multe cazuri să evite diacriticele din motive de simplitate, uniformitate, sau pur şi simplu din lipsa de mijloace necesare pentru codificarea diacriticelor.

Lucrarea de față prezintă o metodă de restaurare a diacriticelor bazată pe învățarea la nivel de literă, şi nu la nivel de cuvânt. Avantajul principal al acestei metode este faptul că oferă posibilitatea de generalizare dincolo de cuvinte. Metoda este folositoare mai ales pentru limbile pentru care resursele disponibile sunt limitate, în speță limbi care nu au dicționare electronice mari cu diacritice. Limbi cunoscute şi bine studiate, precum franceza si spaniola, pot de asemenea beneficia de aceasta metodă pentru procesarea cuvintelor necunoscute.

Experimentele prezentate în această lucrare adresează în principal problema restaurării diacriticelor în texte în limba română. Precizia observată pe limba română este de 99%, măsurată la nivel de literă. Experimente similare au fost efectuate pe alte trei limbi, şi anume poloneză, maghiară şi cehă, de asemenea cu rezultate foarte bune. Avantajul principal al metodei este faptul că nu necesită nici o etapă de preprocesare, ci numai un corpus relativ mic format din texte cu diacritice. Datorită simplității algoritmului, viteza de procesare este foarte mare, de aproximativ 20 pagini de text pe secundă, măsurată pe un calculator cu un procesor Pentium III cu frecvența de 500MHz şi 250MB memorie.

^ Practic, metoda propusă încearcă să învețe reguli aplicabile la nivel de literă. în loc de a învăța reguli care se aplică la nivel de cuvânt, cum ar fi *"anuncio se scrie anuncio atunci când are funcția de verb",* dorim să învățăm reguli aplicabile la nivel de literă, cum ar fi *"s urmat de i şi spațiu şi precedat de spațiu se scrie ş".* Astfel de reguli, învățate la nivel de literă, sunt mai generale şi au aplicabilitate mai mare, în special în cazurile în care dicționarele disponibile sunt de dimensiune redusă, când se întâlnesc multe cuvinte necunoscute în textul dat, sau când procesoare pentru analiză morfologică sau sintactică nu sunt la îdemână.

Este evident că în analiza limbajului literele constituie nivelul cu granularitatea cea mai scăzută, şi de aceea au şi cel mai mare potențial de generalizare. în loc de aproximativ 150.000 de unități candidate potențiale pentru algoritm (mărimea aproximativă a vocabularului de uz general a unei limbi), vom

şvea mai mult sau mai puțin 26 caractere pe baza cărora se vor constitui intrare pentru algoritmul de dezambiguare[1].

3. Experimente

Scopul experimentelor descrise în această lucrare este de a învățarea la nivel de literă este posibilă şi poate rezolva, cu prec problema restaurării diacriticelor. Pe lângă faptul că metoda propusă c problemă de cercetare, scopul învățării la un nivel de granularitate atât este de a oferi o metodă viabilă pentru limbile pentru care resursele isemantice disponibile sunt limitate, şi pentru care restaurarea diacri învățare la nivel de cuvânt este greu de realizat.

3.1. Date

Prezentăm în primul rând experimentele efectuate pe texte română. Limba română nu este o limbă foarte răspândită şi în conseci foarte multe resurse publice disponibile pentru pre-procesare, iar d electronice sunt de dimensiuni relativ mici. în al doilea rând, am avut de problemă specifică de restaurare a diacriticelor într-un dicționar electro englez care conține aproximativ 75.000 de cuvinte, dar are dezav diacriticele lipsesc. Am considerat că este avantajos să studiem proble rării diacriticelor şi să folosim acest dicționar, în loc să ne bazăm pe alte cu diacritice de dimensiuni reduse. în plus, pentru procesoarele pe care le dezvoltăm pentru limba română avem nevoie de numeroase texte ele limba română. De obicei aceste texte nu au diacritice, şi deci r diacriticelor este din nou necesară. Avem de asemenea posibilitatea de eficacitatea acestei metode cu rezultate obținute în experimente ef aceiaşi limbă constând în metode în care învățarea se face la nivel de c

Pentru a aplica metoda descrisă în lucrarea de față, avem deci colecție de texte românești cu diacritice. în acest scop, am colectat "România Literară"[2], un ziar românesc publicat săptămânal, cu articol general de literatură. Ziarul are o versiune care conține diacritice începâ 2000. Colecția disponibilă on-line la data colectării datelor (august 2001 2780 articole. în pasul următor, textul a fost extras din fişierele HTI deosebită a fost acordată doar caracterelor românești. Alte caractere d întâlnite ocazional, cum ar fi e, e etc. au fost transformate în forma lor d fără diacritice, având în vedere că suntem interesați doar de caracterele

[1] *Numărul de litere depinde de limba care se analizează. S-a arătat d aproximativ 85% dintre cuvintele în limba franceză nu au o formă ortografică şi deci numai 20.000 de cuvinte sunt potențial ambigue. Pe de altă parte, • sunt ambigue în limba franceză.*

Accesibil prin http://www.romlit.ro

şi nu de caractere franceze sau din alte limbi. După toate aceste faze premergătoare, am obținut un corpus conținând aproximativ 3 milioane de cuvinte.

Literele mari au fost transformate în litere mici. Cazul literelor â şi / este special în limba română: deşi pronunția lor este identică, folosirea lor este guvernată de reguli bazate pe poziția lor în cuvânt. La începutul cuvântului se foloseşte întotdeauna /, iar â se foloseşte în interiorul cuvântului. Este bine cunoscut faptul că folosirea acestor litere a fost controversată de-a lungul timpului. O lege din anii '60 a schimbat ortografierea de la â la /, singura excepție fiind cuvintele derivate din rădăcina *român.* La începutul anilor "90 ortografia veche a fost reintrodusă, şi astfel s-a ajuns la cazuri de texte inconsistente, în care se întâlnesc scrieri diferite ale aceluiaşi cuvânt. De exemplu, *cîntec* şi *cântec* sunt forme ale aceluiaşi cuvânt care pot fi întâlnite în acelaşi text. Ziarul *"România Literară"* păstrează încă ortografia cu /, cu mici excepții (de exemplu, articole scrise de scriitori invitați care preferă să scrie folosind â în loc de /).

3.2. Algoritmi de învățare

Pentru a rezolva problema restaurării diacriticelor, am ales să folosim un algoritm bazat pe învățarea de instanțe (IBL). Există două motive importante care au stat la baza luării acestei decizii. în primul rând, este faptul demonstrat că excepțiile au un rol important în procesarea limbajelor naturale. Algoritmii de tip IBL sunt recunoscuți pentru faptul că iau în considerare fiecare exemplu de antrenament în luarea unei decizii de clasificare [2], şi deci folosirea acestui tip de algoritmi prezintă un avantaj deosebit în probleme de limbaj natural. în al doilea rând, acest gen de algoritmi sunt foarte eficienți relativ la timpul de antrenament şi testare.

învățarea pe bază de instanțe se desfăşoară în felul următor: în pasul de antrenament, toate exemplele de intrare sunt memorate. în faza de testare, fiecare exemplu din set este comparat cu exemplele memorate şi va primi-clasificarea dată de exemplul memorat de care este cel mai apropiat, distanța fiind dată de măsura specifică aleasă în implementarea folosită. Pentru efectuarea experimentelor propuse, am folosit implementarea TiMBL [3] a acestor algoritmi. în plus, am efectuat experimente asemănătoare şi cu un clasificator pe bază de arbori de decizie, şi anume C4.5 [8]. Arborii de decizie sunt construiți din setul de exemple de antrenament. La fiecare pas este ales un atribut care discriminează cel mai bine exemple din clase diferite (prin valorile sale). Grupele obținute prin diviziunea după acest atribut vor fi din nou împărțite în grupe mai mici şi mai pure, prin alegerea unui nou atribut care discriminează cel mai bine exemplele din grupă. Acest proces continuă până când grupele obținute au un grad de puritate acceptabil, sau mărimea arborelui depăşeşte un prag ales inițial. Rezultatele obținute folosind C4.5 sunt asemănătoare cu cele obținute folosind TiMBL, însă C4.5 are capacitatea de a genera reguli expresive, folositoare pentru implementări practice.

Având în vedere că lucrăm la nivelul literelor, atributul care trebuie învățat este constituit de litera ambiguă. Acesta poate fi oricare din literele ambigue enumerate în Tabelul 1. Pentru limba română avem 4 perechi de litere ambigue: s - ş, t - ț, a - ă, i - î. Literele mari au fost convertite în prealabil în litere mici. Datorită faptului că datele folosite aplică ortografia cu /, nu avem ambiguitatea a - â, ci doar ambiguitatea / - /. Aceasta nu implică însă o pierdere de generalitate. Conversia între cele doua forme de ortografie este simplă şi se poate realiza folosind doar poziția literei în cuvânt, şi prin urmare scrierile diferite nu afectează rezultatul algoritmului.

3.3. Atribute

Atributele folosite în orice algoritm de învățare au un impact foarte mare asupra eficacității algoritmului. După cum am menționat şi în introducere, nu avem posibilitatea de a folosi procesoare care determină partea de vorbire a cuvintelor, şi nici un alt fel de analizoare morfologice sau sintactice. în plus, nu dorim să ne bazăm pe cuvintele învecinate, deoarece avem un număr limitat de date, şi în consecință există şansa de a întâlni un număr mare de cuvinte necunoscute. Prin urmare, ne-am decis asupra folosirii unor atribute foarte simple, pentru extragerea cărora nu este nevoie de nici un fel procesare specială. Vom folosi litere învecinate, cu o notație specială atribuită spațiilor, virgulelor şi punctelor (aceste caractere pot afecta procesul de învățare, fiind considerate caractere speciale de către C4.5 şi/sau TiMBL).

Dacă X este litera a cărui ambiguitate trebuie rezolvată, atributele folosite sunt N litere la stânga şi la dreapta literei ambigue:

$$L_{-N}, \quad L_{-(N-1)}, \quad X, \quad L_1, \quad L_2 \quad L_{(N-1)}, \quad L_N$$

Acest set de atribute se comportă surprinzător de bine, relativ la acuratețe, după cum vom arăta in cele ce urmează.

După cum am menționat anterior, am ales să nu ne bazăm pe nici un tag obținut cu procesoare lexicale sau morfologice, ci doar pe informația care se poate extrage din text neprelucrat. De asemenea, suntem interesați să găsim posibilități de generalizare, astfel încât un corpus limitat să poată fi folosit pentru a genera reguli de reinserare a diacriticelor. în loc de a învăța reguli bazându-ne pe cuvinte, după cum s-a procedat până acum, dorim să învățăm reguli bazate pe litere, pentru că acestea constituie cele mai mici unități în limbaj, şi oferă posibilitatea învățării chiar şi dintr-o colecție mică de texte.

Pentru fiecare pereche ambiguă de litere, parcurgem textul şi generăm toate exemplele posibile întâlnite în text. Atributele într-un exemplu sunt formate folosind N litere la stânga şi la dreapta literei ambigue, şi atributul țintă este însăşi litera ambiguă. Forma generală a exemplelor generate este:

$$L_{-N}, \quad L_{-(N-1)}, \ldots, L_1, L_2 \quad L_{(N-1)}, L_N, X$$

unde ca şi în exemplul anterior, X este litera ambiguă. Prezentăm mai jos exemple de vectori de atribute care constituie date de intrare pentru algoritmul de învățare pentru rezolvarea ambiguității perechii s - ş. CO, DO şi SP sunt codurile care înlocuiesc virgula, punctul şi spațiul.

/, *i,n,SP,(,u,b,SP,i,n,s.*

e,CO,SP,r,o,-,g,a,r,d,ş.

g , a , r , d , i , t , u , l , CO, SP, s.

e,SP,o,r,a,DO,SP,t,o,t,ş.

învățarea se reduce la detectarea corelațiilor între valorile atributelor care caracterizează exemplele de antrenament şi valorile atributelor țintă, şi utilizarea acestora pentru stabilirea valorii atributului țintă din exemplele de testare.

Numărul de exemple extrase din corpus depinde de perechea de litere. Din întregul set de 3 milioane de cuvinte, am obținut 2.161.556 exemple pentru perechea ambiguă a - ă, 2.055.147 pentru perechea / - /, 1.257.458 exemple pentru *t - ţ,* şi în final 866.964 exemple pentru perechea s - s. în fiecare din aceste cazuri, spațiul exemplelor este împărțit în două clase, date de cele 2 variante ale literei ambigue. Metoda de învățare automată va folosi atributele date pentru a găsi reguli de clasificare a exemplelor în cele 2 clase.

3 * 4 ▮ Rezultate

Precizia cea mai ridicată s-a obținut pentru o fereastră de 10 litere în vecinătatea literei ambigue (N = 5). Dată fiind această observație, am considerat că este important să studiem mai în detaliu acest caz, şi să determinăm ratele de învățare pentru cele 4 perechi de litere ambigue. Cu toate acestea, prezentăm rezultate pentru ferestre de diverse dimensiuni, pentru comparație.

Tabelul 2 arată rezultatele obținute pentru N=5. Preciziile raportate în acest tabel sunt obținute folosind algoritmul bazat pe învățarea de instanțe. Am efectuat experimente cu seturi de antrenament de diverse dimensiuni, variind de la 2.000.000 exemple până la 10 exemple, pentru a determina rata de învățare şi dimensiunea minimă a corpusului necesară pentru a obține o precizie satisfăcătoare. în toate aceste experimente s-au folosit seturi de testare conținând 50.000 exemple. Pentru a obține rezultate cât mai exacte am folosit validare încrucişată folosind 10 seturi diferite de test. Tabelul indică de asemenea baza de comparație, definită aici ca fiind precizia obținută când se foloseşte implicit litera cea mai frecventă din fiecare pereche ambiguă.

Rezultatele prezentate în Tabelul 2 sunt reprezentate grafic în Figura 1. Este interesant de observat că cea mai importantă fază a procesului de învățare are loc când se folosesc primele 10 000 exemple. în conformitate cu măsurătorile efectuate, a rezultat că aproximativ 100.000 - 250.000 caractere (aproximativ 25-60 pagini de text) sunt necesare pentru a genera 10.000 exemple cu diacritice, ceea ce constituie un corpus de dimensiune relativ mic. Mai departe, pentru a obține îmbunătățiri de numai 1% este necesar un număr semnificativ de exemple. Tabelul 2 indică de asemenea, în caractere groase, prima precizie care depăşeşte baza de comparație, ca o indicație a dimensiunii minime a setului de antrenament pentru care se observă o formă minimă de învățare. După cum se tabel, numai 1.000 exemple sunt suficiente pentru învățare.

Rezultate obținute în rezolvarea ambiguității literelor cu diacrit română

	Pereche ambiguă			
	a - ă	**a-ă**(2)	i - î	s - ş
Nr.tota! exemple	2.161.566	1.369.517	2.055.147	866.964
Baza comparație	74.70°/c	85.90°/c	88.205	76.53°/c
Exemple de Antrenament	Precizie obținută pe date de test (50.000 exemple)			
2,000,000	96.14%	-	99.69%	-
1,000,000	95.10%	99.14%	99.58%	-
750,000	94.83%	98.97%	99.53%	99.07%
500,000	94.57%	98.79%	99.46%	98.86%
250,000	94.00%	98.37%	99.28%	98.87%
100,000	93.03%	97.56%	98.96%	98.54%
50,000	92.10%	96.86%	98.57%	98.13%
25,000	90.99%	95.75%	98.11%	97.58%
10,000	88.99%	93.75%	97.31%	96.53%
5,000	87.56%	92.76%	96.65%	95.61%
4,000	86.91%	91.86%	96.49%	94.99%
3,000	86.39%	90.99%	96.19%	94.18%
2,000	85.81%	89.93%	95.49%	93.47%
1,000	83.49%	88.36%	93.78%	92.31%
500	80.61%	85.66%	93.07%	90.75%
250	77.89%	83.17%	92.75%	87.41%
100	74.80%	84.04%	91.41%	82.13%
50	72.79%	82.73%	88.05%	86.53%
25	72.45%	81.34%	88.15%	78.26%
10	73.38%	85.90%	88.20%	75.88%

Folosind întregul set de exemple extrase din corpus ambiguității perechii *i* - feste aproape 100% corectă. Pentru acest acum o instanță greşită din 300 instanțe, în timp ce baza de compa instanță greşită din fiecare 8 instanțe, deci o îmbunătățire semnificati

Cel mai slab rezultat este obținut în cazul perechii a - ă. Du rezultatelor, reiese că principalul motiv care cauzează această pr

este faptul că multe substantive în limba română au forma nearticulată terminată în ă şi forma articulată terminată în a. De exemplu, *masa* şi *masă* reprezintă forma articulată şi respectiv nearticulată a substantivului *masă.* De asemenea, timpuri diferite ale aceluiaşi verb se disting numai prin terminaţia în a sau *ă.* Algoritmul de învăţare este deci indus în eroare din cauza folosirii acestor litere în contexte identice. O soluţie simplă constă în evitarea în procesul de învăţare a exemplelor care conţin a sau *ă* la sfârşitul unui cuvânt. Rezultatele obţinute sub aceasta ipoteză simplificatoare sunt raportate in Tabelul 2, în coloana a-ă(2). După cum se arată în tabel, câştigul este de mai mult de 4% in precizie folosind doar această condiţie simplă (câştig care se traduce într-o reducere a erorii de 87%).

Am folosit de asemenea şi algoritmul de învăţare bazat pe arbori de decizie C4.5, cu aceleaşi date de antrenament, fără a observa însă nici o îmbunătăţire comparativ cu rezultatele raportate în Tabelul 2. Dezavantajul folosirii C4.5 pentru această problemă este faptul că faza de învăţare este mult mai lentă decât în cazul folosirii algoritmului TiMBL. Pe de altă parte, C4.5 are capacitatea de a genera reguli expresive. "Dacă L?=e *şi L₂=spaţiu atunci* s"(99.5%), "Dacă *L^t şi L₂=spaţiu atunci s"* (98.7%), "Dacă *L₄=p şi L.^v şi* L₁=f *şi L₂=e atunci* s"(95.5%), sunt exemple de astfel de reguli. L/denotă o literă învecinată în poziţia i relativ la litera ambiguă. Se observă că aceste reguli nu ţin cont de faptul că literele folosite în clasificare aparţin aceluiaşi cuvânt sau nu: Algoritmul de învăţare se bazează pur şi simplu pe litere, indiferent de cuvântul căruia îi aparţin. în consecinţă, pseudo-omonimele (cum ar fi *peste* şi *peşte),* sunt adresate în mod egal de această metodă, pentru că algoritmul are capacitatea de a se extinde dincolo de cuvinte.

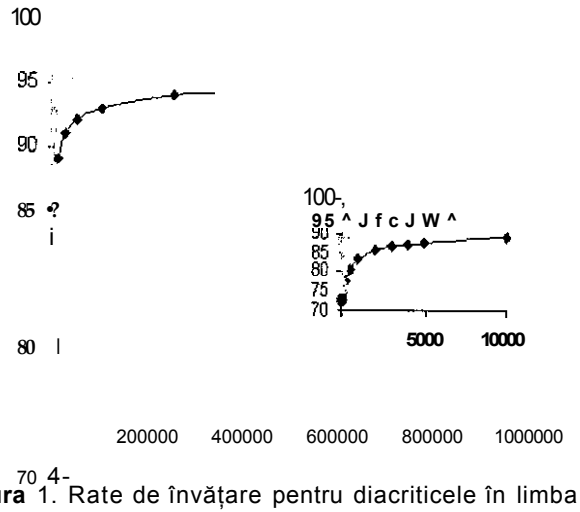


Figura 1. Rate de învăţare pentru diacriticele în limba română. Graficul din mijloc este o reprezentare mărită a zonei 0-10.000

3.5. Ferestre de dimensiune diferită

Am efectuat experimente cu ferestre de diverse dimensiuni, pentru a determina dimensiunea contextului care modelează cel mai bine problema noastră. Pentru aceasta, am considerat ferestre de dimensiune dpi, şase, zece, patrusprezece şi optsprezece litere învecinate (i.e. N = 1,3,5,7,9). Rezultate comparative sunt prezentate în Tabelul 3. Aceste numere trebuie comparate cu primul rând din Tabelul 2 (coloana corespunzătoare valorii N=5 în tabelul de faţă).

Tabel 3

Rezultate comparative obţinute cu ferestre de dimensiuni diferite

Pereche ambiguă	Dimensiune fereastră				
	N=1	N=3	N=5	N=7	N=9
a-ă(2)	88.63%	98.79%	99.14%	99.10%	99.10%
i -î	94.18%	99.13%	99.69%	99.68%	99.43%
s-ş	88.09%	99.06%	99.07%	99.02%	99.00%
t-ţ	89.45%	98.57%	98.75%	98.67%	98.25%

Când nu există suficient context disponibil, o fereastră de dimensiune N=3 poate fi folosită fără a pierde mult din precizie. însă, după cum am specificat şi înainte, cea mai ridicată acurateţe se obţine pentru o fereastră de zece litere înconjurătoare (N=5).

3.6. Comparaţie cu experimente asemănătoare

Rezultatele prezentate în lucrarea de faţă se pot compara cu rezultatele raportate de Tufiş şi Chiţu [10], care au folosit tot limba română în experimentele lor. Tufiş şi Chiţu menţionează că sarcina recuperării diacriticelor în limba română este mai dificilă decât în alte limbi, deoarece în română diacriticele sunt mai intens folosite. După cum raportează în experimentele lor, numai 60% din cuvintele din limba română nu au diacritice, comparat cu studii menţionate în [9] care arată că aproximativ 85% dintre cuvintele limbii franceze se scriu fără accent.

Abordarea prezentată de Tufiş şi Chiţu foloseşte dicţionare, un analizor morfologic, iar învăţarea se face la nivel de cuvinte. Folosind aceste resurse, au obţinut o precizie globală de 97.4%. Nu putem efectua o comparaţie directă a rezultatelor noastre, având în vedere că atât metodele, cât şi modul de evaluare, sunt fundamental diferite. Precizia medie de 99% pe care noi o raportăm este măsurată la nivel de literă, pe când acurateţea raportată în [10] este determinată la nivel de cuvânt[1].

[1] *Diferenţa dintre precizia raportată la nivel de literă şi precizia raportată la nivel de cuvânt rezultă practic din diferenţa de granularitate dintre litere şi cuvinte. Presupunând că un cuvânt conţine L litere ambigue, o singură literă din acest set L a cărui ambiguitate este rezolvată greşit face ca întreg cuvântul să fie considerat greşit, pe când la nivel de litere avem doar o singura eroare din setul L. Pe de altă parte, chiar dacă mai multe litere din*

Metodologia noastră depăşeşte abordările anterioare, prin faptul că s-au obținut precizii şi viteze de procesare ridicate fără a folosi nici un fel de resurse adiționale cum ar fi procesoare pentru analiză morfologică sau sintactică sau dicționare. Din aceste motive, algoritmul se poate aplica oricărei limbi, singura cerință fiind un corpus relativ mic de texte cu diacritice.

4. Alte limbi

Pentru a demonstra generalitatea algoritmului pe care l-am propus, am efectuat experimente pe texte în alte trei limbi europene care fac uz de diacritice: cehă, poloneză şi maghiară. Limbile considerate pentru aceste experimente sunt limbi cu răspândire restrânsă, pentru care resursele publice sunt limitate.

Pentru fiecare dintre aceste limbi am colectat texte cu diacritice disponibile prin Internet. Principalele surse folosite pentru formarea setului de date sunt după cum urmează: (1) pentru cehă, am folosit arhiva ziarului *Lidovky* şi texte literare de *Kafka, Hasek* şi *Capek;* (2) pentru maghiară, arhiva furnizată de către *Digitâlis Irodalmi Akademia* şi un roman de *Petofi Sândor,* (3) pentru poloneză, arhiva ziarului *Wiedza i zycie.* Pe lângă acestea, am mai folosit texte adiționale colectate de pe diverse situri, astfel încât să obținem un corpus de minim un milion de cuvinte pentru fiecare limbă. Asemănător cu procesarea aplicată limbii române, datele au fost convertite în fişiere text, iar literele mari au fost transformate în litere mici. în urma acestei etape de pre-procesare, am obținut un corpus de 1.46 milioane cuvinte pentru cehă, 1.72 milioane cuvinte în maghiară şi 2.5 milioane cuvinte în poloneză.

Algoritmii de învățare şi atributele folosite în procesul de învățare sunt identice cu cele folosite în experimentele efectuate pe limba română, raportate în detaliu în secțiunea precedentă. Tabelul 4 prezintă rezultatele obținute pentru cele trei limbi. Pentru fiecare set de litere ambigue, sunt prezentate în tabel: (1) numărul de exemple obținute din corpusul limbii respective, (2) baza de comparație, măsurată ca fiind precizia ce se poate obține dacă pentru fiecare set ambiguu se foloseşte implicit litera cu frecvența de apariție cea mai ridicată, şi (3) precizia obținută prin aplicarea metodei propusă în lucrarea de față.

Media obținută pentru toate patru limbile studiate (cele trei limbi a căror rezultate sunt prezentate in Tabelul 4, şi limba română) este de 98.17%. Precizia medie măsurată pe fiecare limbă în parte este influențată de mărimea setului de date folosit. Textele colectate pentru cehă şi maghiară conțin aproximativ 1.4-1.7 milioane cuvinte, şi prin urmare precizia obținută în aceste două limbi este mai joasă decât pentru poloneză şi română, pentru care am reuşit să colectăm un corpus de 2.5-3 milioane cuvinte. Estimăm deci posibilitatea creşterii preciziei ca urmare a creşterii dimensiunii corpusului de antrenament.

setul L sunt rezolvate greşit, avem tot o singură eroare la nivel de cuvânt, dar mai multe erori la nivel de literă. Nu este deci foarte clar care ar fi modalitatea corectă de a compara aceste două metode care lucrează la nivele de granularitate diferită.

Tabel 4

Rezultate obținute în restaurarea diacriticelor în trei limbi europene

Set litere ambigue	Număr exemple	Bază comparație	Metodă propusă
Cehă			
a â	649,886	75.01%	96.96%
c *t*	217,570	72,21%	97.08%
d <r	271,070	99.05%	99.86%
e e	768,051	74.59%	97.02%
i i	504,298	60.43%	96.29%
n fi	439,552	98.97%	99.71%
o 6	566,521	99.08%	99.86%
r f	319,352	65.55%	97.60%
s s	380,805	84.44%	98.88%
t f	387,214	99.05%	99.85%
u ti u	264,408	80.89%	93.51%
y y	191,317	65.55%	95.06%
z z	219,082	66.49%	98.70%
		Medie	97.83%
Maghiară			
a â	1,198,294	73.51%)	96.91%
e e	1,306,944	76.34%	96.40%
i i	647,137	89.14%	99.49%
o 6 6 6	678,012	71.15%	96.10%
u u ii îi	207,753	56.00%	97.31%
		Medie	97.04%
Poloneză			
a a_	1,387,019	88.83%	97.07%
c e	657,669	91.50%	99.42%
e e.	1,305,584	89.23%	98.47%
1 *i*	506,041	59.29°/,	98.80%
n ri	878,824[	96.75°/,	99.85%
o 6	1,230,38S)	88.67°/>	99.87%
s *â*	688,67"	88.67°/>	99.83%
z z z	896,90S>	86.26°/>	99.73%
		Medie	99.02%

Este interesant de observat că numărul de diacritice într-o limbă nu influențează precizia medie obținută. Precizia care se obține în cazul limbii maghiare, care are un total de 5 seturi de litere ambigue, este mai scăzută decât precizia care se obține pentru limba cehă, care are un număr impresionant de diacritice (treisprezece). Şi aceasta cu toate că datele colectate pentru limba maghiară sunt mai numeroase decât datele colectate pentru limba cehă.

5. Concluzii

Am descris în lucrarea de față o metodă de restaurare a diacriticelor bazată pe tehnici de învățare la nivelul de literă. Avantajul principal al metodei constă în capacitatea ei de generalizare dincolo de cuvinte. Nu este necesară nici un fel de analiză a textului, şi nu se folosesc nici un fel de procesoare de limbaj sau dicționare. Singura cerință este un corpus relativ mic de texte cu diacritice.

Metoda este folositoare în special pentru limbi pentru care nu sunt disponibile dicționare electronice de dimensiune adecvate, şi nici procesoare pentru analiză morfologică şi/sau sintactică. Mecanismul de învățare foloseşte date de intrare extrase din texte neprelucrate, şi generează rezultate cu o precizie ridicată. Experimente detaliate efectuate pe texte în limba română au arătat că restaurarea diacriticelor în această limbă se poate efectua folosind metoda propusă cu o precizie de peste 99% la nivel de literă. Rezultatele au fost validate prin experimente efectuate pe alte trei limbi europene care fac uz de diacritice: cehă, poloneză şi maghiară. Precizie medie măsurată pe cele patru limbi de studiu este de 98.14%, fapt care demonstrează că metoda este independentă de limbă, în plus, un alt avantaj al metodei este faptul că, datorită simplității sale, viteza de procesare este foarte mare, de până la 20 pagini de text pe secundă.

Referințe bibliografice

[1] Angell, R., Freund G., Willett, P. *Automatic spelling correction using a trigram similarity measure.* Information Processing and Management 19, 4 (1983), 255-261.

[2] Daelemans, W., van den Bosch, A., Zavrel, J. *Forgetting exceptions is harmful in language learning.* Machine Learning 34, 1-3 (1999), 11-34.

[3] Daelemans, W., Zavrel, J., van der Sloot, K., van den Bosch, A. *TiMBL: Tilburg memory based learner, version 4.0, refernce guide.* Tech. Rep., University of Antwerp, 2001.

[4] El-Beze, M., Merialdo, B., Rozeron, B., Derouault, A., *Accentuation automatique des textes par des methodes probabilistes.* Techniques et sciences informatique 16, 6 (1994), 797-815.

[5] Galicia-Haro, S., Bolshakov, I., Gelbukh, A. *A simple Spanish part of speech tagger for detection and correction of accentuation error.* In Text, Speech and Dialogue - Second International Workshop, TSD'99, September 1999, Proceedings (Plzen, Czech Republic, 1999), voi. 1692 of Lecture Notes in Computer Science, Springer, pp. 219-222.

[6] Kilgariff, A., ed., Proceedings of SENSEVAL-2, 2002.

[7] Nagy, G., N., N., and Sabourin, M. *Signes diacritiques: perdus et retrouves.* In Actes du 1er Colloque International Francophone sur l'Ecrit et le Document CIFED '98 (Quebec, Canada, 1998), pp. 404-412.

[8] Quinlan, J. *C4.5: programs formachine learning.* Morgan Kaufman, 1993.

[9] Simard, M. *Automatic insertion of accents in French text.* In Proceedings of the Conference on Empirical Methods in Natural Language Processing EMNLP-3 (Granada, Spain, 1998).

[10] Tufiş, D., Chițu, A. *Automatic diacritics insertion in Romanian texts.* In Proceedings of the International Conference on Computațional Lexicography COMPLEX'99 (Pecs, Hungary, June 1999).

[11] Yarowsky, D. *Corpus-based techniques for restoring accents in Spanish and French texts.* In Natural Language Processing Using Very Large Corpora. Kluwer Academics Publisher, 1999, pp. 99-120.

Contribuţii privind structura statistică de cuvinte în limba română scrisă*

Adriana VLAD şi Adrian MITREA
Universitatea "POLITEHNICA" din Bucureşti
Facultatea de Electronică şi Telecomunicaţii
B-dul. Iuliu Maniu, 1-3, Bucureşti, România
adriana_vlad@yahoo.com

1. Introducere

Această lucrare aparţine unui studiu mai larg dedicat de autori descrierii limbii române ca sursă de informaţie. Punctul de plecare al acestui studiu a fost presupunerea generală conform căreia limba naturală este bine aproximată de un lanţ Markov ergodic multiplu, cu ordin de multiplicitate mai mare decât 30, [1]. Descrierea acestei surse Markov multiple se realizează prin aproximaţii succesive. Investigaţia noastră statistică până în prezent a descris structura de litere, digrame, trigrame, tetragrame, precum şi probabilităţile condiţionate de o literă precedentă, [2]-[8].

Obiectivul principal ale prezentei lucrări este descrierea sursei de informaţie fără memorie având ca simboluri cuvintele limbii române scrise. Aceasta presupune determinarea probabilităţii unui cuvânt (oricare ales), în caz că această probabilitate există. Determinarea probabilităţii a însemnat implicit şi o verificare a ipotezei de staţionaritate a limbii române scrise pe baza structurii de cuvinte; verificarea s-a făcut utilizând o procedură similară cu cea pe care am dezvoltat-o pentru m-grame, [3]-[8] (m-grama este o succesiune de *m* litere consecutive în texte naturale). Metoda noastră statistică de a determina probabilităţile cuvintelor a combinat următoarele tipuri de inferenţe statistice: teoria estimării cu multiple intervale de încredere statistică; test al ipotezei că probabilitatea aparţine unui interval; test de egalitate între două probabilităţi.

Primele două tipuri de inferenţe statistice menţionate (teoria estimării cu multiple intervale de încredere statistica; test al ipotezei ca probabilitatea aparţine unui interval) au fost folosite pentru a decide care este intervalul de încredere statistică *"reprezentativ* pentru probabilitatea cuvântului investigat în textul natural.

O *parte din acest studiu* s-a *desfăşurat în cadrul unuiGrant CNCSIS-MEC (2001-2002) cu tema:* 'Descrierea limbii române scrise ca sursă de informaţie"

Simultan a apărut şi o mulțime *"reprezentativă"* de date *Ud.* extrase din textul natural, corespunzătoare cuvântului investigat (modelul statistic ***i.i.d.*** presupune că datele provin din variabile aleatoare independente statistic şi identic distribuite).

Ultimele doua tipuri de inferențe statistice menționate (test al ipotezei ca probabilitatea aparține unui interval; test de egalitate între două probabilități) ca şi intervalele de încredere statistica *"reprezentative"* şi mulțimile de date *"reprezentative"* obținute în prealabil au fost folosite pentru comparații matematice între texte naturale. Aceste comparații matematice (dincolo de valoarea lor ca atare) au avut scopul principal de a vedea dacă putem vorbi de un model matematic al sursei de cuvinte pentru limba ca ansamblu, pe domenii ale limbii, pe autori etc. Comparațiile s-au făcut în două moduri:

- urmărind probabilitățile unui cuvânt (acelaşi) în texte naturale diferite;
- urmărind probabilitățile cuvintelor situate pe acelaşi rang în texte naturale diferite (se compară probabilitățile asociate unui aceluiaşi rang în ierarhiile frecvențelor relative).

Rezultatele experimentale au adus probe noi în sprijinul ipotezei de staționaritate a limbii române scrise în cadrul unui aceluiaşi domeniu punctând către unele diferențe între domenii diferite.

Investigația noastră (atât privind intervalele de încredere statistică *"reprezentative",* cât şi comparația matematică dintre texte) a avut în vedere şi eroarea statistică de ordinul al doilea. Acest tip de eroare are un rol special în dimensionarea unui nou corpus lingvistic care să satisfacă acuratețea dorită pentru descrierea modelului matematic (sursa de informație de cuvinte).

Lucrarea mai conține si un studiu experimental al uneia dintre cele mai cunoscute legi de tipul rang - frecvență, legea lui Zipf. Este analizat şi un corolar al acesteia, de interes lingvistic.

Analiza experimentală s-a bazat pe corpusul lingvistic global pe care l-am alcătuit în prealabil pentru studiul structurilor de litere, digrame, trigrame şi tetragrame (vezi spre exemplu [6]). Acest corpus este format din 93 de cărți în limba română, scrise cu noua ortografie (introdusă după 1993). Cărțile reprezintă: literatură scrisă de autori români (11 cărți: romane şi nuvele), literatură străină tradusă în română (47 de romane şi nuvele), cărți ştiințifice (drept, medicină, silvicultură, istorie, sociologie etc.) şi altele. Au fost considerate doar cele 31 de litere ale limbii române (A Ă Â B C D E F G H I Î J K L M N O P Q R S Ş T Ț U V W X Y Z) precum şi caracterul spațiu (blanc); orice alte simboluri (cifre, elemente de ortografie sau punctuație) au fost eliminate (suprimate).

Rezultatele experimentale au fost obținute pe diverse corpusuri organizate pe baza celor 93 de cărți:

- Corpusul Mixt Global **(#CMG)** - obținut prin concatenarea aleatoare a celor 93 de cărți; acesta conține un număr de L_c = 8806433 cuvinte dintre care N_c = 202403 sunt distincte.
- Cele două jumătăți ale Corpusului Mixt Global: prima jumătate **(#1JCMG)** şi a doua jumătate **(#2JCMG);** acestea conțin un număr de L_c = 4403217 cuvinte şi respectiv L_c =4403216 dintre care N_c =148853 şi respectiv N_c =137845 sunt distincte.
- Corpusul Literar Global **(#CLG)** - obținut prin concatenarea aleatoare a 58 de cărți (romane şi nuvele scrise de autori români sau traduse în română); acesta conține un număr de L_c =6255235 cuvinte dintre care N_c = 162124 sunt distincte.
- Cele două jumătăți ale Corpusului Literar Global: prima jumătate **(#1JCLG)** şi a doua jumătate **(#2JCLG);** acestea conțin un număr de L_c =3127618 cuvinte şi respectiv L_c= 3127617 dintre care N_c = 116247 şi respectiv N_c = 116860 sunt distincte.
- Corpusul Ştiințific Global **(#CŞG)** - obținut prin concatenarea aleatoare a 11 de cărți; acesta conține un număr de L_c =1049969 cuvinte dintre care N_c = 59093 sunt distincte.

Au fost făcute determinări atât pe o singură carte cât şi pe grupuri de cărți scrise de acelaşi autor. Dintre acestea menționăm:

#1. GeorgeCălinescu, *Bietul Ioanide,* Editura Minerva, Bucureşti, 1995 ISBN 973-21-0432-5 *(voi 1,* ISBN 973-21-0431-7, pag. 1-214),-*{voi. 2* ISBN 973-21-0433-3, pag. 5-256), *{voi. 3,* ISBN 973-21-0434-1, pag 5-238).

#2. Radu Anton Roman, *Precum fumul,* Editura Cartea Românească Bucureşti, 1996, ISBN 973-23-0274-7, pag. 5-283.

#3. Radu Anton Roman, *Z/7e* de *pescuit,* Editura Metropol, Bucureşti, 1996 ISBN 973-562-073-1, pag. 11-302.

#4. John le Carre, *Casa Rusia,* Editura Univers, Bucureşti, 1997, ISBN 973-34-0457-8, pag. 9-355.

#5. John le Carre, *Spionul care venea din frig,* Editura Univers, Bucureşti 1996, ISBN 973-34-0355-5, pag. 9-252, cu ortografie actualizată.

#6. John Le Carre, *Micuța toboşăreasă,* Editura Univers, Bucureşti, 1998 ISBN 973-34-0430-6, pag. 7-443, cu ortografie actualizată.

#7. Alexandr Soljenițîn, *Arhipelagul Gulag,* Editura Univers, Bucureşti, *(voi I,* 1997, ISBN 973-34-0454-3, pag. 7-432), *{voi. II,* 1997, ISBN 973-34-0480-2, pag. 5-474), *(voi. III,* 1998, ISBN 973-34-0497-7, pag. 5-414) cu ortografie actualizată, fără note.

Primul pas în analiza noastră a fost evaluarea frecventelor relative ale cuvintelor din corpusurile menționate anterior. Tabelul 1 conține primele 55 de cuvinte din ierarhia frecvențelor relative din diverse corpusuri.

Tabel 1

Ierarhia frecvențelor relative în câteva corpusuri
0. Rang; 1. Cuvânt; 2. Frecvență relativă (în %)

	#Cl\/IG		#CLG		#1		#4+#5+#6		#6		#CSG	
0	1	2	1	2	1	2	1	2	1	2	1	2
1	de	4,10	de	4,02	de	4,17	de	4,17	de	4,12	de	4,87
2	si	3,20	si	3,12	Şi	2,65	în	2,55	în	2,58	în	3,47
3	în	2,67	în	2,44	în	2,50	si	2,39	si	2,58	si	3,07
4	să	1,62	să	1,87	cu	1,75	să	1,81	0	1,94	a	2,35
5	a	1,47	la	1,52	0	1,62	0	1,73	să	1,69	la	1,52
6	la	1,46	cu	1,50	a	1,47	la	1,55	cu	1,52	se	1,46
7	se	1,39	pe	1,45	la	1,43	cu	1,48	la	1,46	cu	1,21
8	cu	1,38	se	1,43	se	1,42	nu	1,41	se	1,44	care	1,17
9	0	1,30	0	1,41	pe	1,39	pe	1,39	pe	1,41	0	0,87
10	nu	1,28	nu	1,33	nu	1,37	se	1,35	nu	1,27	din	0,85
11	pe	1,27	a	1,17	să	1,33	un	1,18	un	1,25	pe	0,82
12	care	0,98	că	1,05	un	1,26	că	1,08	a	1,04	este	0,79
13	că	0,97	un	0,99	că	1,04	a	1,05	care	0,95	mai	0,75
14	mai	0,95	mai	0,97	lui	0,88	care	0,95	că	0,95	nu	0,73
15	din	0,91	din	0,94	mai	0,86	din	0,93	din	0,91	sau	0,71
16	un	0,87	care	0,89	din	0,86	ce	0,85	mai	0,84	să	0,70
17	ce	0,66	ce	0,69	care	0,84	mai	0,84	ce	0,79	pentru	0,67
18	ca	0,60	ca	0,58	ioanide	0,74	lui	0,68	pentru	0,71	că	0,54
19	pentru	0,54	lui	0,54	era	0,66	era	0,63	ui	0,71	al	0,53
20	ui	0,49	dar	0,51	ce	0,64	pentru	0,63	era	0,64	un	0,50
21	dar	0,45	era	0,51	e	0,54	ca	0,55	charlie	0,59	prin	0,44
22	fi	0,42	pentru	0,48	ca	0,53	ei	0,53	ei	0,57	ca	0,43
23	este	0,42	fi	0,42	fi	0,52	dar	0,50	dar	0,54	fi	0,35
24	era	0,39	când	0,39	pompo-nescu	0,44	fi	0,49	ca	0,53	sunt	0,33
25	sau	0,35	el •	0,38	pentru	0,40	fi	0,43	îi	0,53	ale	0,32
26	e	0,34	e	0,37	când	0,35	ei	0,42	ea	0,51	poate	0,29
27	el	0,34	am	0,35	el	0,33	ea	0,38	el	0,50	sa	0,29
28	al	0,33	ei	0,32	prin	0,27	când	0,34	fi	0,45	au	0,28
29	când	0,33	nici	0,30	am	0,26	nici	0,33	kurtz	0,36	ce	0,27
30	ei	0,29	îi	0,29	după	0,26	cum	0,31	când	0,34	art	0,27
31	am	0,28	mă	0,28	*1*	0,26	e	0,30	nici	0,32	fost	0,24
32	nici	0,28	cum	0,28	al	0,26	aSa	0,30	(cum	0,30	după	0,24
33	prin	0,28	sau	0,25	nici	0,25	dacă	0,29	'Si	0,28	dacă	0,21
34	sa	0,26	fost	0,25	1ară	0,25	îl	0,29	îl	0,28	c	(0,19
35	sunt	3,25	(după	0,24	'i	0,25	charlie	0,29	a	0,28	când	(3,19

36	cum	0,25	sa	0,24	avea	0,24	fost	0,28	aSa	0,26	m	0,18
37	fost	0,25	dacă	0,24	ar	0,23	barley	0,28	dacă	0,25	unei	0,17
38	dacă	0,24	al	0,24	dacă	0,23	al	0,27	e	0,25	cele	0,17
39	după	0,24	ea	0,23	gaittany	0,22	spuse	0,26	oseph	0,24	pot	0,16
40	au	0,23	asa	0,22	însă	0,22	iar	0,26	sau	0,24	are	0,16
41	îi	0,23	îl	0,22	foarte	0,21	însi	0,26	ai	0,23	penală	0,16
42	mă	0,21	însi	0,21	spre	0,20	este	0,26	după	0,23	trebuie	0,16
43	ea	0,21	este	0,21	ei	0,20	după	0,25	te	0,22	această	0,16
44	iar	0,19	au	0,21	asa	0,20	am	0,25	sa	0,22	iui	0,16
45	poate	0,19	sunt	0,20	sunt	0,19	sau	0,24	fără	0,22	acest	0,16
46	asa	0,18	iar	0,20	cum	0,19	ai	0,24	fost	0,22	iar	0,15
47	ar	0,18	fără	0,19	dar	0,19	ar	0,24	ar	0,22	lor	0,15
48	fără	0,18	prin	0,19	hagienuş	0,19	sa	0,23	este	0,21	numai	0,15
49	însi	0,17	ar	0,19	sau	0,18	te	0,20	iar	0,21	dar	0,15
50	îl	0,17	le	0,18	fost	0,18	avea	0,20	le	0,21	mare	0,15
51	le	0,17	asta	0,18	toate	0,18	le	0,20	spuse	0,21	cel	0,14
52	ale	0,17	tot	0,18	este	0,18	leamas	0,20	apoi	0,20	unor	0,14
53	toate	0,17	eu	0,18	sa	0,18	timp	0,20	timp	0,20	fie	0,14
54	va	0,16	acum	0,17	îşi	0,17	apoi	0,19	lor	0,19	va	0,14
55	decât	0,16	până	0,17	qonzalv	0,17	au	0,18	săi	0,19	între	0,13

Un alt rezultat experimental interesant este identificarea unui număr d 162 de cuvinte care se regăsesc în toate cele 93 de cărți ce alcătuiesc corpusu (fie că este vorba de literatură, medicină, drept etc). Deşi sunt doar 162, acest cuvinte au o pondere importantă în textul global **#CMG** acoperind circa 45% di totalul celor 8806433 cuvinte. Aceste cuvinte comune împreună cu rangul lor î ierarhie şi frecvențele lor relative în întreg textul **#CMG** sunt conținute în Tabelul 2

Tabel

Lista cuvintelor comune în toate cele 93 **de cărți**
1. **Cuvânt;** 2. **Rangul cuvântului în ierarhia frecvențelor relative în textul mix global, #CMG;** 3. **Frecvența relativă a cuvântului în textul mixt global, #CMG (în %)**

1	2	3	1	2	3	1	2	3	1	2	3
de	1	4,10	iar	44	0,19	unei	93	0,10	sar	185	0,0
si	2	3,20	poate	45	0,19	atunci	94	0,10	una	187	0,0
în	3	2,67	asa	46	0,18	două	95	0,10	început	188	0,0
să	4	1,62	ar	47	0,18	doar	96	0,10	încât	193	0,0
a	5	1,47	fără	48	0,18	dintre	100	0,10	alte	196	0,0
la	6	1,46	însi	49	0,17	are	101	0,10	acestea	198	0,0
se	7	1,39	îl	50	0,17	face	102	0,10	facă	199	0,0
cu	8	1,38	le	51	0,17	sub	104	0,09	altă	200	0,0
0	9	1,30	ale	52	0,17	nimic	106	0,09	acelaşi	204	0,0

		1,28	toate	53	0,17	fei	107	0,09	desi	206	0,04
pe	11	1,27	va	54	0,16	ia	108	0,09	fac	213	0,04
care	12	0,98	decât	55	0,16	puţin	109	0,09	printre	220	0,04
	13	0,97	tot	56	0,16	între	110	0,09	pare	224	0,04
mai	14	0,95	[lor	57	0,16	intru n	111	0,09	partea	225	0,04
din	15	0,91	spre	58	0,15	cea	112	0,09	afară	226	0,04
un	16	0,87	pana	59	0,15		113	0,08	sus	240	0,04
ce	17	0,66	chiar	60	0,15	săi	116	0,08	faptul	246	0,03
ca	18	0,60	mult	61	0,14	aceea	117	0,08	locul	252	0,03
pentru	19	0,54	cel	63	0,14	ci	119	0,08	adevărat	260	0,03
lui		0,49	fie	65	0,14	fată	126	0,08	tuturor		
dar		0,45	ne	66	0,14	unul	127	0,08	măcar	266	0,03
		0,42	ai	67	0,14	astfel	128	0,08	primul	268	0,03
este	23	0,42	acum	68	0,14	parte	129	0,08	aceeaşi	275	0,03
era		0,39	trebuie	69	0,14	înainte	132	0,07	altfel	277	0,03
sau	25	0,35	cele	70	0,13	pot	134	0,07	noua	299	0,03
	26	0,34	numai	72	0,13	ele	138		acela	302	0,03
	27	0,34	despre	73	0,12	totul	140	0,07	trebui	307	0,03
ai	28	0,33	avea	74	0,12	dată	141	0,07	dintro	330	0,03
când	29	0,33	atât	75	0,12	tOt	143	0,07	dă	358	0,02
ei	30	0,29	această	76	0,12	loc	144	0,07	afla	364	0,02
nici	32	0,28	putea	78	0,12	fiecare	153	0,06	ramane	371	0,02
prin	33	0,28	unde	80	0,12	orice	155	0,06	alt	373	0,02
sa	34	0,26	mtro	81	0,11	spune	165	0,06	pus	377	0,02
sunt	35	0,25	acest	82	0,11	asemenea	166	0,06	întâi	387	0,02
cum	36	0,25	noi	84	0,11	sale	167	0,06	rând	397	0,02
fost	37	0,25	sai	86	0,11	acesta	168	0,06	alta	404	0,02
dacă	38	0,24	cat	87	0,11	lucru	169	0,06	legătură		
după	39	0,24		88	0,11	către	174	0,05		429	0,02
au	40			89	0,11	multe	175	0,05	rândul	601	0,01
n	41	0,23	ceva	91	0,11	celor	178	0,05			
ea	43	0,21	insa	92	0,10	totuşi	184	0,05			

2. Descrierea structurii statistice de cuvinte. Studiu bazat pe multiple intervale de încredere statistică

Fie un text natural **considerat ca succesiune de cuvinte** pe care îl eşantionăm cu o perioadă suficient de mare astfel încât să rupem practic dependenţa dintre observaţiile succesive. Iniţial în investigaţia noastră statistică am considerat această perioadă ca fiind de 200 cuvinte. La fiecare moment de eşantionare am înregistrat observaţia făcută (cuvântul respectiv), conform Fig. 1. Mulţimea de date obţinute în acest fel conţine A/-cuvinte, unde $N = L_c/200$ iar L_c este lungimea textului în cuvinte.

	1	201	401
	CÂND GAITTANY	TĂCU CĂCI ŞTIA	GAITTANY LIPSIT DE
1.	CÂND	CĂCI	LIPSIT...
2.	GAITTANY	STIA	DE
200.		TĂCU	GAITTANY

Figura 1. 200 de mulţimi de date (cuvinte) în model statistic /./.cf. obţinute prin eşantionare periodică a textului natural

Deplasând originea eşantionării în textul natural apar 200 de astfel mulţimi de date experimentale, fiecare în parte de volum A/, Fig. 1.

Fiecare mulţime de *N* observaţii astfel obţinută satisface modelul stati *i.i.d.,* model necesar în aplicarea inferenţelor statistice utilizate. Independenţa e asigurată de mărimea perioadei de eşantionare; distribuţia identică este un rezu al ipotezei de staţionaritate a limbii naturale.

Acceptând ipoteza de staţionaritate a limbii, toate cele 200 de mulţimi date experimentale (compatibile cu modelul *i.i.d.)* extrase din textul natu conform Fig. 1, trebuie să conţină aceeaşi informaţie despre probabilita cuvântului investigat (oricare ar fi acesta).

Atenţie însă, aceste mulţimi de date nu sunt independente între ele

Un prim obiectiv al studiului nostru a fost de a vedea dacă într-adevăr c 200 mulţimi de date confirmă sau nu aceeaşi probabilitate p teore (necunoscută) a cuvântului investigat.

Un răspuns afirmativ ne-ar permite să obţinem un model matematic pe sursa de informaţie de cuvinte asociată limbii române. Pentru a da un răspuns extins o procedură statistică pe care am dezvoltat-o în [3]-[8] pentru /n-grame. P această procedură cele 200 de mulţimi de date experimentale se compară între aplicând repetat un test statistic al ipotezei că probabilitatea aparţine unui inte dat, vezi Anexa 1.

Menţionăm că nu am putut face o comparaţie pe baza unui test mai folosit, anume acela privind egalitatea între doua probabilităţi, întrucât mulţimile date care se compară nu sunt independente între ele.

Procedura a permis în final determinarea unui interval de încred statistică optim care a fost denumit în continuare *"reprezentativ"* pentru cuvâ urmărit şi textul natural. Simultan a apărut şi mulţimea de date experimentale *i.*

"reprezentativă" pentru cuvântul respectiv şi textul natural, mulțime ce va fi folosită în comparații matematice între texte naturale.

2.1. Intervale de încredere statistică *"reprezentative"* pentru probabilitățile cuvintelor. Metodă de determinare şi rezultate experimentale

Scopul acestui subcapitol este de a determina probabilitatea p a unui un cuvânt urmărit.

Fie ***mț*** numărul de apariții ale cuvântului în mulțimea / de date experimentale *i.i.d.* de volum *N, / =* **1*200.** (Aceste mulțimi sunt extrase din textul natural conform Fig. 1.)

Aplicând teoria estimării, fiecare din cele 200 de mulțimi de date conduce la o estimație p;=m;/N a probabilității p necunoscute şi la un interval de încredere statistică al probabilității //=(pi,/;p2,/)» i = 1 + 200. Considerând *N* suficient de mare astfel încât condiția deMoivre - Laplace să fie satisfăcută, *Np(l-p)»l,* limitele intervalului de încredere statistică (inferioară şi superioară) se calculează conform relației (1), [9], [10]:

Plj = Pi~ZalliP&-Pi)i' *P2J = Pi + Za/2^Pi(~Pi)/'* (1)

unde *zj2* este a/2 cuantila legii normale de medie 0 şi dispersie 1. în determinările noastre experimentale am lucrat cu un nivel de încredere statistică de 95%; rezultă *zji* =1.96 .

Cu alte cuvinte putem spune că probabilitatea adevărată p se află în intervalul = *[p\j\P2j)*> cu o încredere statistică egală cu 0,95.

într-o primă etapă a analizei noastre, pentru un anumit eveniment urmărit (apariția unui cuvânt), s-au folosit următoarele mărimi (a se vedea Fig. 2):

- - frecvența relativă a cuvântului pe întreg textul natural considerat (ceea ce înseamnă măsurare din date corelate); p* este raportul între numărul de apariții ale cuvântului în textul natural şi lungimea L_e a textului respectiv (numărul total de cuvinte). Se observă că *p** este media aritmetică a celor 200 de estimații. Subliniem că p* este o mărime importantă pentru orice experimentator.

  Pmin = min£;, / = 1 + 200 valoarea minimă a estimatiilor;

  PmsiX = maxp/, i = 1+200 - valoarea maximă a estimatiilor;

$A_M = \max p_{i_2},,-\min p_{i_n}$, i-1+200 - reuniunea celor 200 de intervale

de încredere statistică;

$A°_M = \max_i p; - \min_i pj,$ $i = 1 + 200$ diferența maximă între două

estimații (intervalul de împrăştiere al estimatiilor);

δ_M =max|p; -/?*[, / = 1 + 200 - diferența maximă între estimațiile *p-*

frecvența relativă *p *;*

S_m =*mm\pi* , / = l + 200 - diferența minimă între estimațiile p_x

frecvența relativă *p **.

Pl *h* *h*

FIA *PA* *Pih*

AM *P200* ^200

Figura 2. Mărimi utilizate în obținerea intervalului de încredere statistică *"reprezentativ"* pentru probabilitate

Următoarele întrebări (probleme) au ghidat analiza noastră teoretică experimentală:

1. **Cât de largi sunt intervalele , S_M şi A_M ?** Intervalele k^e_M S_M sunt importante în analiza împrăştierii estimatiilor în jurul va *p**. Intervalul A_M ne dă o idee despre cel mai mare interval în c s-ar afla p, probabilitatea adevărată, bănuită că există.
2. **Există valori *pi* foarte apropiate de şi cât de apropiate?**

Pentru a răspunde la această întrebare a fost urmărită experimental mărimea δ_m. S_m conduce la estimația p; care este cea mai apropiată de p*, estimație care va fi în continuare notată cu p_A. S-a notat cu A intervalul de încredere statistică asociat estimației p_A conform relației (1).

3. **Cât de multe intervale de încredere statistică // conțin (îmbracă) p*?** Prin presupunerea de staționaritate, ne aşteptăm ca un mare număr de intervale de încredere statistică // să se intersecteze, conținându-l în acelaşi timp pe p*. Nu ne aşteptăm la o proporție de 95%, întrucât cele 200 de mulțimi de date *Ud.* nu sunt independente între ele.

4. Putem găsi un interval de încredere statistică pentru probabilitatea adevărată p, interval care să fie acceptat de toate cele 200 de mulțimi de date experimentale? Se pot găsi mai multe astfel de intervale? Dacă ipoteza de staționaritate este adevărată atunci astfel de intervale trebuie să existe. Dacă intervalul A (definit mai sus în întrebarea 2) este unul dintre aceste intervale, atunci el va fi preferat de experimentator şi va fi considerat ca *"reprezentativ"* pentru probabilitatea cuvântului şi textul analizat.

Pentru a înțelege metoda dezvoltată de noi care răspunde la aceste întrebări şi care conduce la obținerea intervalului de încredere statistică *"reprezentativ"*exemplificăm pentru cuvântul DE în corpusul mixt global **#CMG.**

în Tabelul 3 cele 200 de rânduri corespund celor 200 de mulțimi de date experimentale *i.i.d.* fiind explicitate atât estimațiile p; cât şi intervalele de încredere statistică //, / = 1 + 200. Succesiv, fiecare interval // a fost considerat ca interval de referință şi am aplicat 199 de teste ale ipotezei că probabilitatea aparține intervalului menționat, test descris în Anexa 1. Fiecare dintre cele 199 de teste este aplicat pe o singură mulțime de date experimentală. (Numărul 199 se explică prin faptul că nu se testează şi mulțimea care a produs intervalul de referință.)

în primul rând al tabelului 3, intervalul *1*\ =(3,89; 4,27)xl0⁻^ este intervalul de referință față de care se aplică testele de apartenență a probabilității. Se testează dacă probabilitatea cuvântului DE aparține sau nu intervalului /j pe baza unei singure mulțimi de date *i.i.d.]* aceasta înseamnă că verificăm succesiv fiecare din restul mulțimilor de date, anume / = 2 + 200. Acceptarea ipotezei că probabilitatea cuvântului DE aparține intervalului *1*\ este marcată cu "DA" pentru respectiva mulțime de date (Tabelul 3, rândul 1). în caz contrar, pe poziția respectivă este completat "NU". Numărul total de mulțimi de date care trec testele este conținut în ultima coloană din dreapta. Această procedură se repetă alegând ca referință pe rând toate cele 200 de intervale de încredere statistică //, / = 1 + 200.

Tabel 3

Determinarea intervalului de încredere statistică *"reprezentativ"* A pentru ^ M I L » cuvântului DE în corpusul mixt global #CMG.Este îngroşat rândul 3 care corespunde intervalului *"reprezentativ* A

i	$\frac{\hat{p}_i}{(\times 10^{-2})}$	$I_i(\times 10^{-2})$		Mulțime i										Tota
		$p_{1,i}$	$p_{2,i}$	1	2	3	...	99	...	97	...	200	DA	
1	4,08	3,89	4,27		DA	DA		DA		DA		DA	199	
2	4,06	3,87	4,25	DA		DA		DA		DA		DA	199	
3	**4,10**	**3,91**	**4,29**	**DA**	**DA**			**DA**		**DA**		**DA**	**199**	
...														
94	4,31	4,12	4,51	DA	DA	DA				NU		DA	182	
...														
99	3,78	3,60	3,96	DA	DA	DA		NU				DA	121	
...														
200	4,18	3,99	4,37	DA	DA	DA		NU		DA			198	

Pentru cuvântul DE s-a obținut p* = 0,040986, iar estimația cea m apropiată de *p** a fost $p_3 = p_A$ = 0,040992, S_m/p^* = 0,0002. Pentru estimația se obține intervalul de încredere statistică 95% A = (0,0391; 0,0429). Din rându se observă că intervalul A trece toate cele 199 de teste ale ipotezei probabilitatea cuvântului DE este cuprinsă în interiorul său. Sunt multe intervale care au compatibilitate cu toate mulțimile de date *i.i.d.* (în ultima coloană numă 199 a apărut de 101 ori). Dintre aceste 101 intervale am ales A = (o,0391; 0,042 ca fiind interval de încredere statistică 95% *reprezentativ* pentru probabilita cuvântului DE întrucât este uşor de determinat de oricare experimenta Mulțimea de date *i.i.d.* specificată de indicele i = 3 va fi numită mulțime de d *"reprezentativă"* pentru cuvântul DE în corpusul **#CMG.**

Tabelul 4 conține informații despre elementele analizei pentru primele ze cuvinte din ierarhia frecvențelor relative în corpusul **#CMG.** Exemplificăm pen cuvântul DE care având frecvența relativă p* = 4,10xl0~² este pe primul lo ierarhie. Valoarea p* este cuprinsă în $N(p^*)$= 192 de intervale de încred statistică din cele 200 considerate, (coloana 3); reuniunea celor 200 intervale încredere statistică raportată la p* este A^/p* = 22,24xl0~², (coloana diferența maximă între două estimații raportată la p* este A^{c}_{M}/p* = 13,08xl (coloana 5); diferența maximă între o estimație p, şi p* raportată la p* • <5ₘ/p* = 7,80xl0~², (coloana 6); diferența minimă între o estimație p_i şi

raportată la p* este $S_m / p^* = 0.02xIO^{-2}$, (coloana 7); lărgimea intervalului de încredere statistică *"reprezentativ"* A raportată la p^* este $A/p^* = 9{,}23 \times 10^{-2}$, (coloana 8); există N(A)=101 intervale de încredere statistică la fel de bune ca intervalul A *"reprezentativ",* (coloana 9). Aceste *N(A)* intervale sunt confirmate de toate cele 199 de teste de apartenență a probabilității la interval, prin care s-a făcut verificarea staționarității.

Numărul relativ mare de intervale de încredere statistică confirmate practic de toate cele 199 de teste de apartenență a probabilității la interval - *N(A)* din coloana 9 a tabelului 4 - este o susținere a ideii de staționaritate.

Tabel 4

Rezultate numerice privind mărimile din Fig. 2 pentru cele mai frecvente 10 cuvinte în #CMG.

Cuvânt	p *	N(p*)	A_M/p^*		$^5M/P^*$		A/p*	N(A)
1	2	3	4	5	6	7	8	9
DE	4,10	192	22,24	13,08	7,80	0,02	9,23	101
SI	3,20	198	25,42	14,91	8,77	0,04	10,45	102
IN	2,67	194	24,88	13,43	7,09	0,04	11,43	172
SA	1,62	189	35,74	21,02	11,40	0,05	14,69	122
A	1,47	191	37,11	21,66	11,40	0,05	15,42	124
LA	1,46	185	38,07	22,52	12,89	0,00	15,45	104
SE	1,39	190	39,97	24,27	13,92	0,02	15,89	79
CU	1,38	195	37,40	21,55	11,30	0,05	15,92	132
0	1,30	191	37,52	21,29	12,38	0,02	16,39	120
NU	1,28	189	40,20	23,62	12,78	0,01	16,54	120

Prin centralizarea acestor tipuri de rezultate pentru toate corpusurile analizate şi pentru toate cuvintele pentru care s-a putut face analiza a rezultat Tabelul 5. Concret, studiul experimental a cuprins toate corpusurile prezentate în Introducere. Am putut aplica inferențele statistice doar pentru acele cuvinte pentru care am avut suficiente date; anume $Np^*(1-p^*)>20$, unde *N* este volumul mulțimii de date *i.i.d.* (forma experimentală pentru condiția DeMoivre - Laplace). Cuvintele au fost sortate în ordine descrescătoare a frecvențelor de apariție p*. Această sortare a permis organizarea studiului pe clase de frecvență. Am ales ca limite ale claselor următoarele valori: 5%, 2%, 1%, 0,5%, 0,2%, 0,1% şi 0,05%.

în studiul nostru experimental în aproape toate situațiile (oricare cuvânt urmărit şi orice corpus lingvistic investigat) am găsit o estimație p_A practic egală cu p*. Acest lucru se vede în Tabelul 5, coloana 8 urmărind raportul dintre 8_M şi p*. Pentru toate situațiile analizate am obținut $£_m/p^*<2{,}23\%$. Având în vedere

că studiul experimental a condus şi la obținerea de intervale A *"reprezentative"* i toate situațiile analizate, rezultă că aceste intervale de încredere statistică 95% p fi scrise sub forma:

$$\mathbf{A}=(p_{1A};p_{2A}) = p^*(1+£_r), \qquad e_r s 1.96 x V(1-p^*)/(\#/>^*) \qquad (2)$$

$£_r$ este eroarea relativă cu care se determină probabilitățile.

Exemplificăm citirea Tabelului 5 pentru corpusul #CMG şi clasa a doua d frecvență. Există 8 cuvinte (coloana 3) care au frecvențele relative cuprinse înt (0,01; 0,02). Aceste 8 cuvinte acoperă 11,17% (coloana 4) din totalul aparițiilor c cuvinte din #CMG, $L_c = 8806433$. Celelalte coloane, 5-9, conțin informa referitoare la mărimile din Fig. 2. Astfel coloana 9 conține raportul dintre lungim intervalului A şi p* pentru cuvintele existente în clasa respectivă (limita minimă maximă). Acest raport este practic dublul erorii relative, $£_r$, în determinar probabilității cuvântului; se observă o precizie relativ bună a determinărilor c această clasă, $e_r<8.5xI0^{-2} = 17xI0^{-2}/2$.

în total în #CMG au fost 194 = 3 + 8 + 8 + 24+59 + 92 cuvinte pentru care s putut determina intervalul A *"reprezentativ".* Deşi cele 194 cuvinte reprezintă mică pondere din totalul cuvintelor distincte posibile, ele acoperă 48,87% d L_c =8806433, totalul aparițiilor de cuvinte în corpusul mixt global, #CMG.

Tabe

Rezultate experimentale organizate pe clase de frecvențe relative. Valoril

Clasa de frecvențe	Corpus	Nr.	Aco-perire	A_M//?*	»——r	$8_M/p$	$8_M/P^*$	A/p
1	2	3	4	5	6	7	8	9
$2x10^{-2}<p^*<5x10^{-2}$	#CMG	3	9,97	22-25	13-15	7-9	0,02-0,04	9
	#1JCMG	3	10,04	31-39	18-22	10-13	0,01-0,07	13
	#2JCMG	3	9,90	30-41	17-25	9-14	0,01-0,06	13
	#CLG	3	9,58	27-33	16-19	8-11	0,03-0,04	11
	#1JCLG	3	9,60	39-47	23-27	12-14	0,05-0,10	16
	#2JCLG	3	9,55	38-46	22-25	12-13	0,03-0,10	16
	#CŞG	4	13,76	52-83	27-52	14-27	0,05-0,37	25
$10^{-2}<p^*<2x10^{-2}$	#CMG	8	11,17	36-40	21-24	11-14	0,00-0,05	15
	#1JCMG	9	12,07	49-59	28-33	14-18	0,00-0,20	21
	#2JCMG	8	11,28	46-60	24-37	12-19	0,10-0,16	21
	#CLG	9	12,73	39-52	23-30	12-17	0,00-0,09	10
	#1JCLG	9	12,69	55-78	30-48	16-25	0,02-0,26	23
	#2JCLG	10	13,77	51-75	28-44	14-24	0,00-0,25	23
	#CŞG	4	5,36	110-130	67-78	34-48	0,24-0,59	44
	#CMG	8	6,48	42-61	23-36	12-20	0,00-0,16	1
	#1JCMG	8	5,93	60-92	33-62	17-34	0,11-0,28	2

$5x10^{-3} \le p^* < 10^{-2}$	#2JCMG	8	6,51	58-84	31-52	15-29	0,01-0,26	27-36
	#CLG	9	6,62	53-77	30-45	15-27	0,03-0,27	22-31
	#1JCLG	8	6,15	73-112	40-68	23-42	0,01-0,54	31-44
	#2JCLG	9	6,14	72-107	40-66	24-39	0,06-0,42	32-44
	#CŞG	12	8,46	135-186	71-117	36-64	0,35-1,45	58-78
$2x10^{-3} \le p^* < 5x10^{-3}$	#CMG	24	7,36	59-103	31-63	16-33	0,01-0,43	27-41
	#1JCMG	24	7,17	90-158	51-100	27-59	0,01-0,82	39-60
	#2JCMG	23	7,05	91-143	49-87	25-45	0,04-0,84	38-60
	#CLG	25	7,02	74-123	42-75	21-40	0,05-0,73	32-50
	#1JCLG	23	6,87	116-186	65-116	37-66	0,02-1,33	46-71
	#2JCLG	23	6,40	105-186	57-115	29-66	0,02-1,20	48-71
	#CŞG	2	0,87	198-198	111-112	69-72	1,08-2,14	82-83
$10^{-3} \le p^* < 2x10^{-3}$	#CMG	59	7,78	92-156	46-97	26-61	0,02-1,09	43-61
	#1JCMG	57	7,40	135-224	74-141	41-80	0,02-2,07	61-87
	#2JCMG	61	8,13	135-217	75-134	40-77	0,04-2,23	61-87
	#CLG	58	7,74	114-190	62-120	33-67	0,01-1,41	51-73
	#1JCLG	33	5,21	158-231	81-143	42-85	0,07-1,84	73-91
	#2JCLG	32	5,06	160-220	85-129	46-82	0,04-2,11	72-91
	#CŞG	0	0,00	-	-	-	-	-
$5x10^{-4} \le p^* < 10^{-3}$	#CMG	92	6,11	134-224	72-142	37-82	0,02-2,16	62-90
	#1JCMG	8	0,72	185-221	101-134	52-76	0,30-1,89	88-90
	#2JCMG	9	0,81	194-242	111-150	57-95	0,25-1,89	88-91
	#CLG	52	4,11	162-224	88-131	45-82	0,05-1,95	73-90
	#1JCLG	0	0,00	-	-	-	-	-
	#2JCLG	0	0,00	-	-	-	-	-
	#CŞG	0	0,00	-	-	-	-	-
Total	#CMG	194	48,87					
	#1JCMG	109	43,33					
	#2JCMG	112	43,68					
	#CLG	156	47,80					
	#1JCLG	76	40,52					
	#2JCLG	77	40,92					
	#CŞG	22	28,45					

Tabelul 5 indică şi precizia determinărilor (eroarea relativă *e,)* pentru cuvintele analizate. Această precizie este relativ bună pentru determinările făcute pe corpusul mixt global **#CMG** (pentru cuvinte din primele patru clase de frecvenţă, $e, < 20{,}5 \times 10^{-2} = 41 \times 10^{-2}/2$).

Aplicând procedura descrisă în Cap. 2.1 pentru toate corpusurile lingvistice şi pentru toate cuvintele care au satisfăcut condiţia deMoivre - Laplace au rezultat probe în sprijinul ipotezei de staţionaritate a limbii române scrise.

Intervalele *"reprezentative"* precum şi mulţimile de date *i tative"* determinate pentru un cuvânt anumit şi textul natural consid continuare folosite în Cap. 2.2, pentru a analiza dacă putem vorbi de matematic al sursei de cuvinte pentru limba ca ansamblu, pentru d ale limbii, pentru diverşi autori etc.

Acurateţea în determinarea probabilităţii cuvintelor este dată

- încrederea statistică (95%);
- erorile relative, *e,,* cu care s-au obţinut intervalele A conform Tabelului 5;
- mărimea celor două tipuri de erori statistice care ap apartenenţă a probabilităţii la interval, întrucât acest tip baza validării intervalului A ca *"reprezentativ".*

în ceea ce priveşte testul de apartenenţă a probabilităţii la in fost aplicat pentru un prag statistic $a = 0{,}05$. întrucât testul a fost tr dată (fapt pentru care A a fost validat ca *"reprezentativ')* este impor un control asupra mărimii *(3,* probabilitatea de a accepta date f impune valori mici pentru *li am* avea nevoie de un corpus m exemplu, conform [6, Tabel 4], dacă se doreşte *fi* < 0,3 şi 5 = pentru a investiga cuvinte din primele patru clase de frecvenţă am a un corpus de circa 30 de milioane de cuvinte.

2.2 Comparaţii matematice între diverse texte natural structurii de cuvinte

Investigaţia noastră (privind staţionaritatea) a fost completat matematice privind probabilităţile cuvintelor, pe care le-am c următoarele criterii:

a) Se verifică dacă un acelaşi cuvânt are aceeaşi proba două texte naturale care se compară. Această compara în continuare *comparaţie între cuvinte ca atare.*

b) Se verifică dacă probabilităţile cuvintelor situate pe un ierarhia frecventelor relative din cele două texte su exemplu, pe rangul 20 în corpusul literar global se află iar în corpusul ştiinţific cuvântul UN, vezi Tabelul 1. între cele doua domenii se va urmări dacă probabilita cuvinte (DAR şi UN) este aceeaşi. în cele ce urmeaz criteriu *comparaţie pe baza rangului.*

Toate comparaţiile matematice, atât pe baza criteriului a) criteriului b) au fost făcute folosind următoarele teste statistice:

- *T1* - test al ipotezei ca probabilitatea aparţine unui inte
- 72 - test de egalitate între două probabilităţi, (Anexa 2)

Pentru fiecare din cele două texte naturale care se compară şi pentru fiecare cuvânt investigat s-au determinat în prealabil intervalele *"reprezentative"* precum şi mulţimile de date *i.i.d. "reprezentative".*

Când aplicăm testul 77, intervalul *(a;b)* este intervalul *"reprezentativ"* A din primul text natural implicat în comparaţie, iar mulţimea *[xix₂^>.,xₙ* de date experimentale *i.i.d.* este mulţimea de date *"reprezentativă"* din cel de-al doilea text natural. Testul a fost aplicat în ambele situaţii: corpusl *versus* corpus2 şi corpus2 *versus* corpusl, Tabel 6.

Când aplicam testul 72 se considera pentru comparaţie cele doua mulţimi de date *i.i.d. "reprezentative"* extrase din cele două texte naturale pentru cuvintele care se compară.

Toate testele au fost aplicate pentru un prag de semnificaţie statistică a**=0,05** . Cu alte cuvinte, probabilitatea de a respinge date corecte este mai mică decât 0,05.

Tabel 6

Comparaţii între texte naturale pe baza probabilităţii cuvintelor.
Coloanele 4-9 conţin numărul de cuvinte rejectate de testele statistice

Texte cc>mparate		Nr.	*Comparaţie între cuvinte ca atare*			*Comparaţie pe baza rangului*		
Corpus 1	Corpus 2		Test 77		Test 72	Test 77		Test 72
			1 versus 2	2 versus 1		1 versus 2	2 versus 1	
7	2	*3*	*4*	*5*	*6*	7	*8*	*9*
#1JCLG	#2JCLG	72	0	0	0	0	0	0
#1JCMG	#2JCMG	104	0	0	0	0	0	0
#CLG	#CSG	22	10	18	13	1	16	10

Rezultatele experimentale sunt sintetizate în Tabelul 6. Comparaţiile făcute în cadrul domeniului literar, când se compară cele două jumătăţi de corpus între ele **(#1JCLG** şi **#2JCLG)** nu indică diferenţe între probabilităţi indiferent de testul utilizat (77 sau 72) sau de criteriul utilizat (comparaţii pe baza aceluiaşi cuvânt sau pe baza aceluiaşi rang).

Acelaşi rezultat s-a obţinut şi când s-au comparat cele două jumătăţi ale corpusului mixt global, #1**JCMG** şi #2JCMG.

Exemplificăm în continuare modul de citire al Tabelului 6.

Primele două coloane conţin corpusurile care se compară între ele.

Coloana 3 indică numărul de cuvinte investigate în comparaţii (care au îndeplinit condiţia *Np*(l-p*)*>20 în ambele texte care se compară).

Rezultatele din coloanele 4, 5 şi 6 au fost obţinute aplicând criteriul comparaţiilor "cuvintelor ca atare".

Coloanele 4 şi 5 arată câte cuvinte nu au trecut testul 77 de apartenenţă probabilităţii la interval. Coloana 4 se referă la situaţia când intervalul fix *[a\b)* es intervalul A *"reprezentativ"* din primul corpus al comparaţiei, iar mulţimea de da *i.i.d.* supusă testului este mulţimea *Ud. "reprezentativă"* din al doilea corpu Similar, în coloana 5: intervalul fix *(a\b)* este intervalul A *"reprezentativ"* din doilea corpus al comparaţiei, iar mulţimea de date *i.i.d.* supusă testului es mulţimea *i.i.d. "reprezentativă"* din primul corpus,

Coloana 6 conţine numărul de cuvinte care sunt rejectate de testul 72 egalitate între probabilităţi.

Coloanele 7, 8 şi 9 conţin acelaşi tip de informaţie specificat în coloanele 5 şi 6, cu diferenţa că de această dată se compară cuvintele care ocupă acel rang în loc de cuvintele "ca atare".

Când se compară domenii diferite, spre exemplu literar şi ştiinţific, ap multe diferenţe marcate de ambele teste 77 si 72 şi de cele două criterii comparaţie.

Rezultatele comparaţiilor punctează unele diferenţe între domeniile lite şi ştiinţific. Testele nu au indicat diferenţe când s-au comparat corpusuri organiz după aceeaşi reguli (jumătăţile corpusului mixt global între ele sau jumătăţ corpusului literar global între ele); reamintim că atât corpusul mixt global cât şi literar global au fost obţinute prin concatenarea aleatoare a cărţilor respective.

3. Legea lui Zipf. Studiu experimental

Ierarhiile frecvenţelor relative ale cuvintelor (prezentate în Cap. 1 şi întâ de analiza de staţionaritate din Cap. 2) au constituit o bază de plecare pen studiul nostru experimental asupra legii lui Zipf. în lingvistică legea lui Zipf este u din cele mai cunoscute dependenţe rang - frecvenţă. (Aceste dependenţe ran frecvenţă au fost observate de-a lungul timpului şi în diverse alte dome economie, fizică, biologie, demografie etc. [11], [12].) Obiectivul acestui capito fost de a stabili dacă şi în ce măsură (cu ce acurateţe) limba română scr satisface legea lui Zipf.

Fie un text (corpus) având o lungime de L_c cuvinte, dintre care N_c s distincte. Aceste N_c cuvinte se sortează într-o listă în ordine descrescătoar numărului de apariţii în textul natural. Se notează cu k rangul unui cuvânt în list cu *f(k)* frecvenţa relativă a acestuia (numărul de apariţii raportat la *f(l)> f(2)>...>/(N_c).* (Altfel spus, *f(k)* este de tipul *p** din capito precedente). Legea lui Zipf afirmă că produsul dintre rang şi frecvenţa relativă e constant, [11] - [14]:

$$kf(k)=A$$

Se observă că membrul stâng al ecuației (3) corespunde realității fiind vorba de măsurători efectuate pe texte naturale în timp ce membrul drept corespunde modelului teoretic presupus.

Este ştiut din considerații privind alte limbi naturale că legea Zipf, apreciată ca foarte simplă şi foarte atractivă, funcționează cu aproximație pentru o plajă limitată de ranguri, anume nu prea mici şi nu prea mari. Astfel un prim pas al studiului nostru teoretic şi experimental a fost să reprezentăm grafic dependența rang - frecvență pe tot corpusul de care am dispus (corpusul mixt global, **#CMG).** Fig. 3 prezintă această dependență la scară logaritmică *(f(k) versus k).* La o primă vedere am putea spune că mărimea A din (3) este aproximativ constantă pentru un interval de ranguri *ke* I&$_{m}$i$_{n}$;&maxJ ^{un}*de* k_{m}[$_{n}$ >50. Am limitat studiul la acele ranguri pentru care numărul de apariții ale cuvintelor a fost mai mare decât 50 pentru a beneficia de rezultatele anterioare privind studiul de staționaritate prezentat în Cap. 2. Aceasta face ca rangul k_{m2ix} să depindă de corpusul analizat.

Legea lui Zipf este descrisă în numeroase referințe dintre care în limba română menționăm în special [13] şi [14]. Capitolul de față urmăreşte determinarea constantei legii atât pe corpusul de ansamblu, **#CMG,** cât şi pe diverse texte naturale (grupate după autori sau pe subdomenii ale limbii). Se analizează şi în ce măsură comportamentul real se abate de la cel teoretic.

3.1 Elemente teoretice

3.1.1 Determinarea parametrului legii Zipf prin minimizarea erorii pătratice

Presupunând valabilitatea legii Zipf pentru rangurile *ke* [&$_{m}$i$_{n}$;&max] ne-am propus să determinăm mărimea *A* din condiția de minimizare pe acest interval a următoarei funcții (suma pătratelor erorilor):

$$k\backslash xu\backslash x \qquad i2 \qquad (4)$$

$$k—hrr\backslash\backslash$$

Derivând funcția ***g(A)*** şi egalând cu 0 se obține valoarea mărimii A corespunzând minimului:

$$A = \left| \frac{V \qquad /(*)}{k=\hat{k}\ \bullet\ n\ u\ n} \right| \Big/ \left| \left(k \quad y \quad — \right)_{k=k\ -\ \&} \right| \qquad (5)$$

Valorile k_{m}[$_{n}$ şi £$_{max}$ sunt la dispoziția experimentatorului. Pentru o evaluare a acurateței cu care limba naturală verifică legea lui Zipf definim următoarele tipuri de erori:

• *s* , suma pătratelor erorilor pe intervalul *ke[k$_{m}$i$_{n}$;k$_{max}$]* şi forma ei normată, e_{n} :

$$£_{„}=£/\ \overset{Il\ ia/}{\underset{k=k\ min}{X}} \qquad (6)$$

$$k\text{-}k_{m}\backslash_{n}$$

M **eroarea relativă maximă pe intervalul de optimizare** k **e [fc$_{m}$j$_{n}$;^max]**

$$£M = {}^{£}r(*M) = \overset{max\ e}{\underset{k}{}}\ r\ W \qquad e_{i}(k) = \backslash f(k)\text{-} \qquad \mathbf{U\ J} \qquad (7)$$

3.1.2 Determinarea parametrului legii lui Zipf considerând cazul ideal

Dacă acceptăm legea lui Zipf ca fiind corectă pe întreg domeniul de ranguri *ke\l;N$_{c}$\,* atunci valoarea constantei A se determină prin raționamentul descris în [13], [14]:

$$c + \ln N_{c}$$

unde *c* este constanta lui Euler, egală cu 0,577215 şi N_{c}>50.

Observăm că mărimea A calculată cu relația (8) nu depinde decât de numărul Af$_{c}$ de cuvinte distincte din textul analizat. Prin urmare sunt de aşteptat unele diferențe între evaluările mărimii A pe baza datelor experimentale cu relația (5) şi cazul ideal, pur teoretic, relația (8).

3.1.3 Corolar al legii lui Zipf

Rezultatele experimentale cuprind şi verificarea unui corolar al legii Zipf care se referă la determinarea cotei părți, l_{s}/L_{c} , pe care o acoperă cele mai frecvente s cuvinte într-un text de lungime L_{c} , [13], [14].

$$\frac{l_{s}}{L_{c}} \qquad \frac{c + \ln s}{c + \backslash nN_{c}} \qquad (9$$

Relația (9) este valabilă pentru un număr de cuvinte $s > 50$.

Observăm că valoarea raportului l_{s}/L_{c} nu depinde de mărimea A. D aceea diferențele existente între diversele moduri de evaluare ale mărimii A nu v influența acest raport. în consecință ne aşteptăm la o bună verifica experimentală a acestui corolar.

3.2. Rezultate experimentale şi concluzii

Analiza experimentală a legii lui Zipf a început cu corpusul global **#CM** (vezi Fig. 3) şi a continuat pentru comparație cu o serie de texte naturale incluse acesta (prezentate în Introducere). Rezultatele experimentale sunt concentrate Tabelul 7. Pentru fiecare text analizat Tabelul 7 conține în coloanele 2 şi 3 numă total de cuvinte L_{c} şi numărul cuvintelor distincte 7V$_{c}$. în toate textele analiza

s-au investigat toate cuvintele cu număr de apariţii mai mare decât 50; acesta determină rangul $£_{max}$ corespunzător fiecărui text analizat (coloana 4). $&_{max}$ diferă de la text la text; $&_{min}$ este ales întotdeauna 51. Pentru acest interval de rangurile L^min^maxJ» s-a determinat cu relaţia (5) mărimea A cuprinsă în coloana 5. Coloanele 6 - 9 conţin rezultatele numerice calculate cu relaţiile (6) şi (7) unde mărimea A este cea din coloana 5 (determinată din textul natural respectiv). Coloana 9 conţine rangul k_M pentru care s-a obţinut eroarea relativă maximă e_M .

Ne-am pus problema şi dacă mărimea $A = 0{,}0909$ determinată pentru corpusul mixt global **#CMG,** ar putea fi acceptată drept referinţă pentru limba română. De aceea coloanele 10-13 conţin succesiv mărimile din relaţiile (6) şi (7) unde $A = 0{,}0909$ pentru toate textele naturale analizate. Eroarea relativă maximă e_M este însoţită de rangul corespunzător, k_M .

Tabel 7

Studiu experimental al legii lui Zipf în limba română scrisă

Text	L_c		^•max	A x10²	e x10⁶	x10²	x10²	k_M	£ x10⁶	x10²	£M x10²	ᵏM
1	2	3	4	5	6	7	8	9	10	11	12	13
#CMG	8806433	202403	14543	9,09	0,36	0,22	9,81	286	0,36	0,22	9,81	286
#CLG	6255235	162124	10299	9,60	0,30	0,17	13,93	10136	0,81	0,50	15,43	149
#1JCLG	3127618	116247	5568	9,58	0,26	0,14	10,53	136	0,72	0 44	16,41	136
#2JCLG	3127617	116860	5529	9,74	0,29	0,16	10,02	122	1,11	0,69	17,86	122
#1	226420	26943	466	9,81	0,37	0,22	9,07	68	1,28	0,88	16,81	173
#2	121177	18457	260	10,15	0,15	0,09	8,21	256	1,95	1,48	20,83	256
#3	88827	13768	190	10,07	0,20	0,14	10,17	186	1,60	1,33	22,07	186
#2+#3	210004	25036	484	9,97	0,52	0,29	18,71	478	1,89	1,29	30,18	478
#4	130743	18223	274	10,71	0,26	0,14	8,38	53	4,47	3,35	25,85	110
#5	75698	10351	187	11,56	0,42	0,22	10,86	183	9,22	7,71	40,92	183
#6	197889	23206	399	10,34	0,15	0,08	7,03	121	2,85	1,99	21,73	121
#4+#5+#6	404330	33555	849	10,53	0,20	0,10	8,90	103	4,03	2,62	26,08	103
#7	644794	49434	1195	10,03	0,35	0,18	10,80	477	2,04	1,30	21,49	77

In Fig. 3 sunt prezentate pentru corpusul mixt global două traiectorii, una experimentală (cu 'o') şi cea teoretică (cu '*') conform relaţiei (3) cu parametrul $A = 0{,}0909$ din coloana 5, Tabelul 7. Se observă o bună concordantă a celor două curbe pentru *e [W W I -

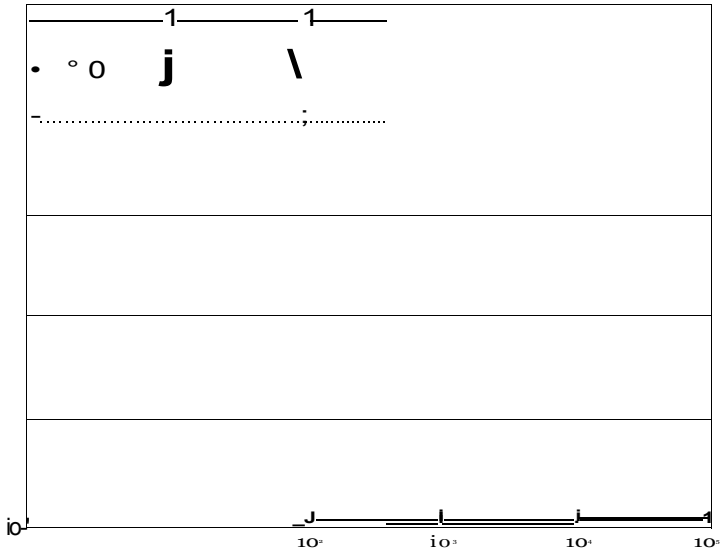


Figura 3. Dependenţa rang - frecvenţă relativă de apariţie a cuvintelor în corpusul mixt global **#CMG** (scară logaritmică *f(k) versus k).* Curba experimentală marcată cu 'o'; curba teoretică, relaţia (3) pentru $A = 0{,}0909$, marcată cu

în cazul ideal, pur teoretic, mărimea *A* poate fi determinată cu relaţia (8) pe baza coloanei 2 din Tabelul 7. Spre exemplu în corpusul global **#CMG,** unde au fost identificate $N_c = 202403$ cuvinte, $A = 0{,}0781$. în corpusul **#CLG,** pentru $N_c = 162124$ cuvinte distincte, aplicând relaţia (8) rezultă $A = 0{,}0795$.

Tabel 8

Valori teoretice, relaţia (9), şi experimentale ale raportului I_s/L_c în corpusul literar global #CLG

X	0,1%	0,05%	0,01%
s	104	189	911
I_s/L_c (experimental)	43,69%	49,64%	64,38%
I_s/L_c (teoretic)	41,53%	46,28%	58,78%

Tabelul 8 conţine date despre cota parte acoperită de cuvintele pentru care *f(k)>Ă* unde A = 0,1%; 0,05%;0,01%, în textul literar global **#CLG.** S-a folosit relaţia (9) unde N_c =162124, iar numărul de cuvinte s corespunzător pragului *X*

este conținut în linia 2 a Tabelului. Se observă o concordanță destul de bună între valorile teoretice şi cele experimentale.

Notă: Din cele 189 cuvinte din corpusul literar global #CLG care au frecvența relativă mai mare decât 0,05%, doar 156 au îndeplinit condiția deMoivre-Laplace şi au fost investigate cu control statistic apărând şi în Tabelul 5.

Ca o remarcă finală legea lui Zipf poate fi considerată ca valabilă şi pentru limba română pentru ranguri nu prea mici şi nu prea mari, fapt susținut de Fig. 3 şi datele din Tabelul 7.

4. Concluzii

Unul din principalele rezultate obținute în cadrul acestei lucrări este de a aduce probe noi privind staționaritatea limbii române scrise, de această dată pe baza structurii de cuvinte. (Ipoteza de staționaritate este inclusă în presupunerea generală conform căreia limbile naturale sunt lanțuri Markov multiple ergodice). Analiza staționarității s-a făcut prin extinderea unei metode dezvoltate de autori pentru studiul structurii statistice de m-grame (litere, digrame, trigrame, tetragrame). în consecință s-au putut obține probabilitățile cuvintelor cu intervale de încredere statistică 95% *"reprezentative".* Aceste intervale pe care le-am numit *"reprezentative"* au avut compatibilitate cu toate mulțimile de date *i.i.d.* obținute prin eşantionarea periodică a textului natural. Simultap au rezultat mulțimile de date *i.i.d. "reprezentative"* pentru cuvântul investigat şi textul natural analizat.

O altă contribuție constă în procedura de comparație matematică între texte naturale facilitată de intervalul *"reprezentativ"* pentru probabilitate şi de mulțimile de date *i.i.d. "reprezentative".* Comparațiile făcute între corpusuri organizate în aceeaşi manieră (literar *versus* literar sau mixt *versus* mixt) au întărit ideea de staționaritate a limbii şi au confirmat modelul matematic prezentat anterior prin intervale de încredere statistică 95% *"reprezentative"* pentru probabilitățile cuvintelor. Au apărut unele diferențe între domeniile literar şi științific.

Rezultatele experimentale dau un plus de semnificație frecvenței relative p*, mărime de care orice experimentator este interesat. Acest plus de semnificație este datorat faptului că în toate situațiile analizate de noi (cuvânt sau text natural) am putut obține o estimație a probabilității practic egală cu *p*,* iar intervalul de încredere statistică asociat acestei estimații a fost confirmat ca interval *"reprezentativ"* pentru probabilitate.

Lucrarea conține totodată şi confirmarea valabilității pentru limba română scrisă a legii lui Zipf (lege de tip rang - frecventă) şi a unui corolar al acesteia de interes lingvistic.

Autorii doresc să mulțumească D-lui dr. ing. Dan T corespondent al Academiei Române, pentru sprijinul ştiințific acor studiul limbii române scrise. Autorii menționează, de asemenea, primite din partea D-lui Prof. dr. ing. Alexandru Şerbănescu d Tehnică Militară.

Referințe bibliografice

[1] C E . Shannon, *"Prediction and Entropy ofPrinted English",* Bell Voi. 30, pp. 50-64, January, 1951.

[2] Adriana Vlad, A. Mitrea, *"Estimating condițional probabili statistical structure in printed Romanian",* in Recent Advance Languge Technology, D. Tufiş, P. Andersen eds., Ed. Acade 1997, (ISBN 973-27-0626-0), pp. 57-72; http://www.raca vlad.html.

[3] Adriana Vlad, A. Mitrea, M. Mitrea, D. Popa, *"Statistical me the natural language stationarity based on the first approxima Printed Romanian",* Proc. VEXTAL'99 (Venezia per il trattan della lingue), Ed. Unipress, (ISBN 88-8098-112-9), pp. 127- 1999, Venezia; http://byron.cgm.unive.it/events/papers/vlad.p

[4] . Adriana Vlad, A. Mitrea, M. Mitrea, *"Verifying Printed Ro Stationarity Based on the Digram Statistical Structure",* Pro Romanian Academy, Series A, Voi. 1, No. 2, pp. 129-139, 200

[5] Vlad Adriana, Mitrea A., Mitrea M., *"Two frequency-rank printed Romanian",* Procesamiento del Languaje Natural, (Sociedad Espanol para el Procesamiento del Language Nat 5948), pp. 153-160, Septiembre, 2000.

[6] Adriana Vlad, A. Mitrea, M. Mitrea, *"The trigram statistical s Romanian",* ROMJIST (Romanian Journal of Informatio Technology), Voi. 4, No. 3, pp. 353-372, 2001.

[7] Adriana Vlad, A. Mitrea, M. Mitrea, *"A Corpus - based Accurately Printed Romanian Obeys Some Universal Laws",* i Corpora: Corpus Linguistics and the Languages of the Wor Rayson, T. McEnery eds., Lincom-Europa Publishing Hous (ISBN 3-89586-872-8), pp. 153-165 .

[8] Adriana Vlad, A. Mitrea, M. Mitrea, *"Estimating tetragram pr multiple data samples from a natural text. Case study: p* Proc. IPMU 2002 - The 9th International Conference Processing and Management of Uncertainty in Knowledge - July 2002, Annecy, France, pp. 1285-1292.

[9] J. Devore, *Probability and Statistics for Engineering and the Sciences,* 2nd ed., Brooks/Cole Publishing Company, Monterey, California, 1987.

[10] Adriana Vlad, B. Badea, M. Mitrea, *Metode Statistice în Prelucrarea Informației. Compendiu şi Aplicații,* Ed. Metropol, Bucureşti, 1999, (ISBN 973-562-104-5).

[11] I. Kanter, D.A. Kessler, *"Markov Processes: Linguistics and Zipfs Law",* Physical Review Letters, Voi, 74, No. 22, pp. 4559-4562, May, 1995.

[12] R. Gunther, L. Levitin, B. Schapiro, P. Wagner, *"Zipfs Law and the Effectof Ranking on Probability Distributions",* Intl. J. of Theoretical Physics, Voi. 35, No.2, pp. 395-417, 1996.

[13] S. Marcus, Ed.Nicolau, S. Stați, *Introducere în lingvistica matematică,* Ed. Ştiințifică, Bucureşti, 1966.

[14] M. Dinu, *Personalitatea limbii române,* Ed. Cartea Românească, Bucureşti, 1996.

Fie / = *(a;b)* un interval în care presupunem că se află probabilitatea p a unui eveniment urmărit. Dispunem de o mulțime de date experimentale -[jq,x2,....*ArL date care satisfac modelul statistic *i.i.d..* Ne interesează dacă datele experimentale *[xi,x2,..-,xN]* confirmă ipoteza că probabilitatea p aparține intervalului / = *(a;b),* pentru un prag de semnificație statistică, *a,* ales.

Procedura de test este următoarea:

Se formulează cele două ipoteze statistice, ipoteza nulă *HQ* şi respectiv ipoteza alternativă *Hi:*

HQ: p aparține intervalului *(a;b); pe(a\b)*

H\: p este în afara intervalului *{a\b)\ p£(a\b).*

Se alege pragul de semnificație *a* (echivalent, nivelul de încredere statistică 1-a). Se calculează estimația *p = m/N,* unde cu *m* s-a notat numărul de succese ale evenimentului în mulțimea de date [*i,jt2,Verificăm dacă estimația *p* se află sau nu în zona de acceptare a datelor. Regiunea de acceptare a datelor este un interval (q;c2) care include *(a\b).* Intervalul (q^) se determină conform relației (10), [3]-[8]:

$$-yjlna\{\backslash -a)iN \text{ exd } \frac{(*-\langle\langle)^{*}}{2a(\backslash -a)IN} \wedge \cdot ? \frac{1}{\wedge 2nb\{\backslash -b)IN} \bullet \exp \frac{(}{2b(\backslash -b)/N} \frac{(x-b}{} \backslash dx = i-a \quad (10)$$

în relația (10) apar două funcții de densitate de probabilitate corespunzătoare legii normale: de medie *a* şi dispersie *a{l-a)/N* şi respectiv de medie *b* şi dispersie *b(l-b)/N.*

Ipoteza nulă *HQ va* fi acceptată dacă şi numai dacă estimația *p* aparține intervalului (q;c2j. în caz contrar, *p<£{ci;c2),* datele se resping ca fiind semnificative pentru pragul de semnificație *a* ales (se acceptă ipoteza *Hi).*

Ca în orice test statistic, pot să apară două tipuri de erori:

Eroarea de tipul (genul) I: Eroarea de a fi respinse date bune, adică să fie respinsă ipoteza H_0 când ea este corectă. Această situație apare atunci când estimația *p* nu satisface testul, adică p6(q;c2j, cu toate că probabilitatea adevărată p este în intervalul *[a\b).* Probabilitatea acestui tip de eroare este mai nnică decât *a.*

Eroarea de tipul (genul) II: Eroarea de a fi acceptate date false, adică să fie acceptată H_0 când ea este, de fapt, falsă. Această situație apare atunci când estimația *p* satisface testul, pe(q;c₂j, cu toate că probabilitatea adevărată p a evenimentului nu aparține intervalului *{a\b), pe{a;b).* Pentru *a* si *N* fixate probabilitatea acestui tip de eroare depinde de valoarea adevărată necunoscută *p* şi se calculează cu relația:

$$p7ip(l-p)/N \int_{2p(l-p)/N}^{exd} \backslash dx, \qquad P<£(a;b).$$

/3{p) este mare atunci când p este la stânga lui a (sau la dreapta lui *b),* dar foarte aproape de a (respectiv de £>). Practic, deranjante sunt situațiile în care *p<(l-6)>a* sau *p>(l + d)-b,* cu toate că testul este trecut, adică /?e(q;c₂J. Valoarea *5* este determinată (prestabilită) de utilizator, în funcție de cât de mult deranjează această situație.

în studiul nostru asupra staționarității limbii române acest test a fost absolut necesar, vezi Cap. 2. A trebuit să stabilim dacă probabilitatea p a unui anumit cuvânt este aceeaşi când dispunem de diverse mulțimi de date experimentale extrase dintr-un acelaşi text natural (unde mulțimile sunt compatibile cu modelul statistic *i.i.d.,* dar nu sunt independente între ele). Testul a fost folosit şi în comparații între texte naturale.

Anexa 2. Test de egalitate între două probabilități - T2

Dispunem de două mulțimi de date experimentale în model statistic *i.i.d.,* de volume *N[,* respectiv N_2. Notând cu *m* numărul de succese (apariții) ale unui eveniment în prima mulțime de date experimentale, estimația probabilității este *Pi=(m[fNi).* Similar, pentru a doua mulțime de date experimentale, estimația probabilității este $p_2 = (m_2/N_2)$. Urmărim să stabilim dacă cele două estimații *p* şi p_2 provin din aceeaşi probabilitate teoretică, respectiv $p\text{-}p_2$.

Procedura de test este următoarea:

Se formulează cele două ipoteze statistice, ipoteza nulă HQ şi respectiv ipoteza alternativă *Hy:*

H_0: cele două probabilități teoretice sunt egale $p\ = p_2$;

H\: cele două probabilități teoretice sunt diferite $p\ * p_2$.

Se alege pragul de semnificație statistică *a.*

Se construieşte o valoare de test z conform, [9], [10]:

$$z = - \frac{P\text{-}P2}{\sqrt{\frac{P\backslash(l-P\backslash)}{} , \frac{P2(\sim P2)}{N_2}}}$$

Valoarea z depinde de datele experimentale prin estimațiile p_x şi p_{2}. In condițiile în care ipoteza HQ este adevărată z provine dintr-o variabilă aleatoare a cărei iege de repartiție este practic legea normală standard.

întrucât $p\backslash$ şi p_2 sunt necunoscute, se consideră $p\backslash = p_2 = \frac{m\backslash + m_2}{N_{(}+N_2}$

în aceste condiții valoarea de test z devine:

$$z = \frac{\backslash + Z}{N\backslash N_2} \frac{m\backslash N_2\text{-}m_2N\backslash}{\sqrt{\{m\backslash + m_2)(Ni + N_2\text{-}m\backslash - m_2)}} \qquad (11)$$

Ipoteza nulă HQ va fi acceptată (se va considera că probabilitățile sunt egale, *p\= p₂)* dacă şi numai dacă $\backslash z\backslash \leq z_{/2}$ *(z₀/₂* corespunde pragului de semnificație statistică *a* ales; am folosit $z_{a/2} = 1,96$*).* în caz contrar se respinge ipoteza de egalitate a celor două probabilități pentru pragul de semnificație statistică *a* ales.

Această procedură de test a fost folosită când am comparat între ele diverse texte naturale.

Dezambiguizarea automată a cuvintelor din corpusuri paralele folosind echivalenții de traducere

Dan TUFIŞ
Institutul de Cercetări pentru Inteligența Artificială, Academia Română
Calea 13 Septembrie, nr. 13, 74311, sector 5, Bucureşti
tufiş@racai.ro

Rezumat

Corpusurile paralele constituie surse de cunoştințe extrem de valoroase, traducerea unui text reprezentând o succesiune de decizii lingvistice pe care traducătorul le ia în vederea asigurării unei transpuneri cât mai naturale şi mai fidele a semnificației din textul sursă în textul tradus. Explicitarea şi extragerea acestor cunoştințe prin metode algoritmice, formalizarea şi reutilizarea lor ulterioară constituie provocări ale inteligenței artificiale, subiecte de interes fierbinte în cercetarea actuală. Lucrarea prezintă o serie de contribuții în această direcție, prezentând mai întâi o metodă originală de identificare a echivalenților lexicali de traducere a cuvintelor dintr-un corpus paralel (extrăgând deci un dicționar multilingv) şi apoi o metodă extrem de promițătoare pentru identificare automată a diferitelor sensuri ale cuvintelor polisemantice.

1. Motivații

Evoluția ştiințifică şi tehnologică este o sursă permanentă de formare a noi termeni sau a noi sensuri specializate pentru cuvintele existente. în domeniul lexicografiei multilinguale, păstrarea în actualitate a dicționarelor bi- şi multilingve fără a apela la tehnologiile informatice, cu precădere cele din sfera ingineriei lingvistice, este aproape imposibilă. O serie de studii în domeniul traducerii automate au arătat că principalele probleme în *acceptabilitatea* traducerilor automate şi cu atât mai mult al celor implicând pre- sau post-editare umană, nu sunt legate de probleme de natură sintactică (topică, acorduri, structură frazală) ci ele se regăsesc cu precădere în sfera lexicală, mai precis a semanticii lexicale. Evaluarea sistemelor existente de prelucrare a limbajului natural şi mai ales a

celor de traducere automată (cu variantele ce presupun intervenția expertului uman) a condus la identificarea unor puncte sensibile, deficitare (pentru o interesantă trecere în revistă a problemelor privind evaluarea sistemelor de prelucrare a limbajului natural şi a sistemelor de traducere a se vedea http://www.isi.edu/natural-language/mteval/). De pildă, traducerea greşită a unui cuvânt sau a unei expresii într-o frază perfectă din punct de vedere sintactic este percepută de imensa majoritate a consumatorilor de traduceri, în special de natură ştiințifică, ca mult mai gravă decât un dezacord gramatical sau vreo altă abatere de la norma gramaticii. S-a invocat pe bună dreptate că dificultatea majoră a prelucrării automate a limbajului este rezolvarea ambiguităților lexicale, a omonimiilor şi a polisemiei ce apar în orice text (scris sau vorbit). Spre deosebire de oameni, care de multe ori nici nu conştientizează aceste fenomene (ele sunt "obturate" fie de contextul textului, fie de cunoştințele de "bun simț" ale fiecărei persoane), procesoarele artificiale de limbaj natural încearcă rezolvarea ambiguităților printr-un proces inteligent de alegere, dintr-un spațiu al tuturor soluțiilor posibile în raport cu o modelare a limbajului, a soluției care respectă cel mai bine restricțiile modelului. Raportarea la modelul limbajului este esențială întrucât dificultatea procesului de prelucrare este cu atât mai mare cu cât modelul este mai complex: spațiul de căutare a soluțiilor poate creşte exponențial, iar procesul decizional poate deveni nedeterminist sau de complexitate neoperațională.

Rezolvarea algoritmică eficientă a omografiei a cunoscut spectaculoase progrese în ultimii 10-15 ani, dar identificarea automată a sensului pe care îl are un anumit cuvânt polisemantic într-un context dat este încă o problemă nerezolvată satisfăcător şi, prin urmare, un subiect "fierbinte" de cercetare. Problema identificării sensului cu care este utilizat un cuvânt este vitală în traducerea automată, întrucât se cunoaşte faptul că de foarte multe ori un cuvânt polisemantic dintr-o limbă se traduce într-o altă limbă prin cuvinte diferite, în funcție de sensul considerat. Este interesant de remarcat că dacă un cuvânt polisemantic din limba sursă se traduce printr-un singur cuvânt polisemantic în limba țintă, sau altfel spus toate sensurile cuvântului de tradus se regăsesc în cuvântul reprezentând traducerea sa, necesitatea identificării sensului de utilizare al cuvântului sursă nu mai este obligatorie (cel puțin la nivelul fazei de transfer lexical) cu excepția situației în care. diferitele sensuri ale cuvântului țintă se realizează lingvistic prin structuri de subcategorizare distincte.

în această lucrare vom prezenta în prima parte o metodă de extragere automată a echivalenților de traducere şi vom descrie apoi procedura de discriminare a sensurilor cuvintelor din corpusuri paralele pe baza echivalenților de traducere.

2. Echivalenți de traducere

2.1. Noțiuni preliminare

O pereche de texte în două limbi diferite L_A şi L_B, astfel încât unul reprezintă traducerea celuilalt constituie cea ce se numeşte un *bitext* Un bitext suficient de mare constituie un corpus paralel. L_A şi L_B se numesc echivalenți de traducere. Noțiunea de echivalență de traducere se poate rafina ia niveluri subtextuale, de pildă la nivelul paragrafului, al propoziției sau chiar la nivel lexical, al cuvântului sau al expresiei. în continuare elementul de aliniere lexicală îl vom numi, generic, *atom lexical* sau simplu *atom.* Un bitext în care echivalenții de traducere sunt explicitați se numeşte un bitext *aliniat.* Cea mai mică unitate textuală la nivelul căreia se realizează alinierea defineşte *granularitatea* echivalenților de traducere. Echivalenții lexicali de traducere (obiectul nostru de interes în această lucrare) depind evident de bitextul din care sunt extraşi iar procesul de extragere a lor devine echivalent cu extragerea unui dicționar bilingv, specific unui anumit domeniu[1]. Extragerea unui dicționar de echivalenți de traducere dintr-un bitext este în fond un proces de explicitare a dicționarului mental folosit de translatorul (sau translatorii) textului original.

Presupoziția fundamentală în încercarea de a alinia corpusurile paralele este că *aceeaşi* semnificație este exprimată în două sau mai multe limbi. Definirea identității de înțeles între două sau mai multe reprezentări ale (presupus) aceluiaşi lucru este o binecunoscută problemă filozofică care rămâne deschisă chiar în domenii mult mai precise decât cel al limbii (de pildă în ingineria software). Prin urmare, noțiunea de echivalent de traducere este un concept vag, şi pentru operaționalizarea sa în domenii ca traducerea automată, terminologie, managementul multilingual al documentelor şi altele asemenea avem nevoie de o definiție precisă în termeni direct cuantificabili. Una dintre cele mai larg acceptate definiții a echivalenței de traducere este cea folosită în [1]: "the translation equivalence defines a (symmetric) relation that holds between two different language texts such that expressions appearing in corresponding parts of the two texts are reciprocal translations. These expressions are called *translation equivalents".*

Majoritatea abordărilor moderne în extragerea automată a echivalenților de traducere[2], sprijinite de forța de calcul din ce în ce mai mare a calculatoarelor, utilizează metode statistice şi pot fi clasificate în două mari categorii:

^*Posibilitatea de a genera automat dicționare bilingve în domenii specializate, coroborată cu performanțele tot mai bune ale programelor de clasificare automată a textelor, deschide noi perspective traducerii automate şi în general prelucrăriimultilinguale a textelor.*

In continuare, dacă nu vom specifica altminteri, prin "echivalenți de traducere" vom înțelege implicit "echivalenți lexicali de traducere".

- paradigma *"presupune şi testează"* [2], [3] etc, se bazează pe generarea unei mulțimi de potențiali echivalenți de traducere (spațiul ipotezelor) din care se selectează ulterior, pe baza unor teste de independență statistică, echivalenții de traducere. Selectarea fiecărui echivalent de traducere se face independent de echivalenții extraşi anterior (procesul poate fi considerat ca fiind unul de optimizare locală).
- ® paradigma *"modelului de limbă" [4],* [5], [6] etc. presupune construirea unui model statistic al bitextului, model ai cărui parametri se estimează prin metode de optimizare globală. în această abordare un candidat supus estimării nu mai este o pereche de atomi lexicali ci o mulțime de perechi, numită *asignare* [4].

Există susținători şi critici ai ambelor abordări şi o discuție a avantajelor şi dezavantajelor lor este prezentată în [6]. în esență,' paradigma *"presupune şi testează"* este mult mai eficientă din punct de vedere computațional deoarece presupune investigarea unui spațiu al soluțiilor proporțional cu N^2, unde N este maximul dintre numerele de articole lexicale distincte din cele două părți ale bitextului, dar echivalenții de traducere cu număr mic de apariții sunt de obicei pierduți. Paradigma *"modelului de limbă"* este extrem de costisitoare din punct de vedere computațional întrucât spațiul de căutare al soluțiilor este teoretic proporțional cu N!, în schimb având potențialitatea identificării corecte chiar şi a echivalenților de traducere cu o singură apariție în bitext *(hapax-legomena).* în [4] sunt prezentați o serie de algoritmi foarte eficienți, bazați pe o serie de ipoteze simplificatoare dar raționale, ce permit ignorarea unor mari regiuni din spațiul de căutare, regiuni în care este improbabil să existe soluții acceptabile.

Metoda descrisă aici poate fi încadrată în categoria abordărilor de tip *"presupune şi testează".* Algoritmul generează mai întâi o listă de candidați şi apoi succesiv, alege din această listă perechile cu cele mai mari scoruri de co-ocurență în regiuni corespondente ale bitextului. După cum se va vedea în continuare, acest algoritm nu are nevoie de un dicționar bilingv inițial, dar dacă acesta există, utilizarea sa poate spori substanțial viteza şi acuratețea prelucrării.

2.2. Ipoteza corespondenței lexicale 1:1

în general, un cuvânt dintr-un segment ce apare într-o parte a bitextului se traduce în segmentul corespunzător din cea de a doua parte a bitextului tot printr-un singur cuvânt. Dacă acest lucru s-ar întâmpla întotdeauna, problema alinierii lexicale a unui bitext ar fi substanțial mai simplă decât în realitate. Din păcate ipoteza *"cuvânt la cuvânt"* nu este adevărată în foarte multe cazuri, astfel încât adoptarea ei ca premisă de calcul nu pare foarte promițătoare. Dificultatea poate fi însă ocolită prin considerarea ipotezei conform căreia un articol lexical dintr-o limbă se traduce în cealaltă tot printr-un singur articol lexical. Un articol lexical f este reprezentat fie de un cuvânt, fie de o secvență de cuvinte (sintagmă, compus, l expresie). Această formulare, cunoscută sub numele de *"ipoteza corespondenței ^' lexicale 1:1",* adoptată ca premisă computațională, simplifică mult problema țintă a ' alinierii lexicale a unui bitext, dar introduce probleme noi şi anume definirea şi respectiv recunoaşterea automată a articolelor lexicale. Din fericire aceste probleme sunt reductibile la contexte monolingve şi au soluții simple şi foarte eficiente. Un program capabil să realizeze recunoaşterea automată a articolelor lexicale se numeşte *segmentator lexical.* Un segmentator lexical este în general independent de limbă, iar funcționarea sa este controlată prin resurse specifice (dicționare conținând cuvinte, secvențe de cuvinte sau expresii regulate definite peste un vocabular limitat). în [7] este discutată structura resurselor necesare segmentării lexicale a textelor în limba română cu ajutorul segmentatorului *MtSeg,* dezvoltat la Universitatea Aix-en-Provence în cadrul proiectului european "Multext".

Adoptarea *"ipotezei corespondenței lexicale 1:1"* reduce dramatic complexitatea problemei extragerii echivalenților lexicali, indiferent de paradigma ..... în care este abordată rezolvarea (a se vedea [7], [8] pentru detalii). Trebuie menționat însă că o segmentare lexicală perfectă (din punctul de vedere al utilității ei într-un context multilingv) este practic imposibilă din cauza incompletitudinii inerente a oricărui dicționar frazai. în [9], [8] se arată cum poate fi surmontată această incompletitudirie a resurselor necesare segmentării lexicale.

2.3. Etape de preprocesare

2.3.1 Alinierea frazată

înainte de extragerea propriu-zisă a echivalenților de traducere, un corpus paralel este supus unor prelucrări preliminare, de aducere a bitextului într-o formă normalizată. După ce fiecare parte a bitextului a fost supusă segmentării lexicale, urmează etapa de aliniere la nivelul propoziției a corpusului paralel. Pentru acest scop, am utilizat o variantă puțin modificată a algoritmului prezentat şi documentat în [10]. în [7] este descris procesul de aliniere la nivel de frază şi furnizate exemple şi statistici pentru diferite perechi de limbi prezente în corpusul paralel multilingv "1984", conținând traduceri, în şase limbi, ale romanului omonim al lui George Orwell. Acolo arătam că, în marea majoritate a cazurilor, traducerile din limba engleză s-au realizat în celelalte limbi păstrând corespondența de 1:1 la nivelul frazei[1], cu alte cuvinte, aproape întotdeauna o frază din textul englezesc a fost tradusă ca o singură frază în celelalte limbi reprezentate în corpusul paralel. Algoritmul de aliniere la nivelul frazei poate depista şi acele cazuri în care traducerea s-a realizat fără păstrarea corespondenței 1:1. Astfel, au fost cazuri în care două fraze sursă au fost traduse printr-o singură frază, sau invers, când o frază din limba engleză a fost tradusă prin 2

[1] *Noțiunea de frază este luată aici în sensul ei larg, al unei propoziții sau fraze (enunț terminat cu un semn de punctuație din categoria celor finale: punct, punct şi virgulă, două puncte, semnul exclamării, semnul întrebării, trei puncte).*

sau chiar 3 fraze în celelalte limbi. în cele ce urmează, indiferent de tipul de aliniere (1:1, 2:1, 1:2 etc.) vom numi porțiunile aliniate la nivelul frazai, *unități de traducere* (UT).

Rațiunea acestei etape de prelucrare constă în intuiția comună că elementele lexicale aflate în relație de echivalență de traducere se regăsesc în frazele ce se constituie în unități de traducere. Pe de altă parte, procesul de aliniere la nivelul frazei este mult mai simplu, pentru că în general indiferent de perechile de limbi considerate într-un bitext ordinea frazelor dintr-o limbă este păstrată în cealaltă limbă. Această ipoteză, operaționalizată de un algoritm de optimizare dinamică de genul celui descris în [10], permite printre altele şi identificarea porțiunilor netraduse într-una din limbi (alinieri de tipul N:0 sau 0:M).

O altă ipoteză simplificatoare pentru procesul identificării echivalenților lexicali de traducere se bazează pe observația că în marea majoritate a traducerilor, categoriile gramaticale din limba sursă se conservă în limba țintă [1]. Cu alte cuvinte, un verb se traduce de obicei printr-un verb, un substantiv printr-un substantiv ş.a.m.d. Melamed a numit o astfel de pereche de traducere, pereche de tip V, distingând-o de perechile de tip P, în care atomii lexicali în cele două limbi au categorii gramaticale diferite. Melamed, distinge şi o a treia categorie de perechi de traducere, tipul I, perechile de traducere incomplete, rezultate ca urmare a unei segmentări lexicale parțiale şi a utilizării "ipotezei de aliniere lexicală 1:1". Considerațiile lui Melamed referitoare la distribuția celor trei tipuri de traduceri lexicale sunt foarte bine confirmate de experimentele noastre, în ciuda faptului că textul nostru este un text literar în timp ce textul său este un text politic (dezbaterile din Parlamentul Canadian) conținând traduceri literale, mult mai puțin afectate de personalitatea literară a traducătorului. Ceea ce este demn de remarcat este că perechile de tip P nu conțin categorii gramaticale arbitrare, şi că de la o pereche de limbi la alta, se pot identifica regularități în alternanța categoriilor gramaticale la traducere (de ex. participiu-adjectiv, gerunziu-substantiv, gerunziu-adjectiv). Astfel de regularități pot fi abstractizate prin expresii regulate, efectul net fiind ca multe din perechile de tip P pot fi asimilate (algoritmic) perechilor de tip V. Prin urmare, necesitatea identificării rapide şi precise a categoriei gramaticale (şi eventual al altor trăsături morfologice sau lexicale) pentru atomii lexicali dintr-un bitext impune o altă prelucrare preliminară, respectiv etichetarea morfo-lexicală, prelucrare pe care o prezentăm în secțiunea următoare.

2.3.2 *Etichetatea morfo-lexicală şi lematizarea*

Etichetarea morfo-lexicală este procesul prin care fiecărui articol lexical dintr-un text arbitrar i se atribuie un cod morfo-lexical unic dintr-o mulțime specifică articolului lexical respectiv, numită clasa sa de ambiguitate. Codul'morfo-lexical reprezintă o reprezentare compactă, şi de obicei standardizată, a proprietăților

morfologice şi lexicale ce caracterizează apariția unui atom lexical într-un text. Clasa de ambiguitate a unui atom lexical reprezintă mulțimea tuturor interpretărilor posibile în orice context legal al atomului respectiv. De exemplu cuvântul *"urâțr* are cel puțin 8 interpretări posibile, putând fi substantiv, adjectiv sau verb. Lema sa poate fi una dintre "urât" (substantiv sau adjectiv), "a urâți" sau "a urî" (verb).

urâţi	urâţi	Vmnp	(inf.: A *urâți* înseamnă a face să devină urât)
urâţi	urâţi	Vmis3s	(ind., perf simplu, sing., pers. 3: El *urâți* totul în viaţa ei)
urâţi	urâţi	Vmm-2s	(imp., sing: Prietene, nu *urâți* singurul lucru frumos din viaţa lui!)
urâţi	urî	Vmip2p	(ind., prez., p1, pers. 2: De pomană îi *urâți* pe ei, ceilalţi sunt de vină)
urâţi	urî	Vmsp2p	(subj., prez., pl., pers. 2: Voi ar trebui să *urâți* tot ce e împotriva vieţii)
urâţi	urî	Vmm-2p	(imp., pl.: Nu-i *urâți* pe apărătorii planetei!)
urâţi	urât	Afpmp-n	(adj., mase. pl., neart. : Doi câini *urâți* şi răi păzeau intrarea.)
urâţi	urât	Ncmp-n	(subs. corn., mase. pl., neart.: Nişte *urâți* m-au băgat în sperieţi.)

Aşadar, clasa de ambiguitate a cuvântului "i/râff este mulțimea (Vmnp, Vmis3s, Vmm-2s, Vmm-2p, Vmip2p, Vmsp2p, Afpmp-n, Ncmp-n), iar etichetarea morfo-lexicală a acestui cuvânt înseamnă a alege, în funcție de contextul apariției sale, unul şi numai unul dintre cele 8 coduri reprezentând interpetarea contextuală a cuvântului. în cercetările anterioare am dezvoltat o metodă statistică de etichetare morfo-lexicală [11], numită etichetarea cu două niveluri şi modele de limbă combinate (TT-CLAM: tiered-tagging with combined language models), bazată pe programul TnT al lui Thorsten Brants [12] de prelucrare a modelelor markov cu legături ascunse de ordin 2 (3-gram HMM), program ce poate fi descărcat de la adresa www.coli.uni-sb.de/~thorsten/tnt/. Abordarea TT-CLAM a arătat că texte arbitrare în limba română pot fi etichetate morfo-lexical în mod corect în peste 98.5% din cazuri şi că atunci când de interes este numai categoria gramaticală, procentul de etichetare corectă depăşeşte 99.5%. Metoda TT-CLAM s-a dovedit independentă de limbă, rezultate mai bune decât în alte abordări fiind raportate în literatura de specialitate pentru limbi foarte diferite de limba română: limba maghiară [13], [15] limba germană [16], [17].

Lematizarea este procesul prin care o formă flexionară a unui articol lexical (cuvânt sau expresie) este redusă la forma normală de dicționar. Lematizarea se poate realiza fie printr-un proces de analiză morfologică fie prin căutarea într-o bază de date lexicale, conținând cuvinte în formă flexionară însoțite de analiza lor morfologică şi de forma lernă. Lematizarea se realizează în acest caz prin

identificarea în baza de date a lemei pentru care forma flexionară şi analiza morfo-lexicală sunt identice cu cele din textul de lematizat, care desigur a fost în prealabil etichetat. Pentru limba română, noi am experimentat cu ambele metode şi datorită vitezei mult superioare, am optat pentru varianta a doua.

în figura de mai jos este exemplificat rezultatul prelucrărilor preliminare discutate în această secțiune (segmentare lexicală, aliniere frazală, etichetare morfo-lexicală şi lematizare) pentru începutul bitextului Englez-Român din corpusul multilingv "1984". Prima linie arată că în limba română, fraza cu identificatorul Oro.1.2.2.1, reprezintă traducerea a două fraze din textul englezesc, respectiv a celor cu identificatorii Oen.1.1.1.1 şi Oen.1.1.1.2 (avem deci o aliniere de tip 1:2). Liniile următoare, specifică pentru fiecare articol lexical din fiecare limbă tipul său (TOK, LSPLIT, DATE, ABR etc), forma ocurență, lerna, codul morfo-lexical şi categoria gramaticală (ultimele trei separate prin caracterul T).

<linktargets="Oro.1.2.2.1; Oen.1.1.1.1 Oen.1.1.1.2">

(<S FROM="Oro.1.2.2.1">

LSPLIT	într-	întru\Spsay\S	TOK	It	it\Pp3ns\P
TOK	o	un\Tifsr\T	TOK	was	be\Vmis3s\AUX
TOK	zi	zi\Ncfsrn\N	TOK	a	a\Di\D
TOK	senină	senin\Afpfsrn\A	TOK	bright	bright\AAA

</S>

(<S FROM="Oen.1.1.1.1">

</S>

<S FROM="Oen.1.1.1.2">

</S>

Figura **1:** Bitext preprocesat pentru extracția echivalenților lexicali de traducere

O descriere a principiilor de codificare morfo-lexicală, în conformitate cu recomandările EAGLES poate fi găsită în [18]. Codificarea specifică pentru limba română, conformă cu standardul respectiv este pe larg descrisă în [19].

2.4. Un prim algoritm de extragere automată a echivalenților lexicali de traducere

Există, aşa cum am văzut mai sus, mai multe ipoteze simplificatoare care permit ținerea sub control a complexității problemei extragerii automate a echivalenților de traducere. Nici una dintre aceste ipoteze nu este satisfăcută întotdeauna, dar situațiile în care ele nu sunt adevărate sunt suficient de rare astfel încât adoptarea lor nu alterează semnificativ valoarea rezultatelor. Trebuie subliniat faptul că ipotezele simplificatoare folosite de noi, disc rezumate în continuare, în general nu afectează precizia dicționarelor bilingve extrase ci completitudinea lor. Altfel spus, o corecte (echivalenți de traducere reali), deşi prezente în bitext, po Precizia şi completitudinea (în limba engleză aceşti termeni sunt se definesc în mod standard astfel:

PREC=(număr de echivalenți corect extraşi)/(număr tot extraşi)

COMP=(număr de echivalenți corect extraşi)/(număr tot existenți în bitext)

Mai trebuie precizat şi faptul că ipotezele simplificatoare e nu împiedică recuperarea ulterioară a echivalenților negăsiți din acestor ipoteze de lucru. în [9] sunt discutate metode de re echivalenți de traducere ce nu respectă ipoteza "echivalenței lexic

- ipoteza "echivalenței lexicale 1:1"; ea stă la baza maj cunoscute: [20], [21], [6], [22], [23], [1] etc. Aşa cu devreme, un articol lexical identificat corespunzător d lexical adecvat diminuează considerabil efectul cor ipoteze;
- un articol lexical polisemantic ce apare de mai mu unitate de traducere este folosit cu acelaşi presupoziție este explicit utilizată de [1] şi implicit d amintiți mai sus;
- un articol lexical dintr-o parte a unității de traducere unui articol lexical în cealaltă parte a UT doar dacă au categorii gramaticale compatibile; în majo compatibilitatea categoriilor gramaticale se reduce cum am specificat anterior, este posibil să corespondențe compatibile (de pildă, verbele la part din limba engleză sunt destul de frecvent traduse în adjective sau substantive, şi reciproc).
- Deşi ordinea cuvintelor nu este un invariant al tradu nici arbitrară; când două sau mai multe perechi de candidează la statutul de echivalenți de traducere, evaluare nu permit departajarea lor, atunci este pr conținând articolele cele mai apropiate în pozițiile lor euristică este, de asemenea, folosită de [23].

Pe baza bitextului preprocesat aşa cum s-a prezen precedentă, primul pas al algoritmului este de a delimita spaț soluțiilor. Acest lucru se realizează prin construcția unei liste a t

posibili (în conformitate cu ipotezele de lucru amintite mai sus). Această listă, pe care o notăm cu TECL (Translation Equivalence Candidates List) conține la rândul ei o mulțime de sub-liste (câte una pentru fiecare categorie gramaticală luată în considerare). Fiecare sublistă conține perechi de forma <token$_s$ token$_T$> unde *tokeris* şi *token$_T$* sunt articole lexicale de categorii gramaticale compatibile şi care au apărut în părțile corespunzătoare ale aceleiaşi unități de traducere. Fie TU^j cea de a j^a unitate de traducere (translation unit). Prin colectarea tuturor articolelor lexicale aparținând aceleiaşi categorii gramaticale POS$_k$ (păstrând ordinea lor relativă şi eliminând duplicatele) se construiesc pentru fiecare TU^j mulțimile ordonate L^{Sj}$_P$oskŞi L^{Tj}$_P$osk- Pentru fiecare POS$_{if}$ie TUVosi produsul cartezian L^{Sj}$_{PO}$si ®L^{Tj}posi- Atunci, definim lista de corespondențe în unitatea de traducere TU^j ca fiind CTU^j (correspondences in the jth translation unit):

$$CTO = \prod_{i=1}^{\text{no.of.pos}} |J^{TU}\ \text{Posi}$$

Cu aceste notații, şi presupunând că bitextul de intrare conține *n* unități de aliniere, atunci TECL se defineşte astfel:

$$TECL = [JCTU^j \quad H$$

TECL conține desigur foarte mult "zgomot" şi cele mai multe perechi candidate (TEC=Translation Equivalence Candidate) sunt extrem de improbabile. Pentru a elimina cât mai multe din perechile TEC improbabile, TECL este filtrată pe baza unor funcții scor ce supun fiecare TEC la o analiză a ipotezei statistice de independență a asocierii articolelor lexicale. Pentru a prezenta funcțiile scor pe care le-am utilizat în experimentele noastre, vom mai defini o serie de notații:

- TEC = <T_s T_T> G TECL, un potențial echivalent de traducere definit ca perechea formată din articolul lexical sursă T_s şi posibila sa traducere T_T în limba țintă;
- nu = numărul de ocurențe ale <T_s T_T> din TECL;
- n_{12} = numărul de perechi <T_s -iT_T> din TECL în care T_s a fost asociat cu un articol lexical diferit de T_T;
- n_2i = numărul de perechi <-.T_s T_T> din TECL în care T_T a fost asociat cu un articol lexical diferit de T_s;
- $n_2$2 = numărul de perechi <-*J*$_s$ -»T_T> din TECL ce nu conțin nici pe T_s şi nici pe T_T;
- n_r = numărul de perechi <T_s *> din TECL în care apare T_s indiferent cu cine este asociat;
- n«i = numărul de perechi <* T_T> din TECL în care apare T_T indiferent cu cine este asociat;

- n_2* = numărul de perechi <-.T_s *> din TECL în care T_s nu apare;
- n^*_2 = numărul de perechi <* -,T_T> din TECL în care T_T nu apare;
- n« = numărul de perechi <* *> din TECL;

Tabela de contingență din figura de mai jos ilustrează aceste notații:

	T_T	-.T_T	
T_s	nu	n_{12}	nr
-T_s	n_{21}	n_{22}	n2*
	n_M	n^*_2	n~

$n_r = nn + n_{12}$, $n_2^* = n_{21} + n_{22}$

$n^*i^n < n + n_2i$, $n^*_2 = ni_2 + n_{22}$

$$n\sim = \sum_{;=1}^{2}\sum_{i=1}^{2} ny$$

Figura **2:** Tabela de contingență pentru un potențial echivalent de traducere <T_s T_T>

Pentru ordonarea potențialilor echivalenți de traducere în vederea filtrării (eliminarea candidaților cei mai puțin plauzibili) am realizat experimente folosind 4 funcții de calcul al scorului de echivalență: MI *(informația mutuală),* DICE, LL (log likelihood), and %^{2} (chi-pătrat). Folosind notațiile de mai sus, aceste funcții-scor se definesc în felul următor:

(1) $$MI(TT, T_s) = \log \frac{n^{***}n_n}{rij^{*}\ {}^{*}n^{*}j}$$

(2) $$DICE(T_T, T_s) = \frac{2n_{MI}}{\cdot}$$

(3) $$LL(TT, TS) = 2 * \sum_{y=i}^{2}\sum_{i=i}^{2} "«' log - ^ r - ?'$$

(4) $$X^2(TT, T_s) = \frac{IV * n^*j\ ^}{n^{*}\ ^\ ^}$$

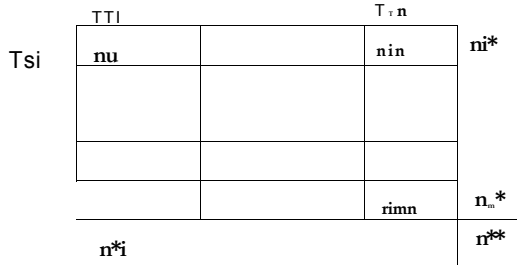
O formulă mai simplă de calcul pentru $\%^2(T_T, T_s)$ este următoarea:

(4') $$\%^2(T\ T) = {}^{n}**({}^{nn}\ {}^{1122}\ {}^{\pi i2}*{}^{\pi 2i})$$

Filtrarea potențialilor echivalenți de traducere se face în raport cu un prag numeric impus scorului calculat cu una dintre funcțiile de mai sus. Toate perechile ce obțin un scor mai mare decât pragul ales sunt considerate plauzibile şi vor fi supuse unor prelucrări suplimentare iar celelalte sunt eliminate. Orice metodă de filtrare statistică va elimina mulți echivalenți falşi de traducere, dar pe lângă aceştia şi un număr de perechi corecte. Alegerea pragului de scor s-a făcut având ca obiectiv minimizarea numărului de perechi corecte dar eliminate în mod greşit şi a numărului de perechi incorecte acceptate ca urmare a scorului superior pragului de selecție. După mai multe experimente, cele mai bune rezultate s-au obținut folosind funcția de scor LL cu limita pragului de acceptanță egală cu 9.

într-o primă variantă, algoritmul nostru de extragere a echivalenților de traducere, având unele asemănări cu algoritmul iterativ prezentat în [23], implementa o strategie de selecție indiferentă la locul şi poziția în corpus a articolelor lexicale apărând în perechea TEC analizată la un anumit moment. O diferență majoră față de algoritmul descris în [23] este că în programul nostru calculul diferitelor probabilități (mai exact al estimaților de probabilitate) şi al scorurilor (testul t) devine nenecesar, conducând la o viteză de prelucrare cu cel puțin un ordin de mărime mai mare. Pornind de la lista filtrată a potențialilor echivalenți de traducere, algoritmul selectează în mod iterativ cei mai plauzibili candidați (vezi mai jos) şi apoi îi şterge din lista inițială. Algoritmul se opreşte după un număr prestabilit de iterații sau mai devreme în cazul în care lista candidaților s-a golit sau dacă nici un candidat nu mai îndeplineşte condiția de selecție.

în iterația *k* a algoritmului se construieşte o matrice de contingență (TBLk) pentru fiecare categorie gramaticală (POS) având dimensiunile $S_m * T_n$ unde S_m şi T_n reprezintă numărul de articole lexicale din limba sursă respectiv țintă care mai există în lista de candidați la pasul *k* (Figura 4). Liniile şi coloanele tabelei sunt indexate cu articolele lexicale (având aceeaşi categorie gramaticală) din limba sursă respectiv limba țintă. Fiecare celulă (i,j) a matricii reprezintă numărul de ocurențe in lista de candidați a perechii <T_si, T_{Tj}>.



$$ny = occ(T_s i, T_{Ti});\ i> = \sum_{j=1}^{n};\ n*f= \pounds n_{sj};\ n- = \pounds\pounds(x^{n} ij)-$$

Figura **4:** Matricea de contingență la pasul *k*

Condiția de selecție la pasul k a mulțimii de echivalenți de traducere este exprimată de relația (5):

(5) $$TP^{k}=\{cT_{si}T_{15}>|V_{Pf}q(n_{u}\ >n_{iq})A(_{nij}\ >n_{Pi})\}$$

Condiția de mai sus constituie esența algoritmului iterativ (numit în [14] algoritmul BASE) şi ea spune că pentru a selecta perechea <T_{Si}, T_{Tj}> drept echivalent de traducere, numărul de asocieri ale lui T_{Si} cu T_{Tj} trebuie să fie mai mare sau cel puțin egal decât numărul de asocieri ale lui T_{Si} cu orice alt T_{Tp} (p*j) şi simultan numărul de asocieri ale lui T_{Tj} cu T_s**j** trebuie să fie mai mare sau cel puțin egal decât numărul de asocieri ale lui T_{Tj} cu orice alt T_{Sq} (q*i). Toate perechile selectate în TP^k sunt eliminate din lista de candidați (ceea ce în matricea de contingență pentru pasul k+1 implică punerea pe 0 a contoarelor de ocurență pentru perechile selectate anterior). Dacă T_s**j** este tradus în mai multe moduri (fie pentru că are sensuri ce se lexicalizează diferit în limba țintă, fie pentru că în limba țintă se folosesc diferiți sinonimi pentru T_r**j)** restul traducerilor sale va fi extras în iterațiile următoare. Algoritmul discutat este schițat în figura 5:

```
procedure BASE(bitext, step; dictionary) is:
   k=1;
   TP(0)={};
   TECL(k)=build-cand(bitext);
   for each POS in TECL do
      loop
         TECL(k)=update(TP(k-1), TECL(k))
```

```
        TBL(k)=build_TEC_table(TECL(k));
        TP(k)= select(TBL(k));  ## relaţia (5) ##
        add(dictionary, TP(k));
        k=k+1;
    until {(TECL(k-1) is empty)or(TP(k-1) is empty)or(k >
step)}
  endfor
  return dictionary
end
```

Figura 5: Algoritmul iterativ de extragere a echivalenților de traducere

2.5. Un algoritm îmbunătățit de extragere automată a echivalenților lexicali de traducere

Una dintre principalele deficienţe ale algoritmului BASE este vulnerabilitatea la ceea ce [1] numeşte *asociaţii indirecte.* Dacă <T_{Sj}, T_{Tj}> are un scor de coocurenţă ridicat iar T_{Tj} apare (dintr-un motiv sau altul) de mai multe ori împreună cu T_{Tk}, s-ar putea ca şi perechea <T_{Si}-, T_{Tk}> să primească un scor ridicat. Deşi, aşa cum observa şi Melamed, în general, asociaţiile indirecte au un scor mai mic decât cele directe (corecte), ele pot obţine totuşi scoruri mai mari decât multe alte perechi corecte ce n-au legătură cu T_{Si}- iar acest lucru nu numai că generează echivalenţi de traducere greşiţi, dar va elimina din competiţie şi echivalenţi corecţi. Prin urmare asociaţiile indirecte afectează atât precizia cât şi completitudinea procesului. Pentru a slăbi această sensibilitate în implementarea algoritmului BASE a fost nevoie de stabilirea unei limite inferioare de ocurenţă pentru fiecare articol lexical luat în consideraţie. Această limită, conduce inevitabil la eliminarea din spaţiul de căutare a soluţiilor a mai mult de 50% dintre echivalenţii de traducere[1]. Deficienţa algoritmului BASE se explică prin faptul că scorurile de coocurenţă sunt calculate în mod global fără a verifica dacă atomii lexicali ai unei perechi evaluate sunt sau nu prezenţi în unităţile de traducere prelucrate.

Pentru diminuarea influenţei asociaţiilor indirecte fără a mai impune un prag de ocurenţă, algoritmul BASE a fost modificat astfel încât ierarhizarea şi alegerea celor mai probabili echivalenţi de traducere se realizează în contextul local al fiecărei unităţi de traducere (deşi scorurile lor se calculează tot la nivelul întregului bitext). Cu această modificare, noul algoritm (BETA) se apropie de algoritmul "competitive linking" al lui Melamed [1]. Candidaţii proveniţi din unitatea de traducere curentă sunt analizaţi prin prisma scorului lor de coocurenţă şi cel cu scorul cel mai mare este selectat. în baza ipotezei corespondenţei lexicale 1:1,

[1] *Pierderea unui aşa mare număr de echivalenţi de traducere desigur nu surprinde întrucât*

dintre candidaţii rămaşi sunt eliminaţi toţi aceia care conţin un lexicale ale perechii câştigătoare. Dintre candidaţii care rămân filtrare, se alege din nou cel cu scorul cel mai bun şi iar se e conţinând unul dintre articolele lexicale din perechea selectată. Pr până când nici un echivalent de traducere nu mai poate fi extras traducere curentă, caz în care algoritmul trece la prelucrarea urm traducere.

Eliminarea pragului de ocurenţă a îmbunătăţit substanţial şi calitatea dicţionarelor de traducere (o detaliată comparaţie a pe analiză cantitativă şi calitativă a dicţionarelor extrase automat est dar a ridicat problema decelării între candidaţii cu una sau dou care scorul de coocurenţă este statistic nesemnificativ. în ac frecvenţei a fost înlocuit cu o combinaţie între un scor de similarit un scor de proximitate relativă. Funcţia de similaritate ortografic COGN(T_s, T_T), este o variantă a funcţiei XXDICE descrisă în [21 este un şir de *m* caractere $o t i a_2 . . . a_m$ and T_T un şir de *n* caractere construiesc două noi şiruri T_s şi T'_T prin inserarea în T_s şi T_T a u de caractere speciale astfel încât în final şirurile T'_s şi T'_T au ac (max (m, n)<p<m+n) şi un număr maxim de caractere poziţional ic caracter din T'_s şi Pi un caracter din T'_T care se potrivesc corespondenţă. Fie 8(ai) numărul de caractere speciale conse imediat caracterul oţj şi 8(**PJ**) numărul de caractere speciale c caracterul Pi . Fie *q* numărul de caractere care se potrivesc în ce aceste notaţii, măsura de similaritate COGN(T_s, T_T) se defineşte a

(6) $$COGN(T_s,T_T) = \frac{m+n}{} \quad if$$

it

Limita de relevanţă a scorului de similaritate a fost empir Această valoare este dependentă într-o oarecare măsură de p considerată în procesul de extragere a echivalenţilor de traduce efectivă a testului de similaritate include şi o serie de norma testate (eliminarea unor afixe, reducerea consoanelor duble, ig create de diacritice etc.) normalizări care depind de morfologi parte.

Cel de al doilea criteriu de evaluare a plauzabilităţii u scorul de proximitate, DIST(T_s, T_T) definit după cum urmează:

Dacă((<T_s, T_T>e L^{Sj}_{posk} ®L^{Tj}_{posk})&(T_s este al n-lea elem este al m-lea element în L^{Tj}_{posk})) atunci DIST(T_s, T_T)=|n-m|

Filtrul COGN(T_s, T_T) este mult mai semnificativ din punct de vedere lingvistic' decât DIST(T_s, T_T), astfel încât scorul de similaritate are precedență asupra celui de proximitate. Funcția DIST(T_s, T_T) este invocată doar atunci când COGN(T_s, T_T)=0 (deci când atomii lexicali nu prezintă similaritate ortografică) şi perechea <T_s, T_T> nu reprezintă o pereche singulară în corpus (hapax-legomena), sau când mai multe perechi candidate au obținut acelaşi scor de similaritate.

Algoritmul BETA este schițat mai jos:

```
procedure BETA(bitext;dictionary) is:
   dictionary={};
   TECL(k)=build-cand(bitext);
   for each POS in TECL do
      for each TXJ\os in TECL do
         finish=false;
         loop
            best__cand=get_the_highest_scored_pairs(TU^X_pos);
            conflicting_cand=select___conflicts(best_cand);
            non_conflicting__cand = best_cand¥conflicting_cand;
            best_cand=conflicting_cand;
            if cardinal(best_cand)=0 then finish=true;
            else
               if cardinal(best_cand)>1 then
               best__card=filtered(best__cand);
               endif;
            best__pairs = non_conflicting__cand + best_cand
            add(dictionary,best_pairs);
            TU
              POS
            endif;
            remove_pairs_containing_tokens_in_best_pairs(TU_pos);
         until { (TU^1_POS={} lor(finish=true)}
      endfor
   endfor
   return  dictionary
end

procedure filtered(best_cand) is:
   result = get_best__COGN_j3Core (best_cand);
   if (cardinal(result)=0)&(non-hapax(best_cand))then
      result = get_best_DIST__score (best_cand);
   else if cardinal(result)>1
            SUlt = get best DI
      endff       -   - ST_score(best_cand);
   endif
```

'^£S3S55S8S5£3S£

```
   return result;
end
```

Din corpusul paralel multilingv "1984" am extras 6 bitexte conținând textu în limba engleză şi traducerea în una din cele 6 limbi amintite. Fiecare bitext a fos prelucrat conform celor prezentate în acest capitol şi au fost extrase 6 dicționar bilingve, din care s-a obținut şi un dicționar multilingv în 7 limbi (cele 6 plu engleza). în [8] este furnizată o analiză contrastivă cu alte sisteme de acest tip vitezei de prelucrare. Timpul mediu de extragere a unui dicționar bilingv di corpusul paralel multilingv "1984" (circa 110.000 de cuvinte în fiecare limbă) este minute. Eşantioane ale acestor dicționare pot fi consultate la adresa http.V/www.racai.roMufis/BilingualLexicons/AutomaticallvExtractedBiiingual Lexicons.html.

3. Dezambiguizarea sensurilor lexicale folosind echivalențele de traducere

3.1. Ambiguitatea limbajului natural

Este binecunoscut faptul că una dintre cele mai dificile probleme î prelucrarea automată a limbajului natural este ambiguitatea sa inerent Ambiguitatea se manifestă la toate nivelurile tradiționale ale analizei de limba nivelul fonetic şi/sau lexical, sintactic, semantic sau discursiv. Ambiguitate prezentă pe fiecare nivel generează exploziv ambiguități pe nivelurile următoar De pildă, omofonia sau omografia prezentă pe primul palier, la nivelul unuia sau mai multor cuvinte va produce secvențe lexicale diferite (combinația tutur interpretărilor posibile la acest palier) pentru intrarea fazei de analiză sintactic Fiecare secvență poate conduce, din pricina unor ambiguități de natură structural la interpretări sintactice multiple, după cum o serie de secvențe lexicale vor putea abandonate pe motivul contrazicerii unor restricții postulate de modelul sintactic limbii prelucrate. Fiecare dintre interpretările sintactice posibile poate la rândul să conducă la multiple interpretări semantice, în virtutea multiplelor sensuri pe ca le poate avea fiecare element frazai al unei analize sintactice. Desigu interpretarea semantică poate elimina unele structuri sintactice generate în faz anterioară pe baza încălcării unor restricții semantice (valabile în orice univers d discurs sau specifice unor domenii discursive de interes). în sfârşit, în analiza d discurs, în care contextul interpretativ transcende limita propoziției, ambiguități rămase se presupun a putea fi rezolvate prin utilizarea restricțiilor pragmatic motivate fie de principii generale ale dialogului (coeziune, coerență), fie de natu bine precizată a unui univers de discurs (modelată prin cunoştințe extra-lingvistic despre entitățile universului de discurs). De pildă, în [24] rezolvarea anaforelo proces tipic analizei de discurs, este modelată în termenii identificării căilor d

accesibilitate a entităților menționate în discurs ("vene ale discursului"), care la rândul lor sunt formal definite pe baza principiilor generale ale coeziunii şi coerenței unui text.

Rezultă din cele spuse până aici că identificarea şi rezolvarea timpurie, la fiecare nivel de prelucrare, a ambiguităților este un imperativ al oricărui demers computațional privind prelucrarea limbajului natural. Şi cum cuvântul (sau mai exact spus, atomul lexical) este elementul primar în prelucrarea limbajului o mare parte a eforturilor de cercetare este îndreptată spre nivelul lexical al prelucrărilor. Metodele de etichetare morfo-lexicală (tagging), printre care etichetarea cu două niveluri şi modele de limbă combinate - amintită în capitolul 2, permit rezolvarea cu mare acuratețe a ambiguităților categoriale şi intracategoriale. De pildă cuvântul *vin* poate fi atât substantiv cât şi verb (ambiguitate categorială), iar ca verb, el conține ambiguitatea intracategorială de persoană, număr şi mod ("indicativ + persoana I + număr singular", "conjunctiv + persoana I + număr singular" sau "indicativ + persoana III + număr plural"). Un program de etichetare morfo-lexicală "instruit" corect pentru limba română este capabil să rezolve, în contextul apariției sale, astfel de ambiguități morfo-lexicale.

Curentul lexicalist, predominant în modelarea sintactică a limbajului natural, presupune precizarea în descrierea de dicționar a fiecărui cuvânt a proprietăților şi restricțiilor sale distribuționale sau colocaționale relevante pentru analiza sintactică. Pe baza acestor descrieri lexicalizate şi a contextului local, multe din potențialele ambiguități structurale pot fi eliminate, înaintea unei costisitoare analize sintactice, prin tehnici cunoscute sub numele de analiză sintactică parțială *(parțial parsing* sau *shallow parsing).*

Un cuvânt omograf, chiar după ce a fost corect clasificat din punctul de vedere al categoriei sale gramaticale şi al proprietăților sale distribuționale sau colocaționale, poate rămâne ambiguu din punct de vedere semantic. Identificarea sensului cu care este utilizat cuvântul polisemantic într-un context dat este desigur de mare interes. Există însă diferite grade de rafinare a noțiunii de sens, iar natura aplicației pentru care identificarea sensului este necesară poate impune o accepție a noțiunii de sens diferită de cea utilizată într-un dicționar explicativ. Să luăm, de pildă, problema traducerii automate. întrucât în imensa majoritate a cazurilor rezultatul traducerii este destinat uzului uman, ceea ce este important este ca în textul tradus să nu apară ambiguități suplimentare față de cele din textul sursă. Cu alte cuvinte, dacă o analiză algoritmică evidențiază în limba sursă o serie de ambiguități, pornind de la premiza că textul este admisibil pentru vorbitorii nativi ai limbii textului sursă, de cele mai multe ori este nenaturală o traducere ce încearcă să evite total ambiguitatea identificată. La nivel lexical, aceasta revine la a spune că dacă diferitele sensuri ale unui cuvânt din limba sursă nu se lexicalizează prin cuvinte diferite în limba țintă, este neproductivă o încercare a diferențierii sensului contextual, atâta timp cât indiferent care ar fi el, traducerea cuvântului respectiv în limba țintă este aceeaşi. De exemplu, cuvântul englezesc "bottle" are în Wordnetl .5 [25] două sensuri (ca substantiv) anume de vas de sticlă sau plastic cilindric cu un gât îngust şi fără mâner, respectiv cantitatea de substanță conținută într-un astfel de vas. Ambele sensuri se regăsesc în cuvântul românesc "sticlă" (care însă include şi alte sensuri lexicalizate în engleză prin cuvîntul "glass"). în acest caz, a încerca eliminarea ambiguității la traducerea textului "He drank only a bottle of beer" în limba română, de pildă prin utilizarea unei parafraze de genul "El băuse doar conținutul unei sticle de bere", este nenecesară. Orice vorbitor al limbi române va găsi traducerea "El băuse doar o sticlă de bere" mult mai naturală ş desigur nu va avea dificultăți în a înțelege despre ce este vorba.

Acelaşi gen de considerații se poate face şi în raport cu ambiguitățile sintactice pure. Celebrul exemplu " I saw the Statue of Libery flying over New York conține cel puțin 4 ambiguități, dar dacă de pildă rezolvarea omografului *saw* (am văzut / tai cu fierăstrăul) este esențială în traducere, rezolvarea ambiguități structurale poate fi lăsată în sarcina minții celui ce citeşte textul: "Am văzut Statuia Libertății zburând deasupra New York-ului ", căci dacă cititorul englez nu are dificultăți în a înțelege cine şi cum zbura, e plauzibil că nici cititorul român (de exemplu) nu le va avea. Aceasta nu înseamnă că nu există ambiguități structurale a căror nerezolvare prealabilă să nu conducă la traduceri hazlii sau chiar incomprehensibile. Ideea este că metodele formale de analiză a limbajului modelabile algoritmic, explicitează de multe ori ambiguități greu de conştientizat de omul obişnuit, iar luarea în considerare a factorului uman poate simplifica mult prelucrările automate. Reconsiderarea conceptului de traducere automată în accepțiunea clasică (MT) în favoarea unor concepte mai realiste de tipul HAMT (human assisted machine translation) sau MAHT (machine assisted human translation) a relevat faptul că, în numeroase ocazii, posteditarea umană a unui text tradus automat introduce ambiguități care, deşi nu sunt sezizabile uşor la lectură, pot fi totuşi puse în evidență de algoritmii de analiză.

Cercetările moderne în domeniul dezambiguizării automate, în context, a sensurilor cuvintelor sunt motivate şi de alte aplicații informatice, cum ar fi clasificarea după conținut a volumelor mari de texte, regăsirea mai precisă a documentelor electronice, rezumarea automată a textelor, extragerea de cunoştințe din texte, crearea de ontologii. Această direcție de cercetare, identificată în literatura engleză prin acronimul WSD (Word Sense Disambiguation), constituie de câțiva ani obiectul unor conferințe specializate şi chiar a unor competiții de evaluare (SENSEVAL, ajunsă la a treia ediție) a soluțiilor propuse de specialişti din întreaga lume.

Primii care au sugerat ideea că, pentru obiectivele WSD, sensurile ce trebuie diferențiate sunt cele care se lexicalizează într-o altă limbă prin cuvinte diferite au fost Resnik and Yarowsky [26]. Intuitiv, se poate presupune că, dacă un cuvânt din limba sursă se traduce în limba țintă în mai multe feluri şi aces

traduceri nu sunt sinonimice, atunci trebuie să existe o motivație conceptuală. Analizând un număr suficient de mare de limbi şi de texte, e plauzibil, afirmau cei doi specialişti, să identificăm diferențierile lexicale semnificative care delimitează sensurile unui cuvânt. Aceste sensuri sunt numite de cei doi *"sensuri tari"*. Inabilitatea de a identifica corect sensurile tari este principala sursă a erorilor inacceptabile în orice aplicație multilingvală. Utilizarea textelor paralele pentru WSD [27], [28], [29], în scopul identificării proprietăților semantice a lexemelor şi a relațiilor dintre ele [30] a folosit implicit sau explicit noțiunea de "sens tare", Mai recent, pe baza echivalenților de traducere extraşi din corpusul "1984" prin procedura noastră, descrisă în capitolul precedent, Ide [31] a arătat că diferențele de traducere în 5 limbi (din 4 familii diferite) pot constitui un criteriu extrem de eficace în identificarea sensurilor tari în limba de pornire (în acest caz, engleza). Resnik and Yarowsky [32] au folosit în schimb traducerile unor propoziții izolate în limba engleză efectuate de vorbitori nativi ai limbilor țintă, dar în mare concluziile studiului lor au fost aceleaşi cu ale lui Ide. în ambele studii amintite referința pentru limba engleză a fost WordNet [33] şi deşi rezultatele lor sunt promițătoare, mai ales pentru sensurile tari, ele se bazează pe o mulțime prestabilită de sensuri. Date fiind divergențele semnificative între distincțiile de sens realizate în dicționarele (monolingve) existente, precum şi inexistența unui acord general asupra gradului de rafinare a descrierilor de sens în practica lexicografică internațională, raportarea la un inventar prestabilit de sensuri, cel puțin din perspectiva prelucrării automate a limbajului, nu pare a fi o soluție optimă. în continuare, vom prezenta o abordare alternativă, detaliată în [34], [35].

3.2. Discriminarea automată a sensurilor lexicale: metodologia

Metoda pe care o vom descrie este menită a identifica sensurile distincte cu care unul sau mai multe cuvinte apar într-un text dat. întrucât este foarte improbabil ca într-un text omogen, chiar foarte lung (de pildă un roman), un cuvânt să fie folosit în toate sensurile sale, metoda desigur va identifica, prin analiza textuală descrisă în continuare, doar acel sens sau acele sensuri cu care este folosit cuvântul respectiv în textul prelucrat. La limită prin prelucrarea unor texte foarte diferite este posibil teoretic să fie identificate toate sensurile atestate ale unui anumit cuvânt.

Din punct de vedere metodologic, studiul nostru s-a bazat pe corpusul paralel multilingv "1984" şi pe dicționarul multilingv extras din acest corpus. Cele 7 limbi ale experimentului nostru fac parte din patru familii: germanică (engleza), romanică (româna), slavică (bulgara, ceha şi slovena) şi ugro-finică (estoniana, maghiara). Deşi corpusul conține un text beletristic, textul orwelian ca şi traducerile sale în celelalte limbi nu sunt foarte stilizate şi, ca atare, oferă un eşantion rezonabil de limbă modernă, comună. Mai mult, traducerile textului original, efectuate de translatori avizați (unii dintre ei fiind apreciați scriitori), reflectă riguros originalul: pentru mai mult de 95% din textul englezesc o frază sursă este tradusă * în celelalte limbi tot ca o singură frază ;I corpusul "1984" sunt prezentate în tabela . Tipurile de alinieri frazale existente în de mai jos şi discutate în [7]:

Estoniană-Engleză

Tip	Nr.	Proc
3-1	2	0.030321%
2-2	3	0.045482%
2-1	60	0.909642%
1-3	1	0.015161%
1-2	100	1.516070%
1-1	**6426**	**97.422681%**
1-0	1	0.015161%
0-2	1	0.015161%
0-1	2	0.030321%

Maghiară-Engleză

Tip	Nr.	Proc
7-0	1	0.014997%
4-1	1	0.014997%
3-1	7	0.104979%
3-0	1	0.014997%
2-1	108	1.619676%
1-6	1	0.014997%
1-5	1	0.014997%
1-2	46	0.689862%
1-1	**6479**	**97.165573%**
0-4	**1**	0.014997%
0-2	3	0.044991%
0-1	19	0.284943%

Română-Engleză

Tip	Nr.	Proc
3-1	3	0.046656%
2-4	1	0.015552%
2-3	3	0.046656%
2-2	2	0.031104%
2-1	85	1.321928%
2-0	1	0.015552%
1-5	1	0.015552%
1-3	14	0.217729%
1-2	259	4.027994%
1-1	**6047**	**94.043551%**
0-3	2	0.031104%
0-2	2	0.031104%
0-1	10	0.155521%

Bulgară- Engleză

2-2	2	0.030017%
2-1	23	0.345190%
1-2	72	1.080594%
1-1	**6558**	**98.424134%**
0-1	8	0.120066%

Cehă- Engleză

4-1	1	0.015029%
3-1	2	0.030057%
2-1	109	1.638112%
1-3	2	0.030057%
1-2	81	1.217313%
1-1	**6438**	**96.753832%**
0-1	21	0.315600%

Slovenă- Engleză

3-3	1	0.014970%
2-1	48	0.718563%
1-5	1	0.014970%
1-2	53	0.793413%
1-1	**6572**	**98.383234%**
1-0	2	0.029940%
0-1	3	0.044910%

Figura 6: Distribuția tipurilor de aliniere frazală în corpusul paralel "1984"

Alinierile de tipul N:M reprezintă situațiile în care M fraze din limba englez au fost traduse cu N fraze în limba respectivă. Un caz particular îl reprezint situațiile de omisiune în traducere (0:M) sau de inserare de text fără coresponden în original (N:0).

3.3. Experimentul inițial

Textul original "1984" conține 7.069 leme diferite, iar dicționarul multiling extras prin metoda descrisă în prima parte a acestei lucrări conține 1.233 de intră Aceste intrări au fost reținute respectând condiția ca un articol lexical din limb engleză să aibă traduceri (eventual multiple) în cât mai multe limbi țintă. Condi

impusă dicționarului multilingv este foarte restrictivă, având în vedere că majoritatea dicționarelor bilingve extrase automat conțin între 6000 şi 7000 de intrări. Intrări tipice (parțiale) în dicționarul multilingv sunt ilustrate în figura 7. O informație suplimentară, ce nu apare în exemplificarea din figura 7, este mulțimea tuturor unităților de traducere din corpusul paralel în care cuvântul englezesc a fost tradus prin echivalenții săi listați în dicționar. Dintre aceste intrări, au fost selectate 845 pentru care s-au găsit una sau mai multe traduceri în toate limbile. Dintre acestea, s-a ales o mulțime de 33 de substantive, acoperind toate gamele de frecvență şi ambiguitate, cu care s-a realizat experimentul ale cărui rezultate au fost validate de experți umani [34].

Engleză	Categorie	Bulgară	Cehă	Estoniană	Maghiară	Română	Slovenă
finally	R	HaKpaa	nakonec konecne	lopuks viimaks	vegul	în_cele_din_urmă până_la_urmă	koncen nazadnje
wealth	N	6oraTCTBo 6uaro	bohatstvi	joukus rikkus	jolet gazdagsâg	avuție bogăție 1	blaginja bogastvo

Figura **7:** Exemple de echivalenți de traducere identificați în corpusul paralel "1984"

Pentru fiecare substantiv din acest eşantion au fost extrase toate frazele englezeşti în care apare, împreună cu toate frazele corespunzătoare din celelalte limbi şi pentru fiecare ocurență a sa a fost construit un vector binar reprezentând toate traducerile posibile ale cuvântului respectiv. O valoare 1 în poziția *n* a acestui vector semnifică faptul că acea ocurență a fost tradusă prin cuvântul ce reprezintă a A?-a traducere posibilă. O valoare 0 semnifică faptul că a n-a traducere posibilă nu a fost folosită. De pildă pentru substantivul *"wealth"* (vezi figura 7) au fost depistate 11 traduceri posibile (2 în bulgară, estoniană, maghiară, română şi slovenă, 1 în cehă). Un vector asociat oricărei ocurențe a lui *wealth* va avea prin urmare 11 poziții. Astfel, dacă a m-a apariție în textul original al romanului "1984" a cuvântului *wealth* are ataşat vectorul 10101010101 acest lucru semnifică faptul că în varianta bulgărească el a fost tradus cu *6oeamcmeo,* în cea cehă cu ***bohatstvi,*** în cea estoniană cu *rikkus,* în cea maghiară cu *gazdagsâg,* în cea română cu *bogăție* iar în cea slovenă cu *bogastvo.* Vectorii astfel definiți au fost prelucrați cu un algoritm de clasificare de tip aglomerativ [36], clasele rezultate fiind

considerate a reprezenta sensuri distincte în care cuvântul curent a fost folosit de a lungul romanului. Clasele produse de algoritm au fost comparate cu clasel rezultate prin dezambiguizarea manuală efectuată, independent, de 2 vorbito nativi ai limbii engleze. Dezambiguizarea manuală a fost efectuată utilizân numerotarea sensurilor din WordNet 1.6.

Pentru a putea compara rezultatele produse de dezambiguizatorii uman (numiți în continuare adnotatori) cu cele produse de algoritmul nostru, datele a fost normalizate în feiui următor: pentru fiecare adnotator şi pentru algoritm, fiecar din cele 33 de cuvinte a fost reprezentat printr-un vector binar de lungime *n,* und *n* este numărul de ocurențe ale cuvântului în tot corpusul. Pozițiile în vect reprezintă o asignare de tip "DA/NU" indicând dacă ocurența respectivă a fo clasificată la fel de către adnotatori, respectiv algoritm. Rezultatele acestui pri experiment sunt rezumate în tabelul din figura 8 indicând procentul de acord înt clasificările propuse de algoritm şi cele ale fiecărui adnotator, acordul dintre c doi adnotatori şi acordul dintre toți cei trei clasificatori.

Algoritm/Adnotator 1	66.7%
Algoritm /Adnotator 2	63.6%
Adnotator 1/Adnotator 2	76.3%
Algoritm /Adnotator 1/ Adnotator 2	53.4%

Figura **8:** Concordanța între diferite clasificări

3.4. Cel de-al doilea experiment

Rezultatele primului experiment au arătat că metoda discrimin sensurilor folosind echivalenții de traducere este foarte competitivă, acurate procesului fiind comparabilă (şi uneori superioară) cu performanțele obținute de cercetători ce au folosit ca referință acelaşi dicționar (Wordnet). Mai m diferențele de acord asupra clasificării dintre cei 2 adnotatori pe de o parte şi dir fiecare adnotator şi algoritm pe de altă parte este de numai 10-13%, ceea ce nou este foarte competitiv în raport cu scorurile obținute în alte experimente.

Pentru a valida aceste rezultate empirice, în cea de a doua fază experimentului a fost luat în considerare un număr dublu de substantive (76) di cele "dificile", adică cu grad de ambiguitate mare, atât din clasa celor abstracte şi a celor concrete (de exemplu, "thought", "stuff, "meaning", "feeling" respe "hand", "boot", "glass", "girl" etc). Am ales acele substantive care au apărut puțin de 10 ori în corpus (pentru a elimina efectul de "insuficiență a datelor") ş plus care au cel puțin 5 traduceri în cele 6 limbi țintă. Restricția de 10 aparit

S S K ^ ^ W Pe care ,-am impus procesului de ^ g

$$LL(T_T, T_s) = 2 * \sum_{7=1}^{} \sum_{i=1} n_{ij, Jog} n > 18$$

aj. -a doi vorbitori nativi e .chetată, în mod independent de 5; c l a s a *ƒ CUVintelorse.ectate * fost a<ci. In tabela din figura 9 sunt reLmtte SP, ș/ a.90ritmul discutat adnotatori: rezumate datele ş. rezultatele de acord între cei 4

Nr. de cuvânte	76
Nr. ocurenţe	2399
Număr mediu de ocurenţe/cuvânt	32
Nr. de sensuri găsite de adnotatorul 1	241
Nr. de sensuri găsite de adnotatorul 2	280
Nr. de sensuri găsite de adnotatorul 3	213
Nr. de sensuri găsite de adnotatorul 4	232
Nr. de sensuri găsite împreună de toţi adnotatorii	345
Numărul mediu de sensuri pe cuvânt	4.53
Procent de acord între adnotatori	
Acord total (4/4)	54.27
75% acord total (3/4)	28.13
50% acord total (2/4)	16.92
Dezacord total	0.66

Figura 9: Datele experimentului şi acordul între 4 adnotatori umani independenți

Rezultatele produse de algoritmul de clasificare şi clasificările realizate de adnotatori prin asignarea sensurilor din Wordnet1.6 au fost de data aceasta normalizate în mod diferit, prin ignorarea etichetei puse de adnotatori şi considerând doar clasele rezultând din această etichetare. Pentru a clarifica acest aspect să urmărim modul în care doi dintre adnotatori au dezambiguizat cele 7 ocurențe ale cuvântului "youth":

Ocurența nr.	1	2	3	4	5	6	7
Adnotatorul 1	1[3]	1	6	3	6	3	1
Adnotatorul 2	1[2]	1	4	2	6	2	1

Figura 10: Acordul de clasificare pentru cuvântul" youth" î umani independenți

Acordul între cei doi adnotatori este doar de 43% (doar au asignate sensuri consensuale); totuşi, ambii adnotatori au 1, 4, şi 6 ca având acelaşi sens, deşi primul le-a etichetat cu s în timp ce al doilea le-a etichetat cu sensul 2. Dacă îns clasificarea celor 3 ocurențe este consistentă, în sensul că decis că ele au acelaşi sens. Acordul de clasificare se dublea datele sunt mult mai uşor de comparat cu rezultatele produse

în acest al doilea experiment am luat în consid momentului optim de oprire a clasificării aglomerative. în pri folosit o distanță minimă predefinită, pentru determinarea n între care se realizează discriminarea. Această soluție nu proprietățile individuale ale cuvintelor (numărul maxim de se Wordnet, frecvența de apariție a cuvântului, numărul mediu de a primit cuvântul în corpus). Noul algoritm de clasificare a fost să-şi calculeze un număr optim de clase[2], optimalitatea fiind j numărul mediu de clase identificate de adnotatori. Drept c folosit distanța minimă dintre clasele existente la fiecare pas d pas de aglomerare, clasele cu cea mai mică distanță relativ clasă mai mare. Procesul începe cu fiecare ocurență într-o opreşte când distanțele relative între clasele existente este Distanța dintre două clase se calculează pe baza ve (centroizii) ai celor două clase (evident depinzând de cuv ocurențe şi de numărul de sensuri ale cuvântului clasificat):

$$dist(v\setminus, V2) = /£_i(v_i(i)-V_2(i))^2$$
$$i=I$$

[1] *Singurul dezacord rămas constă în faptul că Adnotatorul 1 consid având acelaşi sens, în timp ce Adnotatorul 2 atribuie un sens dife realizând o discriminare mai fină între sensurile celor două ocurențe*

[2] *în principiu, limita superioară a numărului de sensuri pe care îl englezesc într-un text este dată de numărul de sensuri listate în W de aşteptat însă nu există în corpusul nostru nici un exemplu polisemantic să fi apărut cu toate sensurile din WordNet*

Cele mai bune rezultate în discriminarea automată au fost obținute pe cale experimentală, impunând drept criteriu de oprire a algoritmului condiția:

$$\frac{mindist(k) - mindist(k+1)}{mindist(k+1)} \wedge \wedge \wedge$$

în care *mindist(k)* reprezintă distanța minimă între clasele existente la pasul *k* de aglomerare.

Pentru medierea opiniilor adnotatorilor am definit o adnotare de referință reprezentând clasificarea majoritară între cei 4. în cazul egalității de voturi, adnotatorul care a fost în cele mai multe cazuri de aceeaşi opinie cu majoritatea a impus clasa. Folosind această clasificare mediată şi raportând-o la clasificarea produsă de algoritm pentru cele 76 de cuvinte, am analizat diferențele de clasificare, considerate ca fiind erori. Marea majoritate a erorilor de clasificare pentru cele 2399 de ocurențe au apărut în cazul cuvintelor pentru care distribuția sensurilor este foarte inegală; ca urmare am adăugat algoritmului o fază suplimentară de postprocesare, în care clasele cu un număr mult mai mic de ocurențe decât clasa cu cele mai multe ocurențe au fost incorporate în ultima. Raportul minim între numărul de ocurențe al celei mai mari clase şi numărul de ocurențe din clasele potențial absorbabile în cea dintâi a fost ales empiric ca fiind 10[1]. Motivația acestei euristici constă în constatarea făcută de mai mulți cercetători în domeniul lingvisticii corpusului (fapt sugerat chiar de Zipf cu peste 50 de ani în urmă) că utilizarea frecventă a unui cuvânt într-un text omogen tinde să-i păstreze sensul.

Cu această nouă euristică încorporată, algoritmul de clasificare a atins cifra de 74,6% acord cu clasificarea mediată. în [35] sunt prezentate alte variante ale algoritmului care au condus prin evaluarea empirică la versiunea sa finală. Clasele produse de fiecare pereche de clasificatori (om sau maşină) au fost evaluate printr-un algoritm ce calculează alinierea claselor astfel încât intersecția lor să fie maximală. Diferențele dintre două clase astfel aliniate au fost considerate dezacorduri de clasificare. Scorul de acord a fost calculat ca fiind raportul dintre suma numărului de ocurențe comune pentru fiecare clasă aliniată şi numărul total al ocurențelor cuvântului respectiv. In tabela din figura 10 este exemplificat modul de calcul al acordului dintre clasificarea produsă de algoritm şi clasificarea mediată a adnotatorilor pentru cuvântul *movement*. Acesta a apărut în text de 40 de ori. Atât algoritmul cât şi cei patru adnotatori au identificat 4 sensuri distincte în care acest cuvânt a fost utilizat. Aşa cum se vede din figura 10, cea mai numeroasă clasă (clasa 1) conține în clasificarea mediată 28 dintre cele 40 de apariții ale cuvântului *movement,* în timp ce clasa corespondentă creată de algoritm conține doar 25 de ocurențe. Dintre acestea, 24 sunt comune cu cele din clasa 1 a clasificării mediate. în conformitate^cu definiția anterioară a scorului de acord, pentru acest exemplu avem

$$\frac{rezultat\ corect}{n\ cazuri} = \frac{24}{34} = 85\%$$

CLASA	1	2	3	4	2.
Clasificare mediată	28	6	3	3	40
Clasificare algoritmică	25	7	6	2	40
Intersecţie	24	6	3	1	34
Precizie			85%		

Figura 11: Clasificarea mediata si cea produsă de aigoritm pentru cuvântu. *movement*

3.5. Rezultate

Rezultatele obținute cu ultima variantă a clasificatorului în cel de-al doilea experiment sunt sintetizate in tabelul din figura 12. Tabelul indică procentul de acord între diverse clasificări: 1, 2, 3, 4, reprezintă clasificările realizate de adnotatorii umani, M reprezintă clasificarea mediată a clasificatorilor umani, A reprezintă clasificarea produsă de algoritm, iar B este referința de bază (baseline) care presupune toate ocurențele unui cuvânt ca având acelaşi sens.

	1	2	3	4	M	A
B	71.1	65.1	76.3	74.1	75.5	81.5
1		78.1	75.6	83.1	88.6	74.4
2			71.3	75.9	82.5	66.9
3				77.3	82.1	77.1
4					90.4	75.9
M					1	77.3

Figura **12:** Acorduri între diverse clasificări

Tabela arată că acordul între adnotatorii umani comparat cu cel dintre algoritm şi adnotatorii umani (cu excepția unuia dintre ei (4), pe care îl suspectăm că a văzut clasificările celorlalți trei şi în consecință şi-a revizuit unele decizii) nu diferă substanțial. Acest lucru demonstrează (cel puțin în raport cu datele experimentului nostru) că dezambiguizarea automată este comparabilă ca acuratețe cu cea efectuată de adnotatori umani. Diferența fundamentală constă

10 este o valoare precaută; experimente viitore, de mai mare anvergură, ar putea furniza argumente pentru coborârea acestui prag.

• faptul că programul a terminat în circa 2 minute clasificarea pentru care adnotatorilor le-au trebuit între 4 şi 5 săptămâni.

Experimentul descris a evaluat dezambiguizarea automată a cuvintelor englezeşti pornind de la traducerea lor în celelalte 6 limbi. Această direcționare a fost impusă doar de disponibilitatea pentru limba engleză a textului dezambiguizat de experți umani (vorbitori nativi ai limbii engleze). întrucât algoritmul de clasificare nu depinde în nici un mod de limba pentru care se realizează dezambiguizarea (limba țintă) şi nici de limbile martor în raport cu care se face acest proces, rezultă că exact acelaşi procedeu descris până aici poate fi folosit pentru dezambiguizarea cuvintelor româneşti folosind echivalenții lor de traducere în engleză, bulgară, cehă, estoniană, maghiară şi slovenă, ori pentru dezambiguizarea cuvintelor bulgăreşti pe baza echivalenților lor de traducere în celelalte 6 limbi. întrucât sensul este (în principiu) un invariant al traducerii, nu pare a se justifica şi pentru celelalte limbi efortul de adnotare umană făcut pentru limba engleză. Este rațional a presupune că rezultate similare (raporturi relative) s-ar obține indiferent de limba țintă şi de limbile martor.

Să mai menționăm şi faptul că există o anumită corelație (factorul Spearman - 0.51) între numărul de sensuri în Wordnet ale unui cuvânt şi nivelul de acord între diferitele clasificări ale ocurențelor sale. Cele mai scăzute scoruri de acord au fost obținute pentru "line" (29 sensuri), "step" (10), position (15), "place" (17) şi "corner" (11). Acorduri perfecte s-au obținut pentru majoritatea cuvintelor cu mai puțin de 5 sensuri, ca de exemplu "hair" (5), "morning" (4), "sister" (4), "tree" (2), and "waist" (2) care toate au fost considerate, atât de adnotatori cât şi de algoritm, a fi fost folosite cu un singur sens în tot textul. Pe de altă parte, gradul de acord pentru câteva cuvinte cu mai puțin de 5 sensuri ("rubbish" (2), "rhyme" (2), "destruction" (3) şi "belief (3)) a fost semnificativ mai mic decât media pentru toate perechile de clasificări (adnotator-adnotator, adnotator-algoritm). Concluzia a fost că pentru unele cuvinte, distincțiile de sens sunt atât de fine în Wordnet, încât chiar vorbitorii nativi (şi cu atât mai mult algoritmul de clasificare) nu pot face diferențieri sistematice de sens ale diferitelor ocurențe ale acestor cuvinte. O astfel de hiperdiferențiere a sensurilor este în imensa majoritate a cazurilor irelevantă pentru aplicațiile de prelucrare a limbajului natural.

4. Concluzii

Rezultatele experimentelor noastre arată că acuratețea discriminării sensurilor pe baza echivalenților de traducere extraşi din corpusuri paralele este comparabilă cu cea produsă de adnotatori umani. întrucât abordarea noastră este complet automatizată ea poate fi folosită la crearea de volume mari de texte, având discriminate sensurile cuvintelor polisemantice. Utilizarea experților umani este prohibitivă sub aspectul costului şi al timpului de realizare a unei asemenea sarcini, iar procentajul suplimentar de acuratețe, presupus de activitatea umană este prea mic pentru a justifica procedurile manuale.

Metoda pe care am descris-o în această lucrare nu etichetează clasele d ocurențe ale unui cuvânt cu un număr de sens ales dintr-un inventar prescris d sensuri iar majoritatea aplicațiilor de prelucrare a limbajului natural (de pild clasificarea textelor, regăsirea informațiilor, rezumarea automată etc.) nici nu a nevoie de această informație suplimentară; pentru aceste tipuri de aplicații es suficient a decide că două sau mai multe ocurențe ale unui cuvânt sunt folosite acelaşi sens sau nu. O etichetare convențională a sensurilor identificate pentru u anumit cuvânt ar putea să se bazeze pe frecvența sensurilor respective (sensul corespunzând clasei cu cele mai multe ocurențe). Evident o astfel de etichetar depinde de registrul lingvistic al textului pe baza căruia se identifică sensur distincte.

O direcție foarte promițătoare [37], [38], [39] o constituie utilizar metodologiei prezentată aici în construcția şi validarea ontologiilor multilingve tip EuroWordNet. Folosind echivalenții de traducere şi clasificarea ocurențe echivalente din punctul de vedere al sensului se poate verifica dacă proiec interlinguală a două sau mai multe dicționare semantice este corectă. Aceas presupune că sensurile cuvintelor extrase ca echivalenți de traducere ai cuvinte englezeşti dezambiguizate să fie puse în corespondență cu acelaşi conce interlingual aparținând Indexului Interlingual (ILI - vezi [39]- în acest volum). cazul contrar (echivalenții de traducere sunt puşi în corespondență cu concep interlinguale diferite) este fie vorba de o eroare propriu-zisă de proiec conceptuală într-unui sau mai multe dintre dicționarele semantice aliniate conceptele interlinguale sunt atât de apropiate semantic încât se poate propu unificarea lor într-un concept mai general cu lexicalizare în mai multe lim Aceasta este esența conceptului de "soft-clustering" definit în comunita EuroWordNet. Față de identificarea prin metode statistice a concepte interlinguale foarte apropiate semantic, analiza prin metoda "echivalenților traducere şi a discriminării sensurilor" a proiecțiilor sensurilor făcute de lexicog profesionişti peste o mulțime de sensuri conceptualizate în ILI este mult robustă. Experimentele preliminare discutate în [37] au arătat că în diferite li pentru care se realizează o ontologie lexicală multilingvă (bulgară, cehă, grea română, sârbă, turcă) există dificultăți identice de proiecție conceptuală sensurilor unor cuvinte din limbile considerate. Faptul că aceleaşi conce interlinguale creează acelaşi tip de dificultate în proiecția sensurilor unor cuv aparținând unor limbi foarte diferite indică cu claritate că acele concepte trel generalizate.

Un alt aspect care merită subliniat este că metodologia prezentată a corelată cu existența a tot mai multe dicționare semantice de tip Wordnet, ce ad la principiul EuroWordNet de aliniere la Indexul Interlingual, va permite dezvolta

de corpusuri adnotate semantic (de tipul SemCor) pentru orice limbă. Tranzitivitatea relațiilor de tip "EQ-SYN" folosite în proiecția sinseturilor unui wordnet monolingv peste ILI, corelată cu echivalența de traducere (relație tot între sensuri) extrasă dintr-un corpus paralel, în care textul dintr-una din limbi este adnotat semantic, permite importul adnotărilor în toate celelalte limbi. Deoarece limba din care se importă adnotarea semantică nu este relevantă pentru această procedură, rezultă că eforturile depuse de-a lungul timpului în crearea celor câteva corpusuri cu adnotare semantică pentru limbile "mari" pot fi valorificate pentru orice altă limbă în care există (sau se creează) traduceri ale textelor din corpusurile adnotate. Mai mult, se poate imagina crearea unui consorțiu multilingv care să aleagă un corpus paralel în cât mai multe limbi cu scopul de a-l adnota semantic. Prin adnotarea independentă, în fiecare limbă, a unor porțiuni distincte din corpusul paralel, folosind o metodologie de genul celei prezentate în această lucrare (şi desigur având un dicționar semantic multilingv de tip EuroWordNet) adnotările secțiunilor monolingve vor putea fi importate în secțiunile corespunzătoare ale tuturor celorlalte texte monolingve, în final putându-se obține adnotarea semantică, consistentă, a întregului text din fiecare limbă a corpusului paralel.

Mulțumiri

Rezultatele prezentate în această lucrare sunt rodul mai multor proiecte internaționale de cercetare desfăşurate la Institutul de Inteligentă Artificială, alături de colegii Ana Măria Barbu, Eduard Barbu, Radu Ion, Cătălin Mititelu, Octavian Popescu. De asemenea/colaborarea cu Nancy Ide de la Universitatea Vassar din Poughkeepsie, SUA, şi cu Tomaz Erjavec de la Institutul "Jozef Ştefan" din Ljubljana, Slovenia, parteneri în proiectele amintite, a fost şi este extrem de productivă. Tuturor le aduc aici cuvenitele mulțumiri.

Referințe bibliografice

[1] Melamed, D. - *"Empirical Methods for Exploiting Parallel Texts",* MIT Press, 2001, 373p.

[2] Gale, W.A., K.W. Church, - "Identifying word correspondences in parallel texts". In Fourth *DARPA Workshop on Speech and Natural Language,* 1991, 152:157

[3] Smadja, F., K.R. McKeown, and V. Hatzivassiloglou - "Translating collocations for bilingual lexicons: A statistical approach". *Computațional Linguistics,* 22/1, 1996, 1:38.

[4] Brown, P., Della Pietra, S. A., Della Pietra, V. J„ Merc mathematics of statistical machine translation: paramete *Computațional Linguisticsl* 9(2): 263-311, 1993.

[5] Kupiec, J. - "An algorithm for finding noun phrase correspond corpora". In *Proceedings of the 31st Annual Meeting of t Computațional Linguistics,* 1993,17:22

[6] Hiemstra, D. - "Deriving a bilingual lexicon for cross lan retrieval". In *Proceedings ofGronics,* 1997, 21:26

[7] Tufiş, D., Barbu, A.M. - "Automatic Learning of Translatio "Romanian Journal on Information Science and Techno Academy, vol.4, no. 3-4, 2001b, 325:351.

[8] Tufiş, D., Barbu, A.M. - "Revealing translators knowledge: in constructing practicai translation lexicons for langu processing", în *International Journal of Speech Technology.* Publishers, no.5, 2002, 199:209.

[9] Tufiş, D. - "Parțial translations recovery in a 1:1 word-ali RACAI Research report, June, 2001b, 32pp.

[10] Gale, W.A., K.W. Church - "A Program for Aligning Sen Corpora". în *Computațional Linguistics,* 19(1), 1993, 75:102

[11] Tufiş, D. - "Tiered Tagging and Combined Classifiers". în (eds) Text, Speech and Dialogue, Lecture Notes in Artificia Springer, 1999, 29:33.

[12] Brants, T. - "TnT - A Statistical Part-of-Speech Tagger", î *Sixth Applied Natural Language Processing Conference, Al* - May 3, 2000, Seattle, WA, 2000

[13] Varadi, T. - The Hungarian National Corpus, Proceedings Palmas, Spain, 2002, 385:389.

[14] Tufiş, D. - "A cheap and fast way to build useful tran *Proceedings of the 19th International Conference on Com COLING2002,* Taipei, China, 2002, 246:251.

[15] Tufiş, D., Dienes, P., Oravecz, C, Vâradi T.,*- "Princip Design for Tiered Tagging of Hungarian" *Proceedir International Conference on Language Resources and* Athens, Greece, 2000,1421:1426

[16] Hinrics, H., Trushkina, J. - "Forging Agreement: Morpholo of Noun Phrases", Proceedings of the Workshop on Tree Theories 2002, Sozopol, Bulgaria, 2002, 1:18.

[17] Erjavec. T. - "An Experiment in Automatic Bi-lingual Lexicon Construction from a Parallel Corpus", Proceedings of the 7th TELRI International Seminar on Corpus Linguistics, Dubrovnik, Croația, 2002.

[18] Erjavec T., Ide, N. - "The Multext-East corpus". în *Proceedings of First International Conference on Language Resources and Evaluation,* Granada, Spain, 1998,971:974

[19]Tufiş, D., Barbu, A.M., Pătraşcu, V., Rotariu, G., Popescu, C. - "Corpora and Corpus-Based Morpho-Lexical Processing", în D. Tufiş, P. Andersen (eds.) "Recent Advances in Romanian Language Technology", Editura Academiei, 1997, 35:56

[20]Kay, M., Roscheisen, M. - "Text-Translation Alignment". în *Computațional Linguistics,* 19/1, 1993, 121:142

[21]Brew, C, McKelvie, D. - "Word-pair extraction for lexicography", 1996, http:///www. ltg.ed.ac.uk/ ~chrisbr/papers/nemplap96

[22]Tiedemann, J. - "Extraction of Translation Equivalents from Parallel Corpora", în *Proceedings of the 11th Nordic Conference on Computațional Linguistics,* Center for Sprogteknologi, Copenhagen, 1998, http://stp.linq.uu.se/ -ioerg/

[23]Ahrenberg, L., M. Andersson, M. Merkel - "A knowledge-lite approach to word alignment", în [40].

[24] Cristea, D., Dima, G. E. - "An Integrating Framework for Anaphora Resolution", Journal on *Information Science and Technology, Romanian Academy Publishing House,* Bucharest, voi. 4, no. 3, 2001, 273:292.

[25]Fellbaum C. - Wordnet: An Electronic Lexical Database, MIT Press, 1998, 423p.

[26]Resnik, P. and Yarowsky, D. - A perspective on word sense disambiguation methods and their evaluation. *ACL-SIGLEX Workshop Tagging Text with Lexical Semantics: Why, What, and How?* Washington, D.C., 1997, 79:86.

[27] Gale, W. A., Church, K. W., Yarowsky, D. (1993). A method for disambiguating word senses in a large corpus. *Computers and the Humanities,* 26, 415:439.

[28]Dagan, I., Itai, A., Schwall, U. - Two languages are more informative than one. *Proceedings of the 29th Annual Meeting of the ACL,* 18-21 Berkeley, California, 1991, 130:137.

[29]Dagan, I., Itai, A. - Word sense disambiguation using a second language monolingual corpus. *Computațional Linguistics,* 20/4,1994, 563:596.

[30]Dyvik, H. (1998). Translations as Semantic Mirrors. *Proceedings of Workshop Multilinguality in the Lexicon II, ECAI98,* Brighton, UK, 1998, 24:44.

[31] Ide, N. - Cross-lingual sense determination: Can it work? *Computers. and the Humanities,* 34/1-2, 1999, 223:234.

[32]Resnik, P. and Yarowsky, D. - Distinguishing systems and distinguishir senses: New evaluation methods for word sense disambiguation. *Journal Natural Language Engineering,* 5(2), 2000,113:133.

[33]Miller, G. A., Beckwith, R. T. Fellbaum, C. D., Gross, D. and Miller, K. J. WordNet: An on-line lexical database. *International Journal of Lexicograph* 3/4,1990,235:244.

[34] Ide, N., Erjavec, T., and Tufiş, D. (2001). Automatic sense tagging usi parallel corpora. *Proceedings of the Sixth Natural Language Processing Pac Rim Symposium,Tokyo,* 2001,83:89.

[35] Ide, N., Erjavec, T., Tufiş, D. - "Sense Discrimination with Parallel Corpora" Proceedings of the SIGLEX Workshop on Word Sense Disambiguatic Recent Successes and Future Directions. ACL2002, July Philadelphia, 200 56:60

[36]Stolcke, A. - Cluster - 2.9. http://www.icsi.berkeley.edu/ftp/global/pub/ stolcke/software/cluster-2.9.tar.Z, 1996.

[37] Tufiş, D. - "Interlingual alignment of parallel semantic lexicons by means automatically extracted translation equivalents", Proceedings of the 7th TEL International Seminar on Corpus Linguistics, Dubrovnik, Croația, 2002.

[38]Tufiş, D., Cristea, D. - "Methodological issues in building the Roman Wordnet and consistency checks in Balkanet", în Proceedings of LREC20 Workshop on Wordnet Structures and Standardisation, Las Palmas, Spa 2002, 35-41.

[39]Tufiş, D., Cristea, D. - "RO-BALKANET - ontologie lexicalizată, în con multilingv, pentru limba română", 2002, în acest volum.

[40]Veronis, J. (ed), Parallel Text Processing. Text, Speech and Langu Technology Series, Kluwer Academic Publishers Voi. 13, 2000.

Referentialitate si cursivitate în relație cu structura de discurs

Dan CRISTEA

Universitatea "Al.I.Cuza" laşi, Facultatea de Informatică
Academia Română, Institutul de Informatică Teoretică - Filiala laşi

dcristea@infoiasi.ro

1. Introducere

în ultimii 25 de ani s-a studiat enorm pentru a se înțelege ce anume face dintr-un text (considerat o secvență de propoziții sintactic corecte) să fie un discurs, aşadar de ce un discurs e coerent şi ce elemente îi atribuie coeziune. Dintre teoriile computaționale ale discursului, trei au avut o influență covârşitoare asupra dezvoltărilor ultimilor ani din acest domeniu: teoria structurilor retorice, teoria stărilor atenționale şi teoria centrelor.

Dezvoltată inițial din perspectiva generării textelor, teoria structurilor retorice *{rhetorical structure theory,* de aici încolo RST), a fost elaborată de Mann şi Thompson ca o teorie a organizării textelor [Mann, Thompson, 1988; Hovy, 1988; Scott, de Souza, 1990]. Ea caracterizează structura de discurs în termeni de relații ce,leagă părți componente ale textului. Unitatea elementară de discurs în RST este, de regulă, o propoziție, aşadar o întindere textuală în care se formulează o predicație. O structură de discurs este descrisă de o **schemă.** Ea grupează o secvență de unități, sau de unități şi scheme, sau o secvență de scheme. O schemă poate fi asemuită cu o regulă a unei gramatici, ea relevând structura de constituenți a unui compus. O schemă constă dintr-o **relație** (27 în RST) care leagă două sau mai multe întinderi de text, fiecare dintre ele având, la rândul său, o structură (constituenții schemei). Un discurs este fie o **unitate,** care este o întindere de text elementar, fără structură, fie o schemă (un text mai lung decât o singură unitate şi care prezintă o structură). Relațiile pot fi de două tipuri: **hipotactice** - dacă argumentele sunt constituenți neegali ca importanță şi **paratactice** (sau echinucleare) - dacă constituenții pe care-i agregă sunt egali ca importanță. între constituenții uniți de relațiile hipotactice există întotdeauna unul singur mai important, numit **nucleu,** ceilalți fiind numiți **sateliți.** La relațiile paratactice, prin convenție, se consideră că toți constituenții sunt nucleari. Satelitul este, în general, mai susceptibil de a fi schimbat sau eliminat complet decât

nucleul, fără ca, prin aceasta, înțelesul discursului să se modifice. Dimpotrivă, înlocuirea sau ştergerea nucleului este o opțiune mult mai drastică, care poate duce la denaturări ale înțelesului. Relațiile hipotactice sunt, în general, cele intenționale, în care o întindere de text comunică un scop şi celelalte exprimă subscopuri ce completează, dezvoltă etc. scopul principal. Pe de altă parte, relațiile paratactice sunt, în general, de natură informațională, simetrice, neputându-se stabili dacă, sau care, componentă predomină.

în RST accentul este pus pe performanța retorică: prin ce mijloace un scriitor (sau vorbitor) reuşeşte să convingă un cititor (ascultător) de intențiile pe care le are de comunicat. Ca produs secundar al liniei principale de investigare în RST, multe eforturi de cercetare care au succedat elaborarea teoriei s-au concentrat asupra îmbunătățirii şi structurării setului de relații propus inițial [Rambow, 1993]. Pe de altă parte, pare extrem de convenabil, inclusiv din punct de vedere computațional, să vedem discursul reprezentat ca un arbore, în care nodurile terminale să reconstituie, în secvența lor, textul. Cu toate acestea RST nu aduce nici o lumină în privința vreunei legături care ar exista între structură şi referențialitate. RST este deci o teorie asupra structurii globale a discursului.

Teoria stărilor atenționate *(atenționai state theory,* AST) [Grosz, Sidner, 1986] reprezintă o dezvoltare a liniei de cercetare în discurs dominată de Barbara Grosz şi Candace Sidner asupra manierii în care focarul ori centrul de discurs *(focus* în engleză) se modifică pe parcursul derulării textului şi a recunoaşterii intențiilor comunicate de discurs [Grosz, 1981; Sidner, 1983]. Grosz şi Sidner nu cred că varietatea atât de mare a intențiilor ce pot fi comunicate de un discurs poate fi condensată într-un număr fix de şabloane retorice exprimate ca relații, cum sugerează RST sau tentative similare acesteia. Teoria se doreşte a fi un model formal, care se distanțează de detaliile ce ar putea fi asociate participanților la discurs. Realizând proiecții corespunzătoare utilizatorului de limbaj, însoțite de detalii specifice, ea s-ar putea regăsi atât în construcția unui sistem automat cât şi într-o teorie psihologică, ambele consumatoare de limbaj natural. Deşi recunoaşte însemnătatea mesajului transmis de un discurs, teoria nu abordează problema înțelesului discursului şi a manierei în care acesta poate fi dedus din elementele constitutive ale textului. Ea este, primordial, o teorie a structurii discursului, prin aceasta plasându-se la baza oricărei tentative de a aborda problema construirii sensului.

Conform lui Grosz şi Sidner intențiile joacă rolul principal în explicarea structurii discursului, în timp ce dinamica atenției joacă rolul principal în explicarea interpretării discursului. Structura discursului are trei componente distincte, dar strâns corelate:

- o **structură lingvistică,** care face ca una sau mai multe propoziții, exprimări *(utterance)* să fie agregate într-un segment de discurs iar limitele dintre segmente să fie indicate de expresii lingvistice, intonație, schimbări ale timpului şi aspectelor verbelor. Segmentul de discurs are însă o definiție recursivă: un segment poate îngloba alte segmente acestea pe altele ş.a.m.d.;
- o **structură intențională,** care face să vedem discursul ca având u scop global (scopul discursului - SD), scopul fundamental a vorbitorului/scriitorului la emiterea discursului. Fiecare segment are u scop al segmentului (scopul segmentului de discurs - SSD), care est un subscop al scopului segmentului din care face el parte. Intuitiv SSD specifică cum contribuie respectivul subsegment la realizare scopului segmentului din care face el parte. Teoria admite că nu exist o listă finită de scopuri ale discursului, care să facă posibilă comparație cu lista categoriilor gramaticale, de exemplu. Confor teoriei, două relații structurale sunt suficiente pentru a compun structura discursului: **relația de dominare** (dacă SSDi domină SS atunci SSD_2 contribuie la SSD_1) sau SSD_2 este intenționată s satisfacă parțial SSD^ şi **relația de satisfacere-precedență** (SS satisface-precede SSD_2 dacă *SSD^* trebuie satisfăcut înainte SSD_2);
- o **stare atenționată,** prin care se asociază fiecărui segment discursului un spațiu al entităților aflate în centrul atenției. Star atenționaiă reprezintă o trăsătură dinamică a discursului, păstrâ **obiecte, proprietăți şi relații** ce sunt importante în fiecare moment interpretării discursului. Starea atenționaiă e modelată printr-un **set spații ale centrelor atenției,** în timp ce schimbările ce pot avea loc starea atenționaiă sunt restricționate de un set de reguli de tranzi care arată condițiile de adăugare şi ştergere a spațiilor. Colecția tutu spațiilor centrelor de atenție ce sunt disponibile în fiecare moment interpretării unui discurs formează o structură a atenției ce a dinamica unei **stive** şi care ar fi capabilă să explice proces implicate în interpretarea discursului, inclusiv accesibilitatea refer țială: domeniul în care trebuie căutate entitățile de discurs referite segmentul corespunzător stării atenționale aflate în vârful stivei e cel al stărilor aflate în stivă.

Structura recursivă a segmentului de discurs din AST permite şi acceptarea unei reprezentări arborescente, în cadrul căreia cele două relații în segmente, de dominare şi de satisfacere-precedență, nu sunt altceva de relațiile fireşti pe orice structură de arbore: cea dintre părinte şi orice fiu al său respectiv, cea de ordine dintre frați. AST se constituie, aşadar, într-o teorie glob asupra structurii şi a coeziunii discursului.

Cercetători precum Moser şi Moore [1996] sau Marcu [1999].pun evidență similarități semnificative între AST şi RST, inclusiv în ceea ce priv

maniera de reprezentare prin arbori a structurii de discurs, ceea ce permite combinarea puterii de reprezentare, mai fine în RST, datorită proliferării relațiilor, cu implicațiile pe care structura le poate avea asupra referențialității, puse în evidență de AST. Utilizând structura de segmente şi stiva, ca mecanism de prelucrare, AST propune o manieră de a rezolva accesibilitatea referinților anaforice printr-o transparență pe verticală, de sus în jos, de-a lungul stărilor atenționale ce se află la un moment dat în stivă. Reprezentarea prin segmente din AST are însă o slăbiciune: modelul stivă nu poate reflecta relația de dominare atunci când scopul dominat corespunde unui segment care apare în text înaintea celui care domină [Ide, Cristea, 2000]. Să remarcăm că defectul este unul de granularitate pentru că identificarea segmentului dominat ce precede pe cel dominator cu însuşi segmentul dominator elimină problema. AST nu e, aşadar, capabilă să reprezinte segmente având o granularitate oricât de fină: coborând de la o granularitate grosieră la una fină, există o limită dincolo de care ne putem aştepta la grave contradicții.

Teoria centrelor *(centering,* CT) [Grosz *et al.,* 1995; Brennan *et al.,* 1987] furnizează explicații convingătoare asupra restricțiilor de utilizare a pronumelor pentru realizarea referinților şi asupra ce anume face un discurs să fie coerent. CT nu se aplică însă dincolo de limitele unui segment (văzut în accepțiunea din AST). Avem de a face, aşadar, cu o teorie locală asupra coeziunii şi coerenței. Deşi nu este definită riguros în teorie, în toate exemplele autorilor unitatea elementară a structurii lingvistice este fraza *[utterance,* exprimare). Abordări ulterioare întrevăd posibilitatea de a considera o segmentare mai fină, la nivel de propoziție (v. [Kameyama, 1998] de exemplu). Noi vom considera drept **unitate** a structurii de discurs acelaşi tip de întindere lexicală ca şi în cazul RST, adică acea întindere ce la nivel sintactic este o propoziție iar la nivel semantic - o predicație. Fiecare unitate de discurs u_n ce intră în compoziția unui segment este caracterizată de o listă de **centre anticipatoare** *(forward-looking)* notată $Cf\{u_n\}$. Centrele listei $Cf(u_n)$ sunt entități semantice ce corespund, la nivelul textului, expresiilor referențiale cuprinse în unitatea u_n. Spunem că o expresie referențială **realizează** un centru. Elementele acestei liste sunt ordonate pentru a reflecta importanța relativă în u_n. Criteriile de ordonare a elementelor listei C_n sunt, în forma originară a teoriei, de natură sintactică, deşi alte abordări le diferențiază în funcție de limbă (v. de exemplu [Walker *et al.,* 1994] pentru japoneză, [de Eugenio, 1990; de Eugenio, 1998] pentru italiană, sau [Strube, Hahn, 1996] pentru germană). Pentru limba engleză autorii CT dau următorul criteriu: subiect > obiect-direct> obiect-indirect > complemente > adjuncți. Elementele listei $C^u_n)$ sunt acele entități despre care se vorbeşte în unitatea u_n şi deci despre care e cel mai probabil că se va continua să se vorbească şi în unitatea următoare, u_{n+1}, dacă aceasta aparține aceluiaşi segment ca şi u_n. Cel mai bine plasat element al listei $C\{u_n)$ se numeşte **centru principal** şi se notează $C_p\{u_n)$. Fiecărei unități îi este asociat un unic **centru retroactiv** *(backward-looking),* notat $C_b(u_n)$. Prin convenție, centrul retroactiv al

primei unități a segmentului este considerat centrul principal, în timp ce, pentru toate celelalte unități ale segmentului, el este cel mai bine plasat element al listei C **a** unității precedente care este de asemenea realizat şi în unitatea curentă.

Teoria face o clasificare a tranzițiilor posibile între unități consecutive, di punctul de vedere al invariantei ori nu a centrelor retroactive şi al identificării ori n a lor cu centrele principale. Astfel, cu excepția cazului în care între unită succesive ale aceluiaşi segment nu există centre comune, următoarele patru tipu de tranziții sunt posibile:

CONTINUARE *(continuing,* CON): $C_b\{u_{n+1}) = C_b\{u_n)$ şi $C_b(u_{n+1}) = C_p\{u_{n+1}$ corespunzând situației în care atât în u_n cât şi în u_{n+1} s vorbeşte despre aceeaşi entitate şi este de aşteptat ca şi î unitatea următoare să se vorbească despre ea.

REȚINERE *{retaining,* RET): $C_b(u_n+i) = C_b(u_n)$ dar $C_b\{u_{n+1}) \pm C_p\{u_{n+1}),$ cărui interpretare este că, deşi atât în u_n cât şi în $i/_{n+}$i s vorbeşte despre aceeaşi entitate, este de aşteptat ca unitatea următoare să se vorbească despre o alta.

SCHIMBARE LINĂ *(smooth-shifting,* SSH): $C_b(u_{n+1}) * C_b(u_n)$ dar $C_b(u_{n+1})$ $C_b\{u_{n+1}),$ cu semnificația că deşi în u_n şi în u_{n+1} nu s vorbeşte despre aceeaşi entitate este de aşteptat ca unitatea următoare să se vorbească despre entitat menționată ultima oară.

SCHIMBARE ABRUPTĂ *(abrupt-shifting,* ASH): $C_b(u_{n+1}) \pm C_b(u_n)$ şi $C_b(u$ $* C_p(u_n+i),$ cu semnificația că în u_n şi în u_n+i nu se vorbeş despre aceeaşi entitate şi este de aşteptat ca în unitat următoare să se vorbească despre o altă entitate dec ultima menționată.

Nucleul CT este concentrat în două reguli, prima enunțând o constrânge asupra formei de realizare a centrelor prin pronume, iar cea de a doua formulâ preferințe asupra secvențelor de tranziții ale centrelor. Regula a doua, cea care referă la coerență, formulează presupunerea că anumite secvențe produc încărcare inferențială în ascultător mai mare decât altele:

Regula 2: Secvențele de continuări sunt preferabile secvențelor rețineri, care sunt preferabile secvențelor de schimbări line, iar acestea s preferabile secvențelor de schimbări bruşte: CON > RET > SSH > ASH.

Dacă ne abținem de a penaliza CT, ca teorie locală, aşadar aplicabilă întinderea unui segment, pe motivul fragilității noțiunii de segment, care are definiție recursivă (un segment este constituit din alte segmente), slăbiciu moştenită de la AST, atunci apare naturală tentativa de a lărgi aplicabilitatea CT întregul discurs, într-o manieră recursivă, pe chiar această structură de segm definită, ea însăşi, recursiv. Teoria nervurilor propune o astfel de generalizare.

Teoria nervurilor *(veins theory,* VT) [Cristea *et al.,* 1998], preluând de la RST diferențierea dată de nuclearitate între argumentele relațiilor retorice dar ignorând, ca şi în AST, numele acestora, relevă o structură "ascunsă" în arborele de discurs, numită **nervură.** Fără a nega structura lingvistică a segmentelor de discurs, cât şi pe cea intenţională a relațiilor dintre scopurile comunicate de segmente şi care, prin echivalarea de care am amintit ([Moser, Moore, 1996; Marcu, 1999]), poate fi recuperată din structura de arbore proprie analizelor RST, VT corectează defectul de accesibilitate al AST înlocuind modelul accesibilității în stivă cu accesibilitatea de-a lungul nervurilor arborelui de discurs şi explicând naturaleţea unor referinţe la distanţă realizate prin mijloace de evocare foarte economice (pronume) [Fox, 1987]. Concluziile VT sunt, de asemenea, stabile la granularitate. în felul acesta VT se constituie într-o teorie globală a coeziunii discursului. VT generalizează totodată partea din CT relativă la încărcarea inferențială (regula a doua), extinzând concluziile ei la întregul discurs, prin aceasta VT constituindu-se şi într-o teorie globală a coerenței.

în secțiunea următoare sunt prezentate argumente lingvistice în favoarea teoriei. Secțiunea 3 prezintă definițiile teoriei, secțiunea 4 enunță conjectura VT relativă la referențialitate, iar secțiunea 5 - conjectura VT referitoare la coerență. Secțiunea 6 descrie rezultate experimentale în sprijinul presupozițiilor VT, secțiunea 7 prezintă o proprietate de granularitate, iar ultima secțiune este dedicată concluziilor şi prezentării unor aplicații ale VT.

2. Intuițiile VT

Noțiunea de nervură s-a născut sintetizând observațiile asupra modului în care se aliniază referințele pe o reprezentare arborescentă a discursului. Considerând organizarea ierarhică dată de structura de arbore şi principiul compoziționalității (v. de exemplu [Marcu, 2000]), care permite ca unități de discurs aflate la distanță să fie frați sub aceeaşi relație, aceste observații au fost următoarele (pentru simplificarea exprimării vom spune că "o unitate A referă o unitate B" şi vom înțelege "o expresie referențială aparținând unei unități A referă o entitate de discurs introdusă de (sau referită dintr-o) unitate B"; de asemenea vom nota cu u_1, u_2, u_3 - unități de discurs iar cu R, R_1, R_2 - relații. Atunci când apar ca argumente ale unei relații, unitățile de discurs vor purta un indice ridicat " sau ^{s}, cu semnificația de nucleu şi respectiv satelit):

- un satelit sau un nucleu poate referi un frate nuclear aflat la stânga: în combinații u'' R u_2, sau u'' R u_2, u_2 poate referi *uu*

Ex. 1

1. Ion a plecat de acasă fără umbrelă

2. deşi dimineață ***0*** *aflase la radio că va ploua.*

Subiectul vid (notat ***0)*** din unitatea 2, un satelit al unității 1 **[Ion]**[1] introdusă de expresia referențială *Ion* din prima unitate.

- un nucleu poate referi un satelit al său aflat la stânga: în co u_2 poate referi u-$_1$. Astfel, în exemplul:

Ex.2

1. Ion i-a dat Măriei o floare.

2. Pentru că ***0*** *s-a simțit frustrată,*

3. soția lui - s-a supărat.

unitatea 2 este un satelit al unității 3. Pe cine desemnează pronur din 2, pe **[Măria]** sau pe **[soția lui Ion]?** într-o interpretare increm la sfârşitul receptării celei de a doua unități avem tendința de timpuriu, subiectul vid **[Măriei]** apreciind totodată bizarul situaț unității 3 are loc însă o reconsiderare a legării ***0->*** **[Măria]** şi expresiei referențiale *soția lui* cu subiectul vid din 2, ambele ir **[soția lui Ion].**

un satelit dreapta al unui nucleu u nu e accesibil dintr-un nuclear sau satelit, al lui u: în combinații $\{u''$ $u_2)^n$ R_2 $u_3{}^n$ s $u_3{}^s$, u_3 poate referi u-\ dar nu u_2.

Ex.3

1. Ion i-a mărturisit Măriei că o iubeşte.

2. El n-&fost niciodată căsătorit

3. şi a trăit până la 40 de ani lângă mama sa.

4. Ea, dimpotrivă, a fost măritată de două ori.

Secvența 2-3-4 oferă o completare la 1. Secvența 2-3 se de CONTRAST (o relație paratactică) față de 4, iar 3 aduce o Structura este deci următoarea: u'' $((u_2$ R_2 $u_3{}^s)^n$ R_3 $u_4{}^n)^s$ în ca CONTRAST. Pentru cei mai mulți cititori, *ea* din unitatea 4 trebuie nu **[mama lui Ion],** deşi **[mama lui Ion]** este entitatea cea mai r poziția unității 4, cu care pronumele feminin se potriveşte în num preferării Măriei în locul mamei este acela că cititorul recunoaşte într-o relație de CONTRAST cu unitatea 2 (relație pusă *dimpotrivă),* ceea ce face ca cele două unități să fie percepute c Apropierea lor nu este însă una liniară, ci ierarhică, pe structură închisă la referință din unitatea 4.

[1] Vom nota prin [text] entitatea de discurs introdusă/referită de expresia n

un nucleu blochează accesibilitatea dintr-un satelit dreapta spre un satelit stânga: în combinații *(uf* R_1 $u_2^n)^n$ R_2 u_3^s, u_3 poate referi u_2 dar nu u_1.

Ex.4

1. *încă înainte cu un an de terminarea mandatului său de preşedinte al firmei*
2. *dl. W. Ross începuse maşinațiile pentru falimentarea acesteia.*
3. *De altfel, circulau vorbe că l-ar fi obținut fraudulos.*

în acest exemplu 1 şi 3 sunt sateliți ai lui 2 (1 este o circumstanțială a lui 2, în timp ce unitatea 3 dă o explicație la purtarea necinstită a lui Rossj. Referința **/=[mandatul lui Ross de preşedinte al firmei]** se deduce cu dificultate, ceea ce face ca întregul discurs să fie defectuos. Dimpotrivă, în următoarea variantă, discursul câştigă în cursivitate:

Ex.5

1. *Dl. W. Ross începuse maşinațiile pentru falimentarea firmei al cărei preşedinte era*
2. *încă înainte cu un an de terminarea mandatului său.*
3. *De altfel, circulau vorbe că l-ar fi obținut fraudulos.*

In Ex. 5 unitatea 2 este un satelit al lui 1, iar 3 - un satelit al lui 2 (aici *de altfel* anunță o parantează la informația asupra mandatului de preşedinte). Referința **/=[mandatul lui Ross de preşedinte al firmei]** poate fi recuperată acum fără dificultate.

Motivația acceptării Ex. 5 şi rejectării Ex. 4, constă nu în depărtarea liniară mai mare a anaforului de antecedent în Ex. 4 decât în Ex. 5, ci în faptul că în Ex. 4, spre deosebire de Ex. 5, accesul anafor-antecedent se face dinspre un satelit către un alt satelit, între ei interpunându-se un nucleu. Să remarcăm, de asemenea, că Ex. 4 poate fi reparat şi dacă se elimină această referință:

Ex. 6

1. *încă înainte cu un an de terminarea mandatului său de preşedinte al firmei*
2. *dl. W. Ross începuse maşinațiile pentru falimentarea acesteia.*
3. *De altfel, circulau vorbe că el ar fi fraudat alegerile.*

3. Definițiile teoriei

Intuiția fundamentală care stă la baza dezvoltărilor unificatoare asupra structurii de discurs şi accesibilității în VT este că distincția specifică RST dintre nuclee şi sateliți constrânge plaja de antecedenți asupra căr anaforii[1]; cu alte cuvinte, distincția nucleu-satelit, corelată cu o st induce pentru fiecare unitate de discurs un domeniu de accesi imediată pentru anaforii pe care-i conține. Mai precis, VT ava pentru fiecare anafor x aparținând unei unități de discurs *u,* x p uşurință examinând doar un subset al mulțimii entităților de disc Dacă antecedentul lui x este plasat într-o unitate de discu domeniului lui *u* atunci legătura anafor-antecedet este refăcută pentru realizarea ei e nevoie de mijloace referențiale tari, cum numele proprii.

Mai mult decât atât, aceeaşi corelație nuclearitate- întregului discurs, permite generalizarea CT dincolo de granit ceea ce face posibilă aplicarea concluziilor CT asupra coerenței

VT se bazează, în mare măsură, pe aceleaşi elemen discurs ca şi RST:

- unitățile de bază ale discursului sunt întinderi de tex *span)* ce nu se intersectează. După cum am preciz vom asimila cu propoziții, la nivel semantic fie predicație (căreia îi corespunde o reprezentare situațională);
- structura unui discurs este reprezentată ca un arbo de RST, dar fără a reduce generalitatea, în VT vo de discurs ca fiind binari (fiecare nod al arborelui are denți) (pentru argumentație, v. [Marcu, 2000] şi 1997]);
- principiul secvențialității [Cristea, Webber, 1997]: de pe frontiera terminală a arborelui corespunde s de discurs ce compune textul[2];
- principiul compoziționalității [Marcu, 2000]: o relați două întinderi de text se aplică, de asemenea, şi nucleare ale întinderilor aflate în relație;
- la fel ca în RST, nuclearitatea nodurilor arborel nodurile fiind clasificate în nuclee (cele mai importa mai puțin importante);

[1] într-o relație anaforică, interpretarea anaforului depinde de c antecedentul fiind plasat în text înaintea anaforului.

[2] Unitățile de discurs întrerupte nuanțează acest principiu. Astfel în următorul: O *datări când treceau unul pe lângă altul pe coridor?! e piezişă*[1] *1 care parcă-l străpunsese*[2]*! şi pentru o clipă fusese cuprins* (G. Orwell, *1984),* unitatea 1 este întreruptă de unitatea 2.

- nodurile terminale ale arborelui reprezintă unități de discurs, în timp ce nodurile neterminale reprezintă relații retorice între întinderi adiacente de text. Spre deosebire de RST, în VT nu interesează numele relațiilor, ceea ce contează fiind topologia arborelui, nuclearitatea nodurilor şi etichetarea nodurilor terminale;
- între fiii fiecărui nod intermediar al arborelui există cel puțin un nod nuclear. Nodul rădăcină, prin convenție, e considerat satelit.

în vizualizarea arborilor vom reprezenta nodurile neterminale prin dreptunghiuri fără nume, pe cele terminale - prin ovaluri etichetate, iar nodurile nucleare vor fi subliniate (v. Figura 1). în definițiile ce urmează vom folosi următoarele convenții de notare:

- *mark(a)* este o funcție care întoarce şirul a în care fiecare simbol este marcat (de exemplu, este poziționat între paranteze);
- *unmark(d)* este funcția inversă lui *markQ,* ce îndepărtează toate marcajele ataşate simbolurilor din expresia *a* (ex. *unmark(mark(a)) = a);*
- *simpl(a)* este funcția care elimină toate simbolurile marcate din expresia argumentului *a* (ex. *simpl(mark(a))* = o, şirul vid, şi *simpl(cx • mark(p) •* y)) = *a - y)*\
- *seq(a, (3)* este o funcție de secvențiere, care întoarce acea permutare a concatenării simbolurilor din *a* şi j3 dată de citirea de la stânga la dreapta a nodurilor corespunzătoare simbolurilor din *a* şi /3 pe frontiera terminală a arborelui. Funcția menține marcajele asupra simbolurilor, dacă acestea există, *seq(0, fi)* = /?; şi *seq(a, seq(P)) = seq(seq(a),* j8) = seg(a, 0);
- *H(n)* şi *V(n)* reprezintă expresiile *head* şi nervură (în engleză - *vein)* ale unui nod n;
- *pref(u, a)* reține prefixul expresiei simbolice *a* până la simbolul *u* inclusiv, o etichetă de nod terminal.

Teoria nervurilor calculează două expresii, pe care le ataşează fiecărui nod al structurii.

3.1 Expresia *head* a unui nod al arborelui

Intenția expresiei *head* a unui nod al arborelui de discurs este de a exprima secvența celor mai importante unități de discurs din întinderea de text acoperită de nod. Ea este o secvență de etichete de unități, calculată după cum urmează:

Definiții

Expresia *head* a unui nod terminal este însăşi eticheta s

2. Expresia *head* a unui nod neterminal este dată de c ordinea apariției lor în arbore de la stânga la dreapt *head ale* descendenților săi nucleari.

Definițiile expresiilor *head* sugerează un proces de calcul de jos în sus în arborele de discurs. Cele mai importante unități proiectate în sus până în primul nod satelit întâlnit.

direcția a calcul expresii

Figura 1: Calculul expresiilor *head*

3.2 Expresia nervurii unui nod al arborelu

Expresia nervurii unui nod intenționează să surprindă se **de** discurs care sunt semnificative pentru a sintetiza[1], în con text, întinderea de text acoperită **de** nod. Pentru orice nod al st nervurii este formată din cele mai importante unități din întinde nod, împreună, eventual, cu alte unități din afara acestei întinderi.

Prin sinteza, sau rezumatul, unei întinderi de text se înțelege un text ideea principală a textului supus sintezei. Indiferent dacă este realizat p prin punerea cap la cap a unor subsecvente ale întinderii originale rezumat trebuie să fie comprehensibil, adică trebuie să poată fi în (printre altele, de exemplu, rezumatul trebuie să conțină toate element rezolvarea anaforilor). Adesea însă, atunci când întinderea este decup mai larg, pentru ca rezumatul să fie comprehensibil, el trebuie să con afara întinderii şi care aparțin contextului. Avem de a face, în acest unei întinderi de text în contextul unei întinderi mai vaste. Să mai obs privințe, "a sintetiza" e analog cu "a înțelege", pentru că ceea ce ne r unui text este o sinteză a lui.

Definițiile care urmează, datorită recursivității lor, vor face posibilă considerarea contextului dat de totalitatea textului din exprimarea "a înțelege, în contextul întregului text, întinderea s" mărginit la întinderea de text acoperită de nodul părinte al celui corespunzător întinderii s. Cu alte cuvinte, la fiecare nivel al structurii, cu excepția rădăcinii, adică întotdeauna unde există două noduri fii sub un nod părinte, cu întinderile celor două noduri fii însumând întinderea nodului părinte, expresia nervură a părintelui conține deja informația care permite înțelegerea/rezumarea întinderii acoperite de el în contextul global. Coborârea pentru înțelegerea/rezumarea subîntinderii acoperite de nodul curent al definiției (unul dintre cele două noduri fii) înseamnă adăugarea şi/sau ştergerea unei secvențe noi/subsecvențe la/din secvența de etichete contribuită de nervura părintelui, în funcție de polaritatea şi poziția specifică a întinderii corespunzătoare nodului fiu curent în întinderea nodului părinte. în continuare, întinderea întregului text, o constantă pentru orice subîntindere, va fi numită contextul total. în figurile 2-^-6, nodurile curente - cele vizate de definițiile curente de nervură - apar în gri. Ele sunt notate simultan cu un dreptunghi şi un oval pentru a sugera că pot fi atât noduri interioare (neterminale), cât şi noduri terminale.

Definiții

1. Expresia nervurii rădăcinii este egală cu expresia sa *head.*

Particularizând intenția pe care o exprimă expresia nervurii unui nod la modul rădăcină obținem: cele mai semnificative unități de discurs necesare înțelegerii/rezumării întinderii acoperite de nod (în cazul de față - întregul text) în contextul total. Cum contextul este aici egal cu textul în totalitatea lui, el poate fi lăsat la o parte în descriere, ceea ce ne lasă cu definiția expresiei *head* a nodului rădăcină.

2. Pentru fiecare nod nuclear, al cărui părinte are nervura *v:*

a. dacă nodul nu are un frate nenuclear în stânga, atunci expresia nervurii este *v(v.* Figura 2);

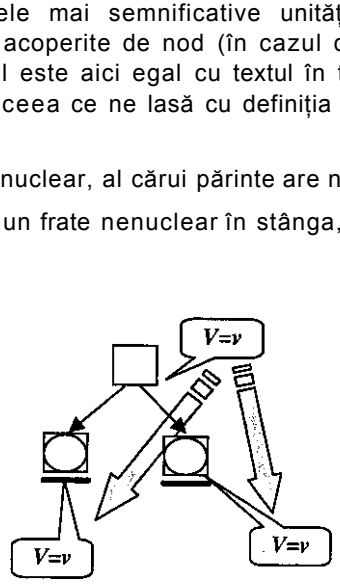


Figura **2:** Expresia nervurii unui nod nuclear fără frate satelit în stânga

Definiția exprimă faptul că secvența de unități necesară înțelegeri rezumării, în contextul total, a unei întinderi nucleare de text ce are ca frate î structură o altă întindere nucleară necesită aceeaşi secvență de unități ca şi ce necesară înțelegerii/rezumării, în contextul total, a reuniunii celor două întinderi. C alte cuvinte', o întindere nucleară ce este frate, în structură, întinderii nuclear curente este esențială înțelegerii/rezumării întinderii curente.

b. dacă nodul are un frate nenuclear în stânga de *head h,* atunci expresi nervurii lui este *seq(mark{h), v)* (v. Figura 3);

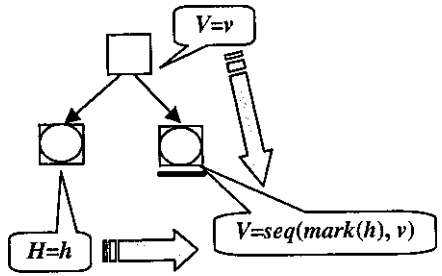


Figura **3:** Expresia nervurii unui nod nuclear având un frate satelit în stânga

Secvența de unități necesară înțelegerii/rezumării, în contextul total, a un întinderi nucleare de text ce are ca frate stânga în structură o întindere nenuclea necesită, suplimentar față de secvența necesară înțelegerii în contextul total întinderii acoperită de nodul părinte (comunicată de expresia nervură a nodu părinte) şi secvența *head* a întinderii frate stângi (adică cele mai importante unit din întinderea stângă). Considerarea, în expresia nervurii întinderii nuclea curente, a expresiei *head* a întinderii nenucleare frate stânga, corespunde, p prisma definiției 2a, cu atribuirea întinderii stângi a calității de a se comporta ca nucleu. Marcarea contribuției satelitului frate stânga prin funcția *markQ* face în această revizuire a nuclearității lui, una cu valoare temporară, după cum se dovedi mai jos, în definiția 3b.

3. Pentru fiecare nod nenuclear de *head h,* al cărui părinte are nervura *v:*

a. dacă nodul este descendentul stâng âl părintelui său, atunci expres nervurii sale este *seq(h, v)*\

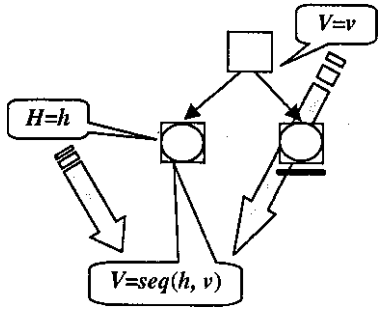


Figura 4: Expresia nervurii unui nod satelit stânga

Definiția exprimă faptul că pentru a înțelege/rezuma, în contextul total, o întindere nenucleară de text ce este descendent stâng, în structură, nodului părinte, la secvența de unități ce exprimă influența contextului total (precizată de expresia nervură a părintelui) trebuie adăugate cele mai importante unități din întinderea proprie (date de expresia *head* proprie). Să observăm că în expresia nervurii nodului părinte, care moşteneşte expresii *head* ale nodurilor superioare, nu poate răzbate influența unui fiu satelit al său, deci numai includerea *head-u\u* fiului satelit, direct în expresia nervurii sale poate completa această influență.

b. dacă nodul este descendentul drept al părintelui său, atunci expresia nervurii lui este *seq(h, simpl(v))*.

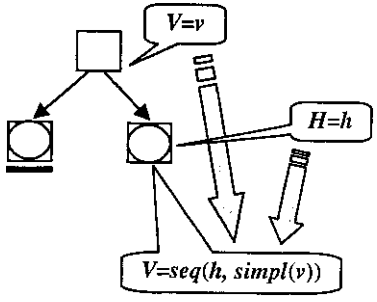


Figura 5: Expresia nervurii unui nod satelit drept

Pentru a înțelege, în contextul total, o întindere nenucleară de text ce este descendent pe dreapta al nodului părinte, la secvența de unități necesară înțelegerii/rezumării contextului total (precizată de expresia nervură a părintelui) şi

din care s-au şters unitățile marcate trebuie adăugate cele mai importante unit din întinderea proprie (date de expresia *head* proprie). în acest fel, dacă expres nervură a nodului părinte nu conține unități marcate (în conformitate cu defini 2b), atunci expresia nervură a unui satelit dreapta nu diferă de expresia nervură aceluiaşi satelit ce ar fi fost poziționat pe stânga (conform definiției 3a). Dacă în nervura părintelui conține unități marcate, atunci acestea dispar din expres nervurii satelitului drept. Conform definiției 2b, unitățile marcate pot fi datorate do unui satelit stânga, frate al celui mai apropiat ascendent nuclear ai întinde curente. Urmează că definiția curentă exprimă o proprietate de blocare accesibilității dinspre un satelit plasat în dreapta unui nucleu către un satelit plas în stânga sa (v. Figura 6).

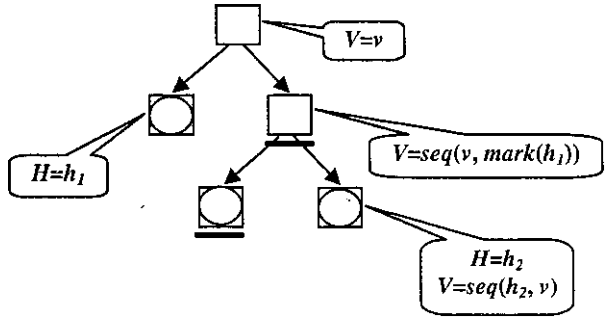


Figura 6: Simplificări în calculul expresiei nervură a unui satelit dreapta:

$V=seq(h_2,\ simpl(seq(v,\ mark_{\hat{}}thi)))\ =\ seq(h_2,\ seq(v))\ =\ seq(h_2,\ v)$

Dacă semnificația expresiei nervurii unui nod oarecare din structură es particularizată la un nod terminal, obținem: **expresia nervurii unei unități discurs reprezintă secvența unităților de discurs care sunt semnificati pentru a înțelege/rezuma, în contextul întregului text, însăşi unitatea discurs în cauză.** Printre altele, aceasta înseamnă că expresia nervurii unei unit de discurs este suficientă pentru a interpreta toate referințele anaforice conținute unitate.

4. Relația dintre structura de discurs si referentialitate

Ipoteza pe care o avansăm este că rezoluția anaforică este caracteriza de două tipuri de procese: **evocative** (sau **imediate)** şi **post-evocative** (sau **int renționale).** Procesele evocative, cele mai frecvente, sunt rapide şi pot fi realiza prin orice mijloace de evocare referențială, inclusiv cele fragile (de tipul subiecte

vide şi pronumelor). Ele dau textului fluenţă şi-l fac coeziv. Cele post-evocative sunt mult mai puţin frecvente decât cele evocative, necesită o încărcare inferenţială mai mare pentru a fi interpretate şi utilizează mijloace referenţiale tari (nume proprii, substantive comune articulate).

Vom asocia spaţiul de căutare al proceselor evocative unui **domeniu de accesibilitate referenţială evocativă** sau **imediată** *(domain of evocative accessibility - dea)* pe baza definiţiei nervurii şi a următoarelor observaţii:

- **relaţia anaforică este de natură semantică, iar nu textuală** [Halliday, Hassan, 1976]: o relaţie anaforică are doi termeni: anaforul şi antecedentul. Anaforul este reprezentat de o expresie referenţială a cărei natură este textuală. Natura semantică a relaţiei anaforice trebuie înţeleasă ca răsfrângându-se asupra antecedentului, care nu trebuie identificat cu o anumită expresie referenţială ce precede în text anaforul, ci cu o reprezentare a acesteia într-un plan semantic, în aşa fel încât semnificaţia anaforului se construieşte din antecedentul însuşi iar nu din semnificaţia lui. în cazul particular al unui lanţ co-referenţial, acest lucru înseamnă că antecedentul este "realizat" repetat în text în aceeaşi entitate de discurs. Expresiile co-referenţiale "ancorează", în diverse poziţii ale textului, o aceeaşi entitate de discurs.
- **dinamica interpretării discursului este incrementală:** un discurs este un text în procesul citirii ori ascultării lui de către un subiect (om sau maşină). Când citirea/ascultarea unui text s-a terminat, discursul este încheiat şi ceea ce rămâne este o reprezentare a lui în memoria subiectului. De asemenea, la un moment dat pe parcursul interpretării unui text, anumite elemente ale discursului pot fi plasate privilegiat în sfera atenţiei [Grosz, Sidner, 1986; Sidner, 1983; Walker, 1996], iar trecerea de la o unitate de discurs la următoarea poate produce schimbări în structura memorată ce configurează sfera atenţionaiă.
- **anafora şi a catafora au o natură cognitivă comună:** din punct de vedere cognitiv, toate referinţele anaforice se fac dinspre expresii referenţiale (entităţi textuale) către entităţi ale discursului (entităţi semantice) deja introduse de discursul trecut. Acest lucru înseamnă că, într-o limbă în care textul se notează de la stânga spre dreapta, nu există jeferinţe anaforice spre dreapta. Distincţia dintre anafora şi cataforă, devine, în această viziune care încearcă să reconstituie procesele cognitive ce stau la baza înţelegerii textelor (cu sau fără scopul simulării lor pe maşină), inutilă. în aceeaşi manieră în care, în cazul unei anafore, un antecedent este o entitate de discurs propusă de o expresie referenţială ce precede anaforul şi pe care anaforul o referă apoi, pronumele ce precede un nume, în cazul unei catafore, propune o reprezentare, mai săracă, pe care numele o referă şi o completează în acelaşi timp [Cristea, Dima, 2001]. Ac interpretării discursului o unică direcţionalitate, care timpului lecturii, şi care este cea a desfăşurării liniare limbile europene, de exemplu, de la stânga la drea referenţialitate trebuie deci să se proiecteze pe acea anafori "noi" către entităţi "vechi", mereu către înapo lecturii.

Ex.7

1. Pentru că 0 n-a *vrut să-şi lase tata singur,*

2. *Ion a renunţat la concediu.*

Expresia referenţială vidă de pe poziţia de subiect a uni propune o entitate de discurs caracterizată cel mult de o descri **number** singular] (ce poate fi atribuită, cel mai probabil, unor sur de natură pragmatică: cineva care nu poate să-şi lase tatăl sing persoană, corelate cu surse de natură sintactică: acordul în num Apoi, substantivul propriu /on, din unitatea 2, referă entitatea co şi o completează până la o reprezentare: **[type** human, **numbe** Ion].

Corelarea definiţiei nervurii cu observaţiile de mai s configurarea domeniului de accesibilitate referenţială evocativă toate unităţile de discurs care preced unitatea în care se referenţială (şi din care au fost îndepărtate eventualele marcaje, rol de memorie temporară):

dea(u) = pref(u, unmark(V(u)).

Definiţia *dea* formalizează prima conjectură a VT (sau pune în legătură accesibilitatea referenţială imediată cu stru **antecedenţii expresiilor referenţiale dintr-o unitate de discurs precădere, printre entităţile de discurs ancorate în unităţile inclusiv u, în expresia nervurii acesteia.**

Paul Cornea [1998] vorbeşte despre recodificarea se rizarea. El pune în evidenţă trei tipuri de memorie, ce apar, de cercetători [Kinntsch, Vârî Dijk, 1975; Schank, Abelson, 197 memoria imediată, memoria de scurtă durată (de termen scurt 'ungă durată (de termen lung - MLT). Memoria imediată este u senzorial al informaţiilor, reţinerea urmelor din ultima jumătate conservă câteva secunde informaţia. Lungimea acestei memor semne (cuvinte, cifre, litere - funcţie de context, v. şi [Miller, 19

apreciază acest "empan" mijlociu la 13^-15 cuvinte, la un lector lent fiind de 8 cuvinte, la unul rapid - de 16-**5**-20 [Richadeau, 1969] - citat în [Cornea, 1998] p. 166).

Construcția structurii de discurs se face dinamic, în actul lecturii. Să ignorăm un posibil proces de multi-interpretare ce poate duce la sintetizarea simultană a mai multor construcții alternative, din care să se selecteze, în urma unui proces de dezambiguizare, una sau mai multe structuri arborescente finale. Arborele însuşi poate fi considerat rezumat în diverse grade, conform capacității de memorare a subiectului. Dacă unitatea curentă este u_n, să notăm AR_n arborele de structură rezumat, la momentul prelucrării unității u_n. Nervura acesteia, culeasă pe AR_n, este $V(u_n)$, iar domeniul ei de accesibilitate imediată $dea(u_n)$. Noi credem că MST poate fi considerată o fereastră de lungime 7±2 semne în directă legătură cu $dea(u_n)$: fie 7±2 unități din această secvență, fie tot atâtea structuri evenimențiale - ca reprezentări ale unităților de discurs, fie încă numai simboluri (cuvinte etc.) culese din acest şir de unități. Tranzitarea la următoarea unitate, u_{n+1}, înseamnă înlocuirea memoriei de scurtă durată $dea(u_n)$ cu $dea\{u_{n+1}\}$. Acest lucru duce uneori la o simplă prelungire a domeniului de accesibilitate precedent, alteori la o alterare a lui prin ştergerea unor unități şi adăugarea altora, de fiecare dată domeniul încheindu-se cu unitatea curentă. MST este aşadar o proiecție a unui şir de unități de discurs (sau de microstructuri ce-şi au suportul în aceste unități) decupate din structura dinamică curentă. Modificările ce apar în şirul MST reflectă schimbările de focalizare, în parcurgerea discursului. Componența acestui şir este influențată de uitare (deci de un proces de abstractizare) şi de modificarea de interes curentă în parcurgerea discursului. Când interesul s-a mutat pe o altă axă, componența nervurii şi, de aici, a domeniului de accesibilitate imediată sunt şi ele actualizate. Includerea sau excluderea din MST a unor unități de discurs în ritmul citirii, pentru că *dea* evoluează eliminând unele unități şi "redeşteptând" altele "uitate", amintesc de procesele de "chemare" în sfera atenției ale memoriei *cash* a lui Walker [Walker, 1996]. Pe de altă parte, structura memorată (rezumată) a discursului este păstrată în MLT şi folosită pentru aducerea în prim plan a unităților de interes curent ce au fost temporar retrogradate de o comutare a atenției într-o altă direcție. Procesele evocative se desfăşoară aşadar în memoria de scurtă durată. Pe de altă parte, procesele post-evocative sunt procese de rezoluție anaforică de natură inferențială, ce presupun un anumit efort de regăsire a unei entități de discurs într-o zonă a memoriei de lungă durată sau evocă entități ale cunoaşterii generice din sfera culturală a subiectului. Noi credem că aceste procese se dezvoltă tot pe structura de discurs dezvoltată deja, ieşind din *dea,* când rezoluția a eşuat acolo.

Dintr-un punct de vedere ce se concentrează asupra relației dintre referențialitate şi structura de discurs, celor două tipuri de procese anaforice pe care le-am pus în evidență le corespund **referințe evocative,** respectiv **post-evocative** (sau **inferențiale).** Diferența dintre ele este că, în cazul primelor, lanțul retroactiv al unităților ce ancorează expresii aflate în relații referențiale

intersectează domeniul de accesibilitate referențială imediată al unității anaforului în cel puțin încă un punct decât unitatea anaforului, pe când în cazul referințelo post-evocative nu există această intersecție dublă. în [Cristea *et al.,* 2000; Cristea 2000] referințele evocative sunt, mai departe, detaliate în **directe** şi **indirecte.**

în referințele directe, a doua unitate de intersecție este unitatea cea ma recentă liniar ce ancorează aceeaşi entitate de discurs ca şi anaforul (în cazu relației de co-referință) sau o entitate corelată funcțional cu aceasta (în cazul une relații de referință funcțională). în referințele indirecte intersecția *dea* cu lanțu co/func-referențial se realizează într-o unitate mai depărtată decât cea mai recenta liniar de unitatea anaforului. în referințele inferențiale lanțul retroactiv al legăturilo anaforice ale anaforului nu intersectează *dea* (în Figura 7 lanțul legăturilo anaforice este reprezentat punctat, iar *dea* printr-o linie groasă).

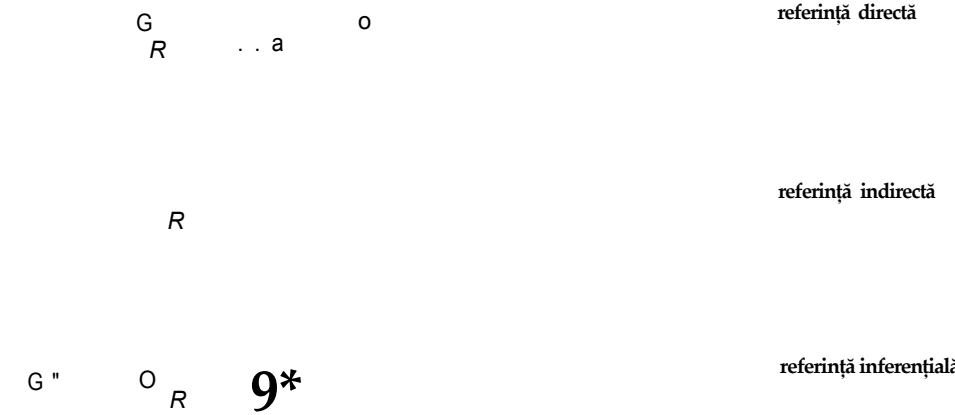


Figura 7: Referințe evocative şi post-evocative

O categorie particulară de referințe post-evocative sunt **referințel pragmatice** (ce pot fi numite şi **pseudo-referințe).** în acest tip de referinț participă expresii referențiale care pot fi interpretate fără un antecedent pentru c interpretarea lor se bazează pe cunoştințe exterioare textului, ce vin di cunoaşterea comună asupra lumii, deci din pragmatică. Deşi există cel puțin încă expresie referențială în text ce realizează aceeaşi entitate de discurs, expresii referențiale pot să nu aibă, în mod necesar, o reprezentare unică, fără ca pr acesta înțelegerea textului să sufere.

Recunoaşterea antecedentului se datorează, în toate cazurile, unor procese de confruntare de şabloane *(pattern-matching)* îmbogăţite cu euristici, în care intervin structura de caracteristici morfo/sintactico/semantice ce definesc anaforul şi structurile de caracteristici ce definesc entităţile de discurs deja introduse [Cristea, Dima, 2001; Cristea *et* a/., 2002a].

5. Relaţia dintre structura de discurs şi cursivitate

5.1. Linii de argumentaţie

Expresiile nervură ale unităţilor ce compun un discurs arată tot atâtea moduri diferite în care poate fi citit acel discurs. Fiecare în parte dă o rezumare a discursului prin prisma unităţii de discurs curente. Atunci când interesul este orientat către un anumit episod al povestirii, putem sări peste pasaje întregi pentru a ne concentra asupra manierei în care elementul de interes se leagă cu ansamblul discursului. în acelaşi fel, putem avea în vedere o altă pistă şi atunci lectura focalizează un alt fir de interes. Acest nou fir poate să aibă elemente în comun cu primul dar poate, de asemenea, să incorporeze şi altele noi. Fiecare fir în parte poate pune în evidenţă anumite particularităţi, legate însă strâns de linia principală a discursului. Toate aceste sub-discursuri sunt coerente şi, în general, nu există referinţe anaforice pentru a căror interpretare să avem nevoie de fragmente aflate în afara rezumatului însuşi. Acest lucru înseamnă că traseele referenţiale ale rezumatului conţin suficiente elemente care să ducă la recuperarea înţelesului anaforilor.

Să luăm următorul text:

Ex. 8

1. *Piton primise-n taină poruncă de la Hera să-l pîndească pe Apolo,*
2. *cînd va trece prin munte,*
3. *şi să-i răpună viaţa.*
4. *Hera-I ura pe fiul cel nou născut al Letei,*
5. *pentru că soţul său, prea puternicul Zeus, ţinea mai mult la dînsul decit la fiii ei: Hefaistos şi Ares.*
6. *Cînd a ajuns Apolo în muntele Parnas,*
7. *dihania uriaşă* s-a *avîntat spre dînsul,*
8. *dornică* să-/ *ucidă.*
9. *Dar zeul şi-a întins arcul.*
10. *A tras prima săgeată.*
11. *Erau doar patru zile de cînd văzuse lumea,*
12. *şi întiia lui săgeată a şi nimerit monstrul.*

Afexandru Mitru - *Legendele Olimpului,* Editura Tineretului, 1966

Structura de discurs a acestui text este cea din Figura 8. Tabela 1 dă expresiile nervură şi domeniile de referenţialitate evocativă ale nodurilor terminale, în coloana *dea(u)* au fost, totodată, marcate în aldine domenii de referenţialitate imediată maximale vis-â-vis de relaţia de incluziune (cele mai lungi trasee *dea).* Astfel dea(1) c dea(2) *<z dea{3)* c ofea(4) c *dea{5) <t dea(6)* ş.a.m.d. Vom numi aceste secvenţe care întrerup lanţuri de incluziuni linii de argumentaţie (/a), în cazul nostru: 1 2, 1 3 4 5, 1 3 6 7, î 3 7 8 şi 1 3 7 9 10 11 12. Dacă /a(u,) precede imediat /a(u_2), atunci în *la{u_2)* se regăsesc domeniile tuturor unităţilor dintre u_y+1 şi u_2. în particular, în *la{u_2)* se regăsesc unităţile ce preced imediat unitatea u, pentru orice *u* între i/$_7$+1 şi u_2, în domeniui lor de accesibilitate imediată (adică acel domeniu care conferă discursului maximum de coerenţă). Cu alte cuvinte, pe *la(u_2)* putem aplica definiţiile CT de calculare a tranziţiilor pentru orice *u* între i/$_1$+1 şi u_2.

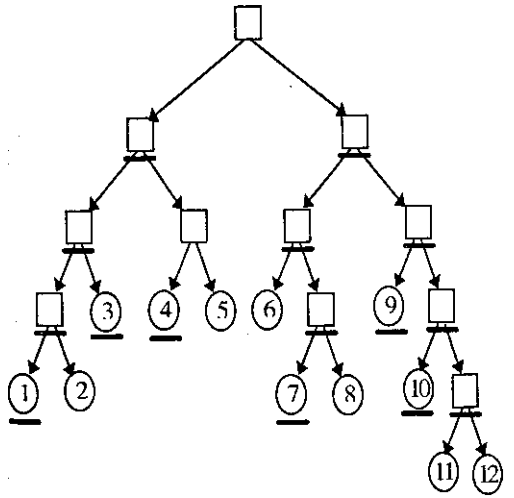


Figura 8: Structura de discurs a Ex. 8

Tabela 1: Nervurile şi domeniile unităţilor din Ex. 8

u	*V(u)*	*dea(u)*
1	1 3 7 9 10 12	1
2	1 2 3 7 9 10 12	**1 2**
3	1 3 7 9 10 12	1 3
4	1 3 4 7 9 10 12	1 34
5	1 3 4 5 7 9 10 12	**13** 4 **5**
6	1 3 6 7 9 10 12	1 36
7	1 3(6)7 9 10 12	**1 367**
8	1 3 7 8 9 10 12	**1 378**
9.	1 3 7 9 10 12	1 3 7 9
10	1 3 7 9 10 12	1 3 7 9 10
11	1 3 7 9 10 11 12	1 3 7 9 10 11
12	1 3 7 9 10 (11) 12	**1 379 10 11 12**

5.2. O generalizare a CT

Urmând recomandările teoriei centrelor, să presupunem că marcăm tranziţiile ce apar între unităţi de discurs cu scoruri care să dea un grad al uşurinţei de prelucrare:

CONTINUARE	(CON)	4
REŢINERE	(RET)	3
SCHIMBARE LINĂ	(SSH)	2
SCHIMBARE ABRUPTĂ	(ASH)	1
LIPSĂ Cb	(-)	0

în felul acesta, tranzițiile line primesc scoruri mari, cele abrupte, scoruri mici. însumând aceste scoruri pentru fiecare unitate a unui segment (segment, în spiritul AST) vom avea un scor al segmentului. Să notăm un scor în spiritul CT al unui segment s cu S_cCT' (CCT de la *Classical Centering Theory).* El ne va da o măsură a uşurinței de interpretare a segmentului: cu cât un segment s, în totalitatea lui, e mai fluent, cu atât scorul lui va fi mai mare şi cu cât el este mai abrupt, mai dificil de interpretat, cu atât scorul lui va fi mai scăzut. în fine, să adunăm aceste scoruri pentru toate segmentele discursului, într-un scor al sumei segmentelor $S_c$crp

$$S_{CCT}- \quad {}_{s}CCT$$

Să ne imaginăm acuma că forțăm nota şi calculăm aceste scoruri şi dincolo de granițele de segment, deci inclusiv în punctele de frontieră dintre segmente. Să notăm acest scor global cu S_cCT'- în scorul global S_cCT' contribuie cu scoruri de tranziții toate unitățile cuprinse între a doua unitate şi ultima. în mod normal tranzițiile în punctele de trecere între segmente ar trebui să fie foarte abrupte, cotate deci slab ori zero, şi deci scorul global ataşat textului n-ar trebui să fie modificat semnificativ. Dacă apare totuşi o diferență, ea trebuie să fie datorată unor tranziții accidentale peste granița de segment. în orice caz trebuie să avem $SCCT^{c} \wedge SCCT$-

Să procedăm acum în mod analog, ca suport folosind de data aceasta liniile de argumentație iar nu secvențele liniare de unități ale segmentelor în sensul clasic. Datorită comportamentului lor similar segmentelor, putem numi liniile de argumentație **segmente în sens ierarhic.** Să notăm $S_{H}CT'$ (HCT de la *Hierarchical Centering Theory)* suma scorurilor unităților aparținând unei linii de argumentație (segment ierarhic) s. Ca să dăm o măsură a fluenței discursului în accepțiunea ierarhică, similară scorului global S_cCT' în calculul scorului global al discursului în sens ierarhic nu va trebui să repetăm contribuțiile unităților ce apar în mai mult decât o singură linie de argumentație. Dacă notăm $S_{H}CT'$ scorul unui segment ierarhic s' în care am păstrat numai unitățile noi față de segmentul anterior, atunci scorul global ierarhic al discursului este:

$$_{>HCT}{}^{G} = I > ; {}_{HCT}$$

Cea de a doua conjectură a VT (a coerenței): Scorul global în sensul ierarhic al unui discurs este mai bun sau cel puțin egal decât scorul global în sensul clasic: $S_{H}CT^{G} \wedge S_cCT^{G}$.

Pentru un anumit detaliu de granularitate în definirea segmentelor în sens clasic, unui segment în sens clasic îi corespunde o secvență de nervură, deci o porțiune a unei linii de argumentație. în spiritul acestei observații, ce conjectură enunță prezumția că tranzițiile la distanță lungă, calcula nervurilor, sunt sistematic mai line decât tranzițiile accidentale la gr segmente. Să notăm că această presupoziție este conformă unor obs de autori precum Passonneau [1995] şi Walker [1998], furnizând explicație pentru rezultatele lor.

în cele ce urmează prezentăm o analiză comparativă clasic- probează ipoteza coerenței, pe discursul din Ex. 8.

Tabela 2: Analiza Ex. 1 în maniera CCT

n	u_n	$C_f(u_n)$	$C_b(u_n$
1	*Piton primise-n taină poruncă de la Hera să-l pîndească pe Apolo,*	**[Piton], [Hera], [Apolo],**	**[Piton**
	cînd 0 va trece prin munte,	*Ø* = **[Apolo], [munte]**	**[Apol**
	şi 0 să-i răpună viața.	o= **[Piton], KApolo], [viața]**	**[Apol**
	\Hera-l ura pe fiul cel nou născut al \Letei,	**[Hera], [Leta],** *fiul cei nou-născut al* **LeteHApolo]**	**[Apo**
	pentru că soțul său, prea puternicul Zeus, ținea mai mult la dînsul decit la fiii ei: Hefaistos şi Ares.	**Zeus], său=[Hera], dmsu/=[Apolo], [Hefaistos], [Ares]**	**[Her**
	\Cînd a ajuns Apolo în muntele Parnas,	**[Apolo], [munte]**	**Apo**
	\dihania uriaşă s-a avîntat spre idînsul,	*dihania* **ur/aşă=[Piton],**	**[Apo**
	\0 (era) dornică să-l ucidă.	**Id/nsuHApolo]**	**[Pit**
	par zeul şi-a întins arcul.	**\o= [Pitoni, I=[Apo»o]**	**[Ap**
10	*0A tras prima săgeată.*	*\zeul* **= [Apolol [arcul]**	**[Ap**
11	*Erau doar patru zile de cînd 0 văzuse lumea,*	**o = [Apolo], [săgeata]**	**[Ap**
12	*\şi întiia lui săgeată a şi nimerit \monstrul.*	**io = [Apolo], [lumea]**	
		[[săgeata] /u/=[Apolo], monsfriv/=[Piton]	**[Ap**

Snzijia da 29/11=2.63 ceea ce înseamnă că tortul, conform a

comportă, în medie, intermediar între o schimbare lină (SSH) şi o reținere (RET), mai apropiat de o reținere.

Dacă luăm în calcul liniile de argumentație indicate de nervuri, pot fi puse în evidență 5 sub-discursuri, în lungul cărora vom calcula, de asemenea, tranzițiile, în tabelele 3^7 de mai jos unitățile pentru care considerăm tranzițiile sunt, de asemenea, indicate în caractere aldine în prima coloană. Să remarcăm că citirea textelor date de liniile de argumentație produce, în toate cazurile, discursuri perfect coerente. în ansamblu, doar câte o tranziție este calculată pentru fiecare unitate, la fel ca şi în interpretarea clasică.

Tabela 3: Analiza HCT a primei linii de argumentație, secvența de unități 1-2

n		*0{u$_n$)*	*C$_b$(u$_n$)*	Traz.	Scor
1	*Piton primise-n taină poruncă de la Hera să-l pîndească pe Apolo,*	**[Piton], [Hera], [Apolo],**	**[Piton]**	-	-
2	*cînd 0 va trece prin munte,*	o **= [Apolo], [munte]**	**[Apolo]**	SSH	2
				Total	**2**

Tabela 4: Analiza HCT a celei de a doua linii de argumentație, secvența de unități 1-3-4-5

n	*u$_n$*		*C$_b$(u$_n$)*	Traz.	Scor
1	*Piton primise-n taină poruncă de la Hera să-l pîndească pe Apolo,*	**[Piton], [Hera], [Apolo],**	**[Piton]**	-	-
3	*şi 0 să-i răpună viața.*	o= **[Piton], /=[Apolo], [viața]**	**[Piton]**	CON	4
4	*Hera-l ura pe fiul cel nou născut al Letei,*	**[Hera], [Leta],** *fiul cel nou-născut al* Lete**/=[Apolo]**	**[Apolo]**	ASH	1
5	*pentru că soțul său, prea puternicul Zeus, ținea mai mult la dînsul decit la fiii ei: Hefaistos şi Ares.*	**[Zeus],** său**=[Hera],** d**/>?si//=[Apolo], [Hefaistos], [Ares]**	**[Hera]**	ASH	1
				Total	**6**

Se constată că tranziția RET a unității 3 către 2 din analiza CCT s-a transformat într-o tranziție CON, pe nervură, dinspre 3 către 1 deşi tranziția RET dinspre 3 spre 4 în CCT devine aici o tranziție ASH, deci mai abruptă, datorită modificării Cb-ului unității 3 din **[Apolo]** în **[Piton].**

Tabela 5- Analiza HCT a celei de a treia linii de argumentație, secv
1-3-6-7

n	*u$_n$*	C/(Un)	*C$_b$(u*
1	*Piton primise-n taină poruncă de la Hera* să-*l pîndească pe Apolo,*	**[Piton], [Hera], [Apolo],**	**[Pito**
3	*şi 0 să-i răpună viața.*	o= **[Piton], HApolo], [viața]**	**[Pito**
6	*Cînd a ajuns Apolo în muntele Parnas,*	**[Apolo], [munte]**	**[Apo**
7	*dihania uriaşă s-a avîntat spre dînsul,*	*dihania* ur/aşă**-[Piton],** dfnsu**/=[Apolo]**	**[Apo**

Tabela 6: Analiza HCT a celei de a patra linii de argumentație, sec
1-3-7-8

n	*u$_n$*	*Ci(u$_n$)*	*C$_b$(*
1	*Piton primise-n taină poruncă de la Hera să-l pîndească pe Apolo,*	**[Piton], [Hera], [Apolo],**	**[Pit**
3	*şi 0 să-i răpună viața.*	o= **[Piton], MApolo], [viața]**	**[Pit**
7	*dihania uriaşă s-a avîntat spre dmsul,*	*dihania* **i**/r/aşă**-[Piton],** *dînsul=[Apo\o]*	**[Pit**
8	*0 (era) dornică să-l ucidă.*	o= **[Piton],/=[Apolo]**	**[Pi**

Se constată că tranziția SSH a unității 7 către 8 din
transformat într-o tranziție CON, pe nervură, tot între 7 şi 8 (C
schTmbat din **[Apolo]** în **[Piton],** pentru că, pe nervura lui 8, pre
lui 7 este acum 3, iar nu 6 ca în secvența liniara).

Tabela 7: Analiza HCT a ultimei linii de argumentație, secvența de unități 1-3-7-9-10-11-12

n	*u.*		$C_n(u_n)$	Traz.	Scor
1	*Piton primise-n taină poruncă de la Hera să-l pîndească pe Apolo,*	**[Piton], [Hera], [Apolo],**	**[Piton]**	-	-
3	*şi 0 să-i răpună viața.*	o= **[Piton], /=[Apolo], [viața]**	**[Piton]**	-	-
7	*dihania uriaşă s-a avîntat spre dînsul,*	*dihania uriaşă=[Piton],* d/nsi*l*/**=[Apolo]**	**[Piton]**	-	-
9	*Dar zeul şi-a întins arcul.*	*zeul* **= [Apolo], [arcul]**	**[Apolo]**	SSH	2
10	*0 A tras prima săgeată.*	o **= [Apolo], [săgeata]**	**[Apolo]**	CON	4
11	*Erau doar patru zile de cînd0 văzuse lumea,*	o= **[Apolo], [lumea]**	**[Apolo]**	CON	4
12	*şi întîia lui săgeată a şi nimerit monstrul.*	**[săgeata]** /u/**=[Apolo],** *monstruHPiton]*	**[Apolo]**	RET	3
				Total	**13**

însumând scorurile tranzițiilor pentru toate liniile de argumentație se obține scorul total: 30, ceea ce corespunde unei tranziții medii a discursului, calculată conform HCT de 30/11=2,72, aşadar o tranziție medie mai bună decât cea calculată conform CCT.

6. Validarea conjecturilor VT

Validarea conjecturilor VT s-a realizat pe corpusuri adnotate la structură si la legături co-referențiale. Astfel în [Cristea *et al.,* 1998] se raportează o investigație efectuată pe texte în limbile engleză, franceză şi română ce au însumat un total de 176 de unități de discurs. Plecând de o adnotare în maniera RST a structurii de discurs, un program a calculat expresiile nervurilor unităților. Pentru verificarea conjecturii coeziunii, utilizând adnotarea legăturilor referențiale s-a calculat apoi procentajul referințelor directe, indirecte şi pragmatice. în medie 99,1% dintre referințe se încadrează acestor trei categorii (87,1% directe, 8,5% indirecte şi 3,5% pragmatice). Pentru verificarea conjecturii coerenței, suplimentar marcajelor de structură şi lanțuri co-referențiale s-au marcat manual, pentru fiecare unitate, C_d-ul, în varianta clasică şi în varianta ierarhică, şi s-au calculat tranzițiile în cele două variante. Scorul S_{HCT} a fost mai bun decât scorul S_{CC}TÎn toate cazurile (scorurile medii pe tranziție au fost de 2,03 în varianta ierarhică fată de 1,89 în cea clasică).

în [Cristea *et al.,* 2000] se raportează experimente care au urmărit să compare potențialul modelelor ierarhice, precum cele bazate pe VT, de a regăsi un antecedent într-o plajă de căutare dată față de modelele lineare (modele ce presupun o parcurgere lineară a textului dinspre unitatea anaforului spre începutul

textului). Pentru aceasta s-au utilizat 30 de texte englezeşti (însumând aproximativ 1560 de unități de discurs), adnotate la structura RST şi lanțuri co-referențiale. Presupunând o plajă de căutare de doar 2 unități, căutarea pe nervură a adus cu aproximativ 16% mai mulți antecedenți decât căutarea liniară. După cum era de aşteptat, pe măsură ce lungimea textului căutat creşte, cele două tipuri de modele se apropie în ceea pe priveşte potențialul de a regăsi legături co-referențiale. O căutare ierarhică înapoi într-o plajă de 5 unități rezolvă potențial doar 70% dintre anafore, pentru ca o performanță potențială de 90% să poată fi atinsă doar dacă se organizează o căutare într-o lungime de 12 unități pe nervură. O altă investigare a urmărit compararea efortului necesar regăsirii unui anumit antecedent în cele două tipuri de abordări (liniară şi ierarhică), unde prin efortul necesar găsirii unui antecedent se înțelege numărul de unități de discurs ce separă, în domeniu, unitatea anaforului de unitatea celei mai recente ancorări în text a unui antecedent. Din nou, modelele ierarhice, de tipul celui dat de VT, s-au dovedit superioare celor liniare: în corpusul folosit în experiment, care a conținut 1200 de expresii referențiale, spațiul de căutare pentru legături co-referențiale s-a redus cu aproximativ 800 de unități.

Un alt tip de investigație empirică [Ide, Cristea, 2000] a urmărit frecvența referințelor evocative în comparație cu cele post-evocative şi depistarea unor corelații între tipul de referințe şi puterea de evocare a anaforilor. Studiul a comparat predicțiile avansate de VT relativ la domeniul de referențialitate evocativă cu cele ale modelului stivă al AST, corelând excepțiile (referințe ce nu se supun prevederilor celor două teorii) cu puterea de evocare a anaforilor (pentru VT excepțiile marchează, evident, referințe din categoria celor inferențiale). într-o ordine descendentă a puterii de evocare (v. şi [Gundel *et al.,* 1993]), tipurile de anafori care dau naştere la excepții sunt, în ordinea descrescătoare a frecvenței: referințe pragmatice > nume proprii > substantive comune > pronume. Pronumele constituie mijloace de referire foarte fragile. Un emitent al unui mesaj utilizează un pronume când e sigur că structura permite recuperarea cu uşurință a entității referită de pronume. Practic, exceptând câteva cazuri în care un pronume putea fi înțeles fără un antecedent *(our* în *our streets,* de exemplu), este imposibilă utilizarea unui pronume pentru a referi o entitate aflată în afara *dea.* La extrema cealaltă se plasează referințele pragmatice ce-şi recuperează antecedentul din cunoştințe exterioare discursului şi numele proprii. Interesant este că această sortare descrescătoare a tipurilor de anafori dată de puterea de evocare se aliniază numărului de excepții raportate în cazul VT (56,3% - pragmatice, 22,7% - nume proprii, 16,0% - substantive comune şi 5,0% - pronume) şi nu are nici o semnificație în cazul AST (0,0% - pragmatice, 26,1% - nume proprii, 39,1% - substantive comune şi 34,8% - pronume). Ea probează corectitudinea conjecturii coeziunii.

7. O proprietate de granularitate

Atunci când arborele de structură al discursului se modifică prin trecerea de la o granularitate mai fină la una mai grosieră, constrângerea de accesibilitate, conjecturată de VT, se păstrează.

Demonstrație

Să presupunem un arbore de discurs D pe care s-au calculat expresiile *head* şi nervură ale nodurilor. O operație de mărire a granularității poate fi efectuată dacă o întindere de text, inițial repartizată în mai multe unități, şi **pentru care există un nod, fie el *n,* care să o acopere strict în structura inițială,** este "compactată" într-o singură unitate de discurs mai mare ce va lua locul nodului *n* din structura inițială. Pentru a vedea în ce măsură o astfel de operație poate afecta accesibilitatea vom investiga rezultatul aplicării ei asupra expresiilor *head* şi nervură.

Definiția expresiei *head,* punctul 1, obligă ca expresia *head* a ceea ce înainte de compactare era un nod interior, fie el *n,* să fi fost dată de concatenarea unui şir de etichete de noduri nucleare aflate în secvența de text subîntinsă de *n.* Să notăm acest nod, după compactare, cu o, etichetă compusă din secvența nodurilor terminale pe care le acoperă. De exemplu, pentru arborele din Fig. 9:

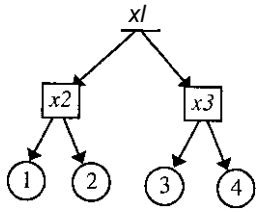


Figura 9: Un subarbore de "compactare"

dacă subarborele cu rădăcina *x2* ar fi compactat, atunci eticheta sa ar trebui să fie notată 1-2, iar dacă întregul arbore aflat sub *x1* ar fi compactat, atunci eticheta sa ar trebui să fie notată 1-2-3-4 (e imposibil să avem un nod notat 2-3).

Acest lucru înseamnă că, aplicând o compactare asupra unui arbore, în expresiile *head* ale nodurilor sale, secvențe de noduri vor fi acum înlocuite cu etichete compuse care conțin cel puțin aceleaşi noduri, eventual mai multe, decât în expresiile originale. De exemplu, presupunând că în arborele de mai sus, nodurile nucleare sunt *x2, x3,* 1 şi 3, atunci, dacă înainte de compactare am fi avut *head(x1)=î* 3, o expresie rezultată din concatenarea a două etichete, după compactarea întregului arbore vom avea *head(x1)=1*-2-3-4, adică o etichetă compusă, dar care include etichetele nodurilor ce apăreau originală. Vom numi astfel de expresii - **expresii contraise** şi *contr\e),* unde e este expresia corespunzătoare de înainte de c deci *contr^* 3) = 1-2-3-4). Să remarcăm că secvențele de etich contrase sunt formate întotdeauna din etichete de noduri ad permite comutarea funcțiilor *seq* si *contr.* seq(confr(e,), *contr*

Vom demonstra mai întâi că expresiile nervură ale noc compactat sunt obținute din expresiile nervură originale prin înl *head* originale cu expresiile contrase. Investigând definițiile expr poate constata că nici o altă modificare nu apare în expresiile n expresiilor contrase. într-adevăr, cazul 1 se transcrie: expresia arborelui compactat reprezintă expresia *head* contrasă a arbor *contr{h),* cu *h* - expresia *head* a rădăcinii arborelui originar.

Să presupunem acum că ne aflăm într-un nod *n* ale că nervură pe arborele originar, necompactat sunt, respectiv *h* şi expresia *head* pe arborele compactat. Considerăm mai întâi ca al nodului rădăcină, a cărui expresie *head* este *contr[h₀),* u expresia *head* pe arborele necompactat. Dacă *n* este nuclea cazului 2 (secțiunea 3.2), avem două subcazuri:

a) *n* nu are un frate nenuclear în stânga: atunci ner nervura părintelui, adică *contr\h₀)',*

b) *n are* un frate nenuclear în stânga de *head contrat):* va fi *seqimarkicontrihi)), contr(ho)) = seqicontrima* confr(seq(A77ar/c(/?₇), *h₀)) = contr(v)',*

Dacă *n* este un nod nenuclear, atunci conform cazului 3

a) *n* este în stânga: nervura sa este *seq(co contr(seq(ho, h)) = contr(v);*

b) *n* este în dreapta: nervura sa este *seq(simpl(c seq{contr(simpl(ho)), contr(h)) = contr[seq{simpl(h₀),*

Folosind inducția, se probează în mod analog că e nodului *n* este o expresie contrasă şi pentru cazul în care *n* este neapărat imediat sub rădăcină, fiu al unui nod de nervură *contr*

Cum expresia accesibilității este definită ca un prefix a din care au fost îndepărtate marcajele, iar nervurile sunt expr eventual conținând mai multe etichete de noduri, înseamnă că pe arborele originar satisface prima conjectură, cu alte cuvinte unitate a unei expresii nervură şi alta ce o precede, după comp de asemenea conjectura, pentru că nici o unitate nu a dispărut

8. Discuții, aplicații ale teoriei

Plecând de la o reprezentare a structurii de discurs similară celei din RST şi în care esențială este distincția dintre nucleu şi satelit, VT defineşte nervura unui nod al arborelui ca secvența de unități ale discursului ce sunt suficiente pentru a rezuma/interpreta întinderea de text acoperită de nod în contextul întregului discurs. Presupunerea principală pe care se bazează noțiunea de nervură este că *referințele inter-unități sunt posibile cu precădere între unități ce se află într-o relație structurală,* chiar dacă acestea sunt dispuse la distanță una de alta în text. Mai departe, referințele se realizează cu precădere spre unități nucleare şi doar în puține cazuri către sateliți, reflectând intuiția că nucleele găzduiesc ideile principale ale discursului. Acest lucru se regăseşte în calculul expresiei nervurii pe arbori (binari) polarizați-stânga (pe orice nivel există un nucleu în stânga), în care orice referință se realizează dinspre un nucleu sau un satelit către un nucleu aflat în stânga (deşi, nu orice nucleu). Făcând uz de echivalarea modelului stivă al lui Grbsz şi Sidner [198Ş] cu structura de arbore utilizată de RST [Mann, Thompson, 1988], similaritate demonstrată de Moser şi Moore [1996] şi Marcu [1999], predicțiile VT asupra accesibilității referențiale sunt consistente cu cele ale modelului stivă. în cazurile în care însă arborele de discurs nu e polarizat-stânga (există cel puțin un satelit care precede nucleul său, deci care apare ca frate stânga pe un nivel al structurii), VT oferă o interpretare mai naturală a accesibilității decât modelul stivă, corectând totodată slăbiciunile acestuia. într-adevăr, într-o secvență /4-satelit, 6-nucleu, deci în care *B* domină *A* în termenii AST, 6 ar trebui să apară în stivă poziționat sub A, deşi el este procesat în secvență după *A*. Totodată, VT formalizează intuiția că într-o secvență de unități A, S, C, unde *A* şi C sunt sateliți ai lui S, C nu poate accesa *A* din cauza interpunerii unui nucleu, ce captează întreaga atenție.

Referențiaiitatea în lungul nervurilor este una naturală, uşor de interpretat şi care, în general, nu necesită mijloace de evocare foarte puternice. Dimpotrivă, ieşirea din acest domeniu incumbă utilizarea unor mijloace de evocare anaforică viguroase. Pe acest criteriu se face distincția dintre referențialitate evocativă şi ne-evocativă (sau inferențială), referințele evocative fiind detaliate în directe şi indirecte, iar între cele ne-evocative remarcându-se referințele pragmatice, ce nu necesită un antecedent pentru înțelegere.

în privința coerenței discursului, VT utilizează domeniile de referențialitate pentru a introduce noțiunea de linie de argumentație şi a deduce din ea pe cea de segment în sens ierarhic ce generalizează segmentul în sens clasic (aşa cum este el utilizat în AST şi CT). Totodată VT avansează conjectura că segmentul în sens ierarhic dă o mai corectă interpretare a porțiunilor de discurs ce se comportă din punctul de vedere al coeziunii şi coerenței ca un tot unitar. Aplicând concluziile CT relative la coerența discursului în lungul segmentelor în sens ierarhic, CT poate fi generalizată pentru a o transforma într-o teorie globală a coerenței.

Au fost trecute în revistă o seamă de experimente care probează că prezumțiile VT sunt corecte şi independente de limbă. Un aspect important îl constituie, de asemenea, faptul că prezumțiile VT sunt stabile la trecerea de la o granularitate mai fină la una mai grosieră în segmentarea discursului.

Aplicațiile VT se înscriu în trei direcții importante: rezoluția anaforei, parsarea discursului şi rezumarea automată. în [Cristea ef a/., 2002a] şi [Cristea *et* a/., 2002b] este descrisă o arhitectură care acționează ca un motor general şi configurabil de rezoluție anaforică. Una dintre componentele oricărui model de rezoluție este o definiție a domeniului de referențialitate. Rezoluția anaforică se realizează, aşadar, ghidată de structura de discurs.

în [Serețan, Cristea, 2002] se propune o abordare inversă, în care cunoştințe asupra legărilor anaforice pot fi utilizate pentru corectarea structurii. Noi credem că procesul de rezoluție anaforică şi de construire a structurii de discurs sunt interdependente într-un asemenea grad încât în analiza de discurs ele trebuie să aibă loc simultan. în interpretarea unui text există o intercondiționare reciprocă între referințe şi structură care trebuie să conducă la obținerea acelei reprezentări în care constrângerile, acționând ca forțe, produc o stare de echilibru, ce trebuie să fie un fel de stare de energie potențială minimă a sistemului. Oamenii dispun de un mecanism cognitiv care le permite să ajungă în mod natural la cea mai plauzibilă interpretare a unui text. Acest lucru este răsplătit de atingerea unei stări mentale "confortabile" ce trebuie să-şi aibă suportul în satisfacerea la maxim a unui sistem de constrângeri. în [Tablan ef a/., 1998] şi [Cristea, 2000] se descrie un mecanism de parsare care modelează acest comportament uman. Prin combinarea unor scoruri contribuite de referințe cu scoruri contribuite de o analiză HCT se obține cea mai fluidă posibil structură de discurs (deci manifestând maximum de coerență) şi care prezintă maximum de referințe pe nervuri (fiind deci cea mai coezivă posibil).

Noțiunea de *head* din VT este similară celei de mulțime de promovare *(promotion set)* pe care Marcu [2000] o utilizează pentru a obține un rezumat ghidat de structura de discurs. Să remarcăm că definiția nervurii presupune rezumarea ca o alternativă a înțelegerii unei unități de discurs în context. Credem că valențele teoriei nervurilor în realizarea unei strategii de rezumare focalizată [Mani, 2001] pe o anumită entitate sau segment de discurs au fost doar tangențial studiate până acum [Sofronie, 1999], [Postolache, 2001] şi merită atenție în abordările viitoare. Credem, de asemenea, ca fiind interesantă o direcție de studiu care să aprecieze maniera în care nervura poate constitui un cadru de sub-specificare a structurii [Schilder, 2001], plecând de la observația că structuri diferite (dau nu fundamental diferite) pot prezenta aceleaşi expresii ale nervurilor unităților componente.

Bibliografie

Brennan, S.E.; Walker Fredman, M. and Pollard, C.J. 1987. A centering approach to pronouns. *Proceedings of the 25th Annual Meeting of ACL,* Stanford, p 155-162.

Cornea, P. 1998. Introducere în teoria lecturii, Editura Polirom, Iaşi.

Cristea, D., and Webber, B.L. 1997. Expectations in incremental Discourse Processing. *Proceedings of the 35th Annual Meeting of the Association for Computațional Linguistics,* Madrid.

Cristea, D., Ide, N., and Romary, L. 1998. Veins Theory: A Model of Global Discourse Cohesion and Coherence, *Proceedings ofthe17th Coling and the 36th Annual Meeting of the ACL (COLING-ACL'98),* Montreal, Canada p.281-185.

Cristea, D., Ide, N., Marcu, D., and Tablan, M.V. 2000. An Empirical Investigation of the Relation Between Discourse Structure and Co-Reference. *Proceedings of the 18th International Conference on Computațional Linguistics COLING'2000,* Saarbrueken, p. 208-214.

Cristea, D. 2000. An Incremental Discourse Parser Architecture. Christodoulakis, D. (Ed.) *Natural Language Processing - NLP 2000,* Second International Conference, Patras, Greece, Lecture Notes in Artificial Intelligence 1835, Springer, p. 162-175.

Cristea, D. and Dima, G.E. 2001. An Integrating Framework for Anaphora Resolution. *Information Science and Technology,* Romanian Academy Publishing House, Bucharest, voi. 4, no. 3-4, p. 259-372.

Cristea, D., Postolache, O.D., Dima, D.E., Barbu C. 2002a. AR-Engine - a framework for unrestricted co-reference resolution. *Proceedings of the Third International Conference on Language Resources and Evaluation, LREC'2002,* Las Palmas, Spain, p. 2000-2006.

Cristea, D., Dima, D.E., Postolache, O.D., Mitkov, R. 2002b. Handling complex anaphora resolution cases. *Proceedings of the Discourse Anaphora and Anaphor Resolution Colloquium,* Lisbon, Portugal.

deEugenio, B. 1990. Centering theory and the Italian pronominal system. *Proceeding of Coling,* p. 270-275.

deEugenio, B. 1998. Centering in Italian. Prince, E., Joshi, A. and Walker, L. (eds.) *Centering in Discourse,* Oxford University Press.

Fox, B. 1987. Discourse Structure and Anaphora. Written and Conversațional English. Cambridge Studies in Linguistics, Cambridge University Press.

Grosz, B.J. 1981. Focusing and description in natural language dialogues. Joshi, A., Webber, B. and Sag, I. (eds.) *Elements of Discourse Understanding,* Cambridge University Press, England, P. 85-105.

Grosz, B.J., Joshi, A.K. and Weinstein, S. 1995 Centering: A framework for modeling the local coherence of discourse. *Computațional Linguistics,* 12(2), p. 203-225.

Grosz, B.J. and Sidner, C. 1986. Attention, intentions, and the structure of discourse. *Computațional Linguistics,* 12(3), p. 175-204.

Gundel, J., Hedberg, N. and Zacharski, R. 1993. Cognitive Status and the Form of Referring Expressions. *Language,* 69, P. 274-307.

Halliday, M.A.K. and Hassan, R. 1976. Cohesion in English, Longman, London and New York.

Hovy, E. 1988. Planning coherent multisentential text. *Proceedings of the 26th ' Annual Meeting of the Association for Computațional Linguistics (ACL-88),* State University of New York, Buffalo, p. 163-169.

Ide, N. and Cristea, D. 2000. A Hierarchical Account of Referențial Accessibility. *Proceedings of the 38th Annual Meeting of the Association for Computațional Linguistics, ACU2000,* Hong Kong, p. 416-424.

Kameyama, M. 1998. Intrasentential Centering: A Case Study. Prince, E., Joshi, A. and Walker, L. (eds.) *Centering in Discourse,* Oxford University Press, p. 89-112.

Kintsch, W. and Van Dijk, T.A. 1975. Comment on se rappelle et on resume les histoires, *Langages,* 40.

Mani, I. 2001. Automatic Summarization. John Benkamin Publishing Company, Amsterdam/Philadelphia.

Mann, W.C. and Thompson, S.A. 1988. Rhetorical Structure Theory: Toward a Funcțional Theory of Text Organization. *Text,* 8(3), p. 243-281.

Marcu, D., 1999. A formal and computațional synthesis of Grosz and Sidner's and Mann and Thompson's theories. *Proceedings of the Workshop on Levels of Representation in Discourse.* Edinburgh.

Marcu, D. 2000. The theory and practice of discourse parsing and summarization, The MIT Press, Cambridge, Massachusetts.

Miller, G. 1956. The magical number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information, *The Psychological Review, voi. 63, p. 81-97.*

Moser, M. and Moore, J.D. 1996. Toward a synthesis of two accounts of discourse structure. *Computațional Linguistics,* 22(3), p. 409-419.

Passonneau, R., J. 1995. Integrating gricean and attentional constraints. *Proceedings of IJCAI.*

Postolache, O. 2001. Sumarizarea textelor. Lucrare de licență. Universitatea "Al.I.Cuza" Iaşi, Facultatea de Informatică.

Rambow, O. (ed.) 1993. Intentionality and Structure in Discourse Relations. *Proceedings of a Workshop Sponsored by the Special Group on Generation of the Association for Computațional Linguistics,* Ohio State University.

Richadeau, F. 1969. La lisibilite. Langage-Typographie-Signes-Lecture, Paris.

Schank, R. and Abelson, R. 1977. Scripts, plâns, goals and understandinq, Hillsdale, N.J.

Schilder, F. 2001. Robust Discourse Parsing Via Discourse Markers, Topicality and Position. *Natural Language Engineering* 1, (1), p.1-22.

Scott, D.R., de Souza, C.S. 1990. Getting the message across in RST-based text generation. Dale,R., Mellish, C. and Zock, M. (eds.) *Current Research in Natural Language Generation,* Academic Press, New York.

Serețan, V. and Cristea, D., 2002. The Use of Referențial Constraints in Structuring Discourse, *Proceedings of the Third International Conference on Language Resources and Evaluation, LREC'2002,* Las Palmas, Spain, p.1231-1237.

Sidner, C. 1983. Focusing in the comprehension of definite anaphora. Brady, M. and Berwick, R.C. (eds.) *Computațional Models of Discourse,* MIT Press.'

Sofronie, V. 1999. Implementări existente în sumarizarea textelor. SumVT. Lucrare de licență. Universitatea "Al.I.Cuza" Iaşi, Facultatea de Informatică.

Strube, M. and Hahn, U. 1996. Funcțional Centering. *Proceedings of the 34th Annual Meeting of the Association for Computațional Linguistics,* Santa Cruz, California., p. 270-277.

. Tablan, M.V., Barbu, C, Popescu, H., Hamza, R.O., Nita, C.I., Bocaniala, C.D., Ciobanu C. and Cristea, D. 1988. Co-operation and Detachment in Discourse Understanding. *Proceedings of the Workshop on Lexical Semantics and Discourse Structure, ESSLLI'98,* Saarbruecken.

Walker, M., lida, M., Cote, S. 1994. Japanese Discourse and the Process of Centering. *Computațional Linguistics,* 20(2), p. 193-232.

Walker, M.A. 1996. Limited attention and discourse structure. *Computațional Linguistics,* 22-2.

Walker, M.A. 1998. Centering, anaphora resolution, and discourse structure. In Marilyn A. Walker, Aravind K. Joshi, and Ellen F. Prince, editors, *Centering in Discourse.* Oxford University Press.

DLIR - un sistem de căutare documentară multilingv

Amalia TODIRAŞCU
INRIA Lorraine, LORIA, Campus scientifique BP 239,
54506 Vandoeuvre-les-Nancy Cedex, France,
todirasc@loria.fr

Abstract

Această lucrare prezintă un sistem de căutare documentară bilingv francez-român pentru un domeniu limitat, cel al securității computerelor. Căutarea şi indexarea documentelor se realizează utilizând o ontologie specifică domeniului. Identificarea instanțelor conceptelor în texte sau în întrebările utilizator se realizează cu ajutorul unor tehnici robuste de analiza limbajului natural, combinate cu o ontologie specifică domeniului.

Introducere

Sistemele de căutare de informații clasice indexează o bază de documente, folosind o listă de cuvinte cheie extrase din documentele respective. Scopul acestor sisteme este de a regăsi documentele care sunt relevante în comparație cu o întrebare lansată de un utilizator. Sistemele de căutare a informațiilor interpretează întrebările utilizatorului, încearcă să găsească un index (unul sau mai multe cuvihte-cheie) care apare în întrebare sau care este similar unui termen din întrebare. Fiecare cuvânt-cheie care aparține indexului este asociat unui document în care cuvântul cheie a fost folosit Răspunsul sistemului conține un număr de documente care sunt relevante în raport cu întrebarea utilizatorului. Fiecare sistem de căutare defineşte un criteriu de relevanță pentru a selecta documentele propuse ca răspuns. Aceste sisteme sunt evaluate pe baza a doi parametri: rapel (numărul de documente regăsite ca răspuns/numărul total de documente relevante care au fost indexate) şi precizie (numărul de documente relevante regăsite de sistem/numărul de documente regăsite). în cazul unui sistem de căutare multilingv, răspunsul la o întrebare poate conține mai multe documente relevante, chiar dacă sunt scrise în alte limbi decât cea în care a fost formulată cererea.

Sistemele de căutare de informații clasice oferă utilizatorului răspunsuri imprecise sau vide. Aceste răspunsuri imprecise apar datorită faptului că majoritatea sistemelor de căutare documentară folosesc doar cuvinte-cheie sau expresii extrase cu ajutorul metodelor statistice, ignorând problemele legate de complexitatea limbajului natural: ambiguitatea (un cuvânt poate avea mai multe sensuri) sau polimorfismul (un concept poate fi exprimat în mai multe moduri). în plus, un sistem care îşi propune să facă o căutare într-o bază de date multilingvă trebuie sa fie capabil să găsească informația cerută în orice document disponibil, indiferent de limba în care a fost scris. Unele sisteme de căutare multilingvă folosesc drept indecşi cuvinte cheie pentru fiecare limbă, alte sisteme propun indexare pe baza unui index comun, alcătuit din concepte.

O alternativă la sistemele de indexare clasice sunt cele care folosesc structuri sintactice sau conceptuale pentru a indexa baza de documente. Acestea nu sunt foarte numeroase, pentru că pe de o parte, ontologiile generice nu sunt disponibile decât în număr prea restrâns (WordNet [19] şi Corelex [3] sunt doar două exemple de resurse libere). Folosirea conceptelor unei ontologii permite asocierea unor termeni din limbi diferite, de aceea am ales o metodă de indexare conceptuală, care va fi în secțiunea 4.

Traducerea termilor care sunt folosiți ca indecşi ridică probleme într-o aplicație de căutare de informații, unui term i se pot asocia mai multe traduceri cu sensuri diferite sau sintagme. Rezolvarea problemelor specifice limbajului natural (ambiguitate, traducere automată) necesită resurse lingvistice importante pentru fiecare limbă care este tratată de către sistem, dacă aplicăm tehnicile clasice de analiza limbajului natural. Tehnicile clasice de analiză sintactică nu sunt adaptate sistemelor de căutare documentară, datorită dimensiunilor prea mari ale bazei documentare şi a resurselor lingvistice necesare. Pe de altă parte, textele specifice necesită adaptarea resurselor lingvistice (dicționare, gramatici locale) folosite de analizoarele sintactice. Tehnicile robuste de analiză sintactică utilizate în domeniul extragerii de informații din texte (GATE [6], FASTUS) sunt dedicate rezolvării unor probleme precise (identificarea numelor proprii, ale grupurilor nominale simple). Printre acestea, automatele cu număr finit de stări [5], colocații [9] sau liste de pattern-uri sintactice (reprezentând structura sintactică a grupului nominal simplu a grupului prepozițional) sunt resursele lingvistice necesare pentru aceste componente. Aceste tehnici au avantajul de a fi robuste, de a putea trata o cantitate importantă de informații în timp real, precum şi de a fi portabile de la un domeniu şi/sau o limbă la alta.

într-o aplicație de căutare de informații pe un domeniu restrâns, utilizatorul doreşte să obțină răspunsuri mai precise decât pentru texte cu caracter general. Aceasta impune folosirea de tehnici adaptate acestui tip de aplicații, bazate pe existența unei baze de cunoştințe din domeniu. Pentru a evita problemele legate de traducerea termilor dintr-o limbă într-alta, propun folosirea unei ontologii specifice domeniului, în vederea ameliorării preciziei. în acest context, voi prezenta

o metodologie de extragere a conceptelor candidat din corpus. Acestea sunt folosite de către un expert uman pentru a îmbogăți o ontologie existentă precum şi pentru a crea o reprezentare sub formă de concepte a documentelor. De asemenea voi prezenta o metodă de indexare a documentelor pe baza acestei ontologii, metodă care modifică metoda clasică de indexare semantică latentă.

2. Ontologii

Noțiunea de ontologie este dificil de definit, mai multe puncte de vedere coexistă. Pentru a simplifica, vom considera că o ontologie este un model restrâns al unui domeniu specific, format din mulțimea claselor de obiecte ce populează acest domeniu şi a relațiilor lor cu celelalte clase din domeniu.

Ontologiile reflectă un anumit grad de subiectivitate din partea expertului ce a definit-o. Fiecare expert poate propune un ansamblu de clase de obiecte ce trebuie incluse în descrierea ontologiei, care poate fi diferit de clasele propuse de alți experți din domeniu.

O problemă a acestor ontologii este legată de portabilitate. O aplicație definită pentru un anumit domeniu dat va trebui adaptată unui alt domeniu prin construirea unei ontologii corespunzătoare. Construirea lor manuală este dificilă şi trebuie ținut cont de posibilele redundanțe, erori, informații care lipsesc sau incoerențe ce pot fi introduse în baza de cunoştințe de către expertul uman care o construieşte.

în ultimii ani, s-au făcut eforturi deosebite pentru a putea reutiliza ontologiile existente: dezvoltarea unor formate standard: (Knowledge Interchange Format - KIF), Ontology Interface Layer - (OIL) [8], dezvoltate în cadrul proiectului Semantic Web (http://www.semweb.org). Aceasta permite reutilizarea ontologiilor existente de către alte sisteme şi aplicații, în ciuda erorilor care pot apare în urma construirii manuale.

Pentru a evita problemele legate de crearea manuală a ontologiilor, au fost propuse mai multe metode semi-automate de extragere a ontologiilor din corpusuri. Acestea disting mai multe etape:

- identificarea termilor (instanțele conceptelor exprimate în limbaj natural);
- identificarea relațiilor între termi;
- identificarea relațiilor între termi şi concepte.

Majoritatea acestor etape necesită validarea rezultatelor de către un expert uman, care va asocia o interpretare claselor de termi şi relațiilor între două mulțimi de termi). Metodele statistice interpretează contextele existente şi

regrupează termii cu contexte identice în aceeaşi clasă [1], [7]. Relaţiile între termi sunt interpretate pe baza informaţiilor de subcategorizare (structura predicat argument) asociate verbelor. Dezavantajul metodelor statistice este acela că necesită corpusuri adnotate de talie importantă pentru a putea învăţa, iar rezultatele (clasele obţinute) nu pot fi întotdeauna interpretate.

în comparaţie cu metodele statistice, metodele bazate pe inferenţe logice propun proceduri semi-automate pentru a verifica validitatea cunoaşterii existente. Conceptele noi, deduse de către regulile de inferenţă, sunt adăugate ierarhiei domeniului dacă sunt coerente cu cunoaşterea existentă. Relaţiile pot fi identificate folosind cunoştinţe legate de subcategorizare [4] sau interpretând relaţiile substantiv-modificator. Supragenerarea de concepte şi costul verificării incoerenţelor şi inconsistentelor cunoaşterii sunt principalele neajunsuri ale metodei. Mai multe formalisme de reprezentare a cunoştinţelor pot fi folosite în astfel de aplicaţii. Am ales logicile terminologice datorită avantajelor pe care le prezintă acestea.

2.1. Logici terminologice

Logicile terminologice (LT) sunt formalisme de reprezentare a cunoştinţelor care sunt derivate din formalismul reţelelor semantice, dar sintaxa şi semantica lor sunt bine definite. Ele combină proprietăţi ale sistemelor orientate-obiect, ale sistemelor de tip frame şi ale logicilor modale.

LT propun o organizare ierarhică a cunoaşterii, pe doua nivele: unul conceptual (T-Box), care descrie clasele abstracte conţinând obiectele relevante pentru modelarea domeniului şi un nivel aserţional (A-Box), conţinând instanţele claselor. Clasele de obiecte (concepte) sunt descrise de relaţii (numite roluri) cu alte concepte, şi cu atributele lor (rolurile cu valori atomice).

2.1.1. Sintaxa şi semantica logicilor terminologice

Operatorii LT sunt inspiraţi de logica de prim ordin:

Operator	Operator Logic	Semantica
D = SOME R C	3x R(y,x)& C(x) & D(y)	Există cel puţin o instanţă a lui C în relaţia R cu o instanţă a lui **D**
D = ALL R C	Vx(R(y,x)**&D(v)** &C(x))	restricţionează co-domeniul relaţiei R
D = AND C1 C2	C1AC2	Conjuncţia de descrieri conceptuale
D = ORC1 G2	C1VC2	Disjuncţia de descrieri conceptuale
C1cC2	C1cC2	Axiom: C1 conţine condiţii necesare pentru C2
D= NOT C	-,C	complementul conceptului C
D = 3n.R.C	3y1...yn (1=<j=<„ R(x, yi)&C(yi)&D(x))	Există cel puţin n obiecte de tip C în relaţia R cu o instanţă a lui D D

Figura **1.** Operatori în LT

Folosind toţi aceşti operatori, sau doar o parte a ac
expresivităţi sunt posibile: definirea conceptelor şi a rolurilor AL
ALL, AND, OR, NOT ca operatori, axiomele conceptuale), p
rolurilor tranzitive (R+), a rolurilor inversabile (I), a ierarhiil
atributelor (f) sau a restricţiilor numerice.

Folosind toţi aceşti operatori, sau doar o parte a ac
expresivităţi sunt posibile: definirea conceptelor şi a rolurilor AL
ALL, AND, OR, NOT ca operatori, axiomele conceptuale), p
rolurilor tranzitive (R+), a rolurilor inversabile (I), a ierarhiil
atributelor (f) sau a restricţiilor numerice.

Unele comenzi LT sunt explicate mai jos. CN este un n
este o descriere conceptuală (orice combinaţie de operatori A
ALL). Comenzile LT sunt inspirate de formalismul KRSS ([2]):

1. (define-concept CN C) - defineşte un nou conce
   conceptuală;
2. (instance IN C) - defineşte o instanţă a unui concep
3. (implies C1 C2) - introduce o nouă axiomă co
   condiţiile C1 necesare pentru descrierea conceptua

LT sunt fragmente decidabile ale logicii de prim
mecanisme logice pentru a identifica subsumarea, regăsirea in
care unesc mai multe concepte. Clasificarea este o ordonare p
concepte, în raport cu relaţia de subsumare. Există algoritr
verificarea coerenţei şi consistenţei cunoştinţelor.

Câteva exemple de comenzi:

(concept-subsumes? C1 C2) testează dacă C1 subsum

(concept-parents C) regăseşte strămoşii direcţi ai conc

(concept-children C) regăseşte fiii direcţi ai lui C

(classify-tbox) calculează toate relaţiile de subsu
conceptele definite în T-Box

(concept-instances C) regăseşte toate instanţele conce

2.2. Logici terminologice pentru sisteme de extrage
a informaţiilor

Rolul cunoştinţelor specifice unui domeniu într-un sis
informaţiilor este acela de a valida reprezentarea semantică a
potenţial relevante, identificate în text prin tehnici de procesare
Aceste entităţi pot fi folosite pentru a adăuga noi concepte la
Cea mai mare parte a sistemelor de extragere a informaţiilor f

robuste pentru identificarea candidaților şi entitățile candidat sunt validate de către o interpretare semantică. Sistemele de extragere a informațiilor pot folosi cunoaştere implicită, cum ar fi relațiile de hiponimie/hiperonimie.

Logicile terminologice prezintă avantajul de a lucra cu date semi-structurate sau incomplete. Nu este necesară definirea explicită a unor valori ca instanțe ale unor concepte. Valorile implicite nu sunt utilizate de către logicile terminologice. Unele valori ale rolurilor sunt lăsate nespecificate ca în următorul exemplu:

(define-concept computer (and physicalobject (some hasOperatingSystem OSystem) (some hasType Type)))

(define-primitive-concept Type)

(define-primitive-concept OSystem)

(instance suni (and computer (some hasType SparcStation)))

în acest exemplu, vom ilustra faptul că definițiile implicite sunt acceptate de către logicile terminologice **(SparcStation** este definit explicit de către o instanță sau un subconcept al conceptului **Type).** Nu este definită explicit nici o instanță a rolului **hasOperatingSystem.**

Aceste proprietăți nu sunt interesante pentru aplicația noastră, dar erorile sunt posibile, iar cunoaşterea domeniului este incompletă.

Relațiile de hiperonimie sau hiponimie sunt tratate cu ajutorul relațiilor de subsumare între conceptele domeniului. De exemplu, dacă un concept candidat este identificat în text ca:

(instance x (and PC (and hasOperatingSystem Linux)))

(define-concept PCcomputer (and computer (some hasType PC)))

x este de asemenea o instanța a conceptului **computer.**

(instance y (and Password (some hasUser Root)))

(define-concept Password (and String (some hasAtr secret) (some hasBelongs User»)

(define-concept System (some hasUser User))

(define-concept Root User)

Pentru aplicația noastră avem nevoie de o logică terminologică care să permită raționament la nivel de instanță, să permită lucrul în contextul unei lumi deschise, precum şi proceduri optimizate de calcul a relațiilor de subsumare sau de clasificare. Printre puținele sisteme care implementează raționament la nivel de instanță am ales RACER ([10]), fiind unul dintre cele mai performante şi mai complete.

în secțiunea următoare voi prezenta metoda de extragere a termilor din texte folosind sistemul DLIR [16]. Textele vor fi traduse într-o reprezentare conceptuală unică, folosită ca index, permițând regăsirea informațiilor în mai multe limbi.

3. Arhitectura

Sistemul DLIR conține mai multe module: un modul de ana robustă, un modul de întreținere a ontologiei domeniului, un modul documentelor bazat pe celelalte două module. în cele ce urmează aceste module în detaliu.

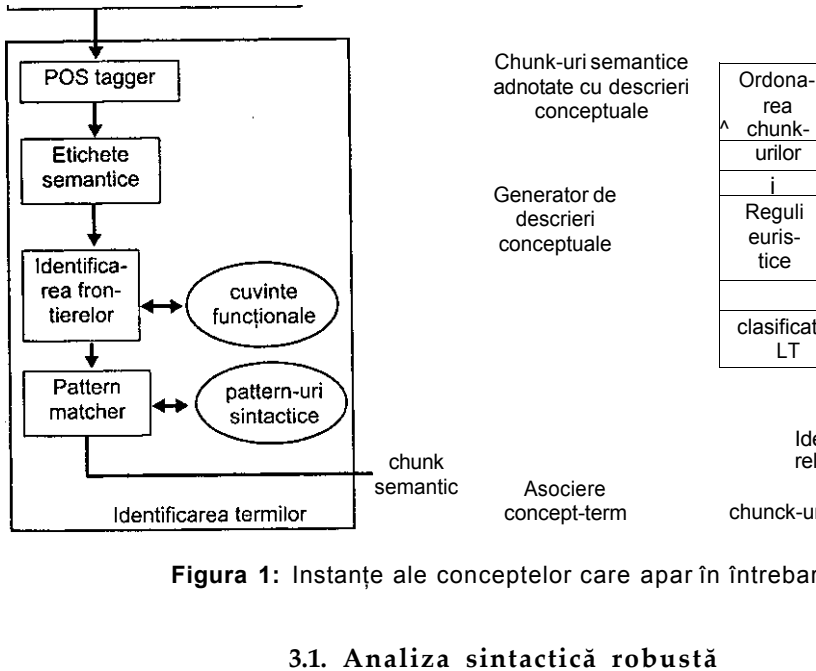


Figura 1: Instanțe ale conceptelor care apar în întrebar

3.1. Analiza sintactică robustă

Acest modul este dedicat identificării termilor posibili, utiliz analiză robuste, şi resurse specifice domeniului (o listă de corespo concepte). Termii sunt combinați conform unor reguli euristice concepte complexe. Aceste concepte sunt validate ulterior, apelâ acces la ontologia domeniului. Eventualele concepte valide sur ontologia existentă. Acest modul conține mai multe submodule implem în Perl şi CLIPS (modulul care aplică regulile de combinare a termilo chunk semantic a fost propusă pentru a identifica termii candidat [16] testată pentru limba franceză, dar cum resursele folosite pentru ident sunt relativ independente de limba pentru care a fost construită posibilă extinderea ei şi pentru limba română, după cum voi arăta mai

3.2. Identificarea chunk-urilor semantice

Scopul principal al acestui modul este acela de a identifica secvențele de cuvinte care corespund celor mai semnificative concepte ale domeniului *(chunk-uri semantice).*

Un *chunk semantic* conține un pattern sintactic simplu (grup substantival simplu, grup verbal) şi este delimitat de doi separatori de clauze.

Separatorii sunt cuvinte funcționale, verbe auxiliare, sau anumite sintagme prepoziționale.

Exemplu."la victime d'une intrusion inattendue"

[victima unei intruziuni neaşteptate]

în acest exemplu, "victima" şi "unei intruziuni neaşteptate" sunt chunk-uri semantice, care conțin informația relevantă.

Modulul conține mai multe submodule: un POS tagger, un tagger semantic, un identificator de frontiere şi un pattern matcher. Identificarea chunk-urilor semantice este bazată pe informația lexicală, propusă de POS tagger.

11 ,

3.2.1. Part-Of-Speech tagging

Modulul care este dedicat identificării părților de vorbire asociate cuvintelor (folosind WinBrill, antrenat pentru franceză pe baza unui set de date propuse de Institut National pour la Langue Frangaise [11]) identifică cuvintele conținut (substantive, adjective, verbe) şi cuvintele funcționale (prepoziții, conjuncții etc).

Taggerul Brill foloseşte un set de reguli contextuale şi lexicale (bazate pe identificarea prefixelor şi a sufixelor), învățate pe baza textelor adnotate, pentru a identifica partea de vorbire pentru cuvintele necunoscute.

Pentru limba română, se foloseşte QTAG adaptat pentru limba română [17], datorită performanțelor foarte bune (98% rezultate corecte).

3.2.2. Tagger-ul semantic

Tagger-ul semantic conține un pattern matcher, care consultă un dicționar de talie redusă. Acesta conține o listă cu cele mai frecvente cuvinte şi un set de sintagme asociate descrierilor conceptuale corespunzătoare.

Setul de descrieri conceptuale a fost stabilit de către un expert pe baza unei liste de cuvinte şi segmente repetate obținute dintr-un corpus reprezentativ (200,000 cuvinte). Un segment repetat este o succesiune de cuvinte care intervin într-un text cel puțin de două ori [14].

Acest modul asociază fiecărui cuvânt conceptul sau descrierea conceptuală din dicționar. Un astfel de dicționar este creat pentru fiecare limbă care este tratată de către sistem.

t
tlltln

3.2.3. Modulul pentru identificarea frontierelor

Acest modul identifică separatorii (cuvinte funcționale s
sintactice mai complexe) care delimitează chunk-urile semantice.
foloseşte rezultatul POS tagger-ului (care identifică cuvintele funcțio
şi un set de sintagme (constituienți sintactici care conțin auxili
compuse). Setul de fraze este construit ca rezultat al studiilor corpu
pentru franceză şi română (200,000 cuvinte pentru fiecare limbă
grupurilor nominale şi prepoziționale (determinanți, prepoziții) sunt
candidați pentru identificarea separatorilor de chunk-uri sema
reprezintă anumite relații potențiale între concepte.

3.2.4. Pattern matcher

Scopul acestui modul este de a identifica nucleul chunk-ur
nucleu care este reprezentat de un grup nominal simplu sau un grup

Exemple. Un grup nominal simplu (în franceză) este identificat a
toarele reguliiN -) NP, N ADJ -) NP,Def N -) NP,Def N ADJ
Pentru limba română, un grup de reguli posibil poate fi: In
Indefart N Adj -) NP, IndefArt N N -) NP

3.2.5. *DLgen*

Acest modul interpretează informația propusă de POS tagge
în mod automat o definiție de concept. Un expert trebuie să veri
acestui modul. Câteva exemple de reguli propuse pentru generarea
simple (valabile pentru ambele limbi):

- S1/N S2/ADJ este asociat definiției (define-concept S
  (SOME hasAtr "S2")))
- S1/N S2/NNP este asociat definiției (define-concep
  hasName "S2"))
- S1/ADJ S2/N este asociat definiției (define-concept S
  (SOME hasAtr "S1")))
- Verbele sunt traduse ca nume de roluri: S1A/B este
  hasS1.

Unele pattern-uri identifică negațiile, chiar dacă este
enumerăm toate posibilitățile şi să detectăm corect domeniul negați

- sans/ADV S1/N este asociat definiției (define-concept n
- nici_unul/ADV S1/N este asociat definiției (define-c
  (NOTS1))

Rezultatele propuse de DLgen sunt 61% corecte datorită faptului că regulile sunt incomplete. Ieşirea este validată de un expert folosind clasificatorul LT pentru a verifica definiţiile conceptuale obţinute în mod automat. Rezultatul este că fiecare structură are asociată o descriere conceptuală.

3.3. Relaţii între termi

Hi

Acest modul foloseşte inferenţele LT, ca şi regulile de sintaxă, pentru a combina descrierile conceptuale asociate fiecărui chunk semantic. Folosim un criteriu de ordonare al chunk-urilor, precum şi reguli de combinare a conceptelor pentru a crea concepte complexe. Descrierile rezultante sunt validate de clasificatorul LT.

3.3.1. Ordonarea chunk-urilor

Modulul interpretează ordinea chunk-urilor şi poziţia chunk-urilor în propoziţie.

Clasificăm chunk-urile în două categorii: **chunk-uri principale** şi **chunk-uri secundare.** Chunk-urile principale corespund noţiunii de nucleu propuse de către teoriile lingvistice clasice.

Chunk-urile secundare joacă rolul unui modificator, care adaugă informaţii
• suplimentare sensului nucleului. Chunk-urile secundare pot lipsi, dar restul propoziţiilor este corect. Aceste exemple de reguli definesc chunk-uri diverse:

- chunk-urile care urmează după un verb la gerunziu sau un auxiliar plus un verb la participiu sunt *chunk-uri secundare;*
- verbele sunt întotdeauna chunk-uri principale.

Exemplu:

'[Main Les atacs Main] [Main ont commence Main] [Second â utiliser les faux comptes Second]'

'atacurile au început utilizând conturi false'

Cele doua chunk-uri principale detectate în exemplul de mai sus sunt subiectul propoziţiei şi verbul principal. Chunk-ul secundar este adnotat astfel pentru că urmează după prepoziţia **â.**

3.3.2. Reguli euristice

Regulile sunt stabilite de către expert pe baza unui studiu asupra corpusului reprezentativ pentru fiecare limbă. Corpusul a fost adnotat cu categoria lexicală propusă de POS tagger şi adnotat manual cu descrierile conceptuale. Setul de reguli euristice este stabilit pe baza unei liste de pattern-uri de forma <*Chunkî* >?x/FW<Cftun/c1 >.

Exemplu de reguli euristice sintactice: dacă o prepoziţie este
două chunk-uri semantice şi prepoziţia asociază u
modificator, atunci putem combina descrierile concep
chunk-uri într-o descriere semantică mai complexă
conceptele fiind cel de modificator:

if (<MainChunk1> <Border> <SecChunk2>)
and (Noun in MainChunkl)
and (Modifier in SecChunk2)
then (and sem(MainChunkl) (some hasModifier

Pentru română, un exemplu de regulă de combinare
următoarea: dacă un verb la gerunziu se găseşte între u
predicativ şi un grup nominal, atunci rolul care leagă conceptel
un modificator.

Fiecare pattern este asociat unui cuvânt ţintă care
pentru aplicarea regulilor. Prepoziţiile, verbele la modul pa
exemple de cuvinte asociate regulilor euristice. Un număr
franceză) şi un număr de 21 de reguli pentru română au fost
Ieşirea acestor reguli va fi o serie de *chunk-uri complexe,* ce
către expert, cu ajutorul ontologiei domeniului, care este ind
Rezultatele propuse de acest modul conţin în mare parte con
o serie de concepte nu sunt identificate. Baza de reguli po
urma studierii unui corpus de dimensiuni mai importante.

3.4. Indexare semantică

O posibilitate de indexare a documentelor este ace
concepte drept index şi nu cuvinte cheie. O metodă efic
reprezintă indexarea semantică latentă. Această metodă c
• document-cuvinte cheie şi foloseşte tehnici de descompuner
metoda valorilor proprii. în acest fel se elimină coloanele şi
(datorat faptului că mulţi termi apar foarte rar). Propunem
care fac parte din ontologie în locul cuvintelor cheie. Este po
de căutare a informaţiilor multilingv să avem diferenţe între o
într-alta. Avantajul este că putem folosi drept index concep
ambelor ontologii. Pentru aplicaţia noastră am folosit o ontolo
care conţine 54 de concepte şi 34 de relaţii.

Numărul de concepte este mai redus decât numărul
în special relaţiile între termi.

Elementele matricii conţin o pondere *weight(C,i)* calc

$$weight(C,i) = \frac{f(C,i)}{\sum_{j=1} f(C,j)}$$

pentru fiecare concept, codificând frecvența instanțelor conceptului în document şi frecvența instanțelor în toate documentele indexate de sistem.

f(C,i) - frecvența conceptului în documentul i;

Conceptele sunt legate prin rolurile dintre acestea. Frecvența unui concept care este situat în ierarhie foarte sus este compus din suma frecvențelor instanțelor sale. Instanțele conceptelor în LT sunt instanțele tuturor subconceptelor şi ale instanțelor sale directe.

Indexarea documentelor se face aplicând metodele de extragere a termilor prezentate în secțiunea precedentă, înainte de a exploata sistemul. Se folosesc conceptele ontologiei care a fost construită manual. O serie de concepte mai generale ar putea fi obținute combinând ontologia specifică domeniului cu WordNet ([16]).

Evaluarea acestui sistem a fost realizată pentru un set restrâns de întrebări (50) numai pentru limba franceză. Rezultatele au fost comparate cu cele furnizate de un sistem care foloseşte cuvinte-cheie pentru indexare. Pentru 74% din întrebări răspunsurile sistemului (rapel şi precizie) au fost comparabile cu cele obținute prin metoda de indexare bazată pe cuvinte-cheie. în celelalte cazuri, răspunsurile au fost mai slabe decât indexarea pe baza de cuvinte cheie. Ontologia folosită este departe de a fi completă, ceea ce a dus la neidentificarea unor termi, de asemenea regulile de formare a conceptelor sau de generare a descrierilor conceptuale sunt incomplete.

4. Concluzii şi perspective

Articolul prezintă o modalitate de a folosi ontologia unui domeniu pentru căutare de informații bilingvă în limbile franceză şi română.

Sistemul integrează tehnici de analiză sintactică robustă pentru extragerea celor mai relevante chunk-uri semantice. Metoda foloseşte o ontologie a domeniului construită manual. Pentru evaluarea pertinentă a metodelor de indexare pe bază de concepte, ontologia va fi actualizată şi extinsă cu ajutorul raționamentelor propuse de logicile terminologice, ca şi folosirea cunoştințelor sintactice, folosite pentru extragerea unei reprezentări semantice pentru texte şi întrebări. Expertul uman trebuie să intervină pentru a decide dacă conceptele identificate în texte pot fi adăugate ontologiei domeniului.

Referințe bibliografice

[1] Assadi, H., Bourigault, D., 2000, Analyse syntaxique e construction d'ontologies â partir des textes. în J.C G.Kassel, D.Bourigault (eds.) - *Ingenierie des conna recentes et nouveaux defis,* Eyrolles Publishing House, pp

[2] Baader, F., Hollunder, B., 1991. A Terminol Representation Systems with Complete Inference Algo *of the Workshop on Processing Declarative Knowledge.*

[3] Buitelaar, P., 1998. CORELEX: Systematic Polysemy and Ph.D. thesis, Brandeis University, Department of Compute

[4] Capponi, N., Toussaint, Y., 2000, Interpretation de cla generalisation de structures predicat-argument. în J.C G.Kassel, D.Bourigault (eds.), Ingenierie des connais *recentes et nouveaux defis,* Eyrolles Publishing House, p

[5] Chanod, J.P., 1999. Natural Language Processing and M.T. Pazienza (ed.), *Information Extraction,* Springer-V pp. 17-31.

[6] Cunningham, H., Wilks, Y., Gaizauskas, R.J., 1996. Ne Trends and Software Infrastructure for NLP. în *conference on New Methods in Natural Language Proc* Bilkent University, Turkey, 1996, pp.1-12.

[7] Daille, B., 1996, Study and Implementation of Combi Automatic Extraction of Terminology. In J.Klavans, P. *Balancing Act* - Combining Symbolic and Statisti Language, MIT Press, pp. 49-66.

[8] Fensel, D. et al., 2000, OIL in a nutshell. în R. Dieng et a Acquisition, Modeling, and Management, *Proceedir Knowledge Acquisition Conference (EKAW-2000),* Lectur Intelligence, LNAI, Springer-Verlag.

[9] Heid, U., 2000, A linguistic bootstrapping approach to candidates from German text, *Terminology,* pp. 161-180

[10] Haarslev, V., Muller R., 2001, Description of the RA Applications, *Proceedings of the International Work Logics* (DL-2001), Stanford, USA, 1.-3. August 2001, pp

[11] Lecomte, J., Le Categoriseur BRILL14-JL5/WINBRIL CNRS report, December 1998.

[12] Riloff, E., Lorenzen, J., 1999, Extraction-based Generating Domain-Specific Role Relationships Aut Strzalkowski, *Natural Language Information Retrieval* Publishers, pp. 167-196.

[13] Riloff, E., Shepherd, J., 1997, A Corpus-Based Approach for Building Semantic Lexicons. în *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing.*

[14] Rousselot, F., Frath, P., Oueslati, R./Extracting concepts and relations from Corpora. In *Proceedings of the Workshop on Corpus-oriented Semantic Analysis, European Conference on Artificial Intelligence,* ECAI 96, Budapest 12 August 1996.

[15] Schimd, H., 1994, Probabilistic Part-of-Speech Tagging Using Decision Trees, *Proceedings of the International Conference on New Methods in Language Processing, Manchester, United Kingdom.*

[16] Todiraşcu, A., 2001, Semantic Indexing for Information Retrieval Systems, Ph.D. Thesis, University Louis Pasteur of Strasbourg, France, March 2001.

[17] Tufiş, D., Mason O., Tagging Romanian Texts: a Case Study for QTAG, a Language Independent Probabilistic Tagger. în *Proceedings of the First International Conference on Language Resources and Evaluation (LREC)* Granada, Spain, 1998, pp. 589-596.

[18] Vilain, M., 1999, Inferential Information Extraction. In M.Pazienza (ed.), *Information Extraction,* LNAI 1714, Springer-Verlag, pp.95-119.

[19] Vossen, P., *Introduction to EuroWordNet,* Kluwer Academic Publisher, 1998.

[20] Zweigenbaum, P., Consortium MENELAS, 1995, MENELAS: Coding and Information Retrieval from Natural Language Patient Discharge Summaries. In M.-F. Laires, M.J. Ladeira, J.-P. Christensen (eds.) - *Advances in Health Telematics,* IOS Press, Amsterdam, pp.82-89.

Mediu hermenofor pentru asistarea învățării unor concepte dintr-o limbă străină

Ştefan TRĂUŞAN-MATU
Universitatea "Politehnica" Bucureşti, Facultatea de Automatică şi Calculatoare, Institutul de Cercetări pentru Inteligență Artificială al Academiei Române
email: trausan@cs.pub.ro, trausan@racai.ro
URL: www.racai.ro/~trausan

1. WWW, o prezență din ce în ce mai comună

în mai puțin de zece ani, rețeaua globală de documente World Wide We (WWW sau, pe scurt web), a devenit omniprezentă şi este posibil ca într-un tim nu prea lung să înlocuiască o mare parte din cărți, televizorul, cinematografu ziarele şi revistele (toate acestea fiind deja disponibile pe web) şi, în plus s furnizeze chiar posibilitatea imersiunii în realități virtuale. Un singur exemplu cre că este suficient: anul trecut rezultatele bacalaureatului au fost publicate pe web.

WWW a atins deja dimensiuni comparabile cu imensa Bibliotecă Congresului SUA. Extinderea sa este datorată uşurinței cu care poate fi parcurs d către oricine are un calculator şi de simplitatea cu care se poate publica ceva p el. Pe de altă parte, costul accesului la resursele web este de cele mai multe o infim.

WWW este un hipertext extins la scara întregului glob prin rețeau mondială de calculatoare Internet. Pe fiecare calculator pot fi plasate unul sau ma multe documente care constituie noduri (pagini) în hipertext. Oriunde într-o astf de pagină poate exista o legătură la o alta pagină, de pe acelaşi calculator sau d pe altul, în acest mod putând fi unite informații aflate în locuri diferite. O nou pagină pentru web poate fi creată uşor chiar de utilizatori nu neapărat profesioniş în informatică, în acest scop existând mai multe editoare de texte specializate.

Termenul de hipertext se pare că provine de la termenul de spați hiperbolic sau hiperspațiu, apărut în 1704 şi folosit de matematicianul F. Kle pentru geometria cu mai multe dimensiuni [Rad91]. Din această perspectivă, u hipertext este un text cu mai multe dimensiuni explicite (față de doar o dimensiun în cazul textului liniar). De fapt, orice text are implicit mai multe dimensiun deoarece, chiar dacă forma de prezentare a unui text este liniară, pe hârtie, în

există o structură implicită, dată de discurs. De asemenea, există conexiuni implicite, subiective între părți ale textului, concepte legate între ele, Hipertextul este o organizare a unui text în care toate aceste legături sunt explicitate şi pot fi exploatate în parcurgerea făcută pe un calculator.

în jurul anului 1962, Douglas Engelbart a dezvoltat primul sistem hipertext, prezentat atunci drept o arhitectură conceptuală destinată creşterii potențialului intelectului uman ("Conceptual Framework for Augmenting Human Intellect") [Eng95]. Sistemul era destinat manipulării de concepte structurate într-o rețea în care arcele sunt relațiile între concepte.

Primul sistem declarat ca fiind hipertext a fost creat de Theodor Nelson în 1967 sub numele de "Xanadu". Nelson îşi propunea atunci să dezvolte un sistem, masiv paralel, destinat muncii creative şi studiului. El a plecat în îndeplinirea acestei idei de la dorința de a găsi cea mai bună abstracție care să unifice literatura şi arta cinematografică.

Sistemele hipertext (hipermedia) pe web permit accesul personalizat la volume imense de informații. în acelaşi timp, însă, ele suferă de problema afluxului de informație cu care este bombardat un utilizator. O soluție este dezvoltarea de instrumente, aplicații, medii informatice pentru facilitarea accesului la cunoştințele dorite pe web. Aceste instrumente trebuie să facilteze înțelegerea, abstractizarea textelor, extragerea informațiilor utile. Acesta este unul din motivele pentru care le-am denumit instrumente hermenofore. Trebuie remarcat faptul că ideea de a considera hipertextele ca intrumente de sprijinire a activităților cognitive a stat chiar la baza conceperii acestora, după cum am precizat în paragrafele anterioare.

în continuare, după o trecere în revistă a problematicii ontologiilor, în secțiunea următoare, se va introduce conceptul de mediu hermenofor, se va justifica necesitatea acestuia şi se vor prezenta caracteristicile acestora. Lucrarea va fi încheiată cu o exemplificare printr-un sistem care are câteva trăsături ale unui mediu hermenofor şi de o secțiune de concluzii.

2. Ontologii

Termenul de "ontologie" a fost, până nu de mult, folosit exclusiv în filosofie, pentru a denumi teoria asupra existenței, mai corect spus, asupra ceea ce consideră că există cel care întocmeşte teoria. Construirea multor sisteme filosofice pleacă de la o ontologie, adică de la definirea categoriilor fundamentale de entități din realitate şi a relațiilor dintre ele. Chiar dacă ontologia nu este întotdeauna explicită, orice demers conceptual construieşte o ontologie, chiar implicit, inconştient.

în ultimii ani, termenul de ontologie este folosit şi în ştiința calculatoarelor. Cea mai frecventă extindere a folosirii acestui concept este în cadrul sistemelor de

inteligență artificială bazate pe cunoştințe. Majoritatea programelor de calculator cu inteligență artificială prelucrează structuri de simboluri, care sunt menite a reprezenta conceptele, cunoştințele referitoare la domeniul considerat. Aceste structuri simbolice sunt grupate într-o aşa numită bază de cunoştințe care constituie, de fapt, un model al domeniului respectiv. în ultimii ani se consideră că această bază de cunoştințe trebuie văzută ca o ontologie, o conceptualizare, o teorie asupra ceea ce există în domeniul avut în vedere. O ontologie este, din această perspectivă, o "specificare a unei conceptualizări ... Termenul este împrumutat din filosofie, unde însemna o considerare sistematică a existenței. în inteligența artificială se referă la precizarea a ceea ce se consideră că « e x i s t ă » [Gru96].'

între concepte pot exista diverse relații. Cea mai importantă relație este probabil cea hiperonimică [WN], taxonomică, între un concept şi unul sau mai multe concepte mai generale, din care derivează, care îl subsumează, din a cărui combinație a fost generat. Prin această relație se pot "moşteni" proprietăți de la conceptul (conceptele) mai general(e) la cel mai particular, dacă aceste proprietăți nu sunt redefinite la conceptul din urmă. Alte relații sunt cea meronimică [WN] ("parte-întreg"), între un concept şi părțile sale sau cea antonimică, între două concepte (adjective) opuse.

O ontologie include, aşadar:

- categoriile, conceptele fundamentale,
- proprietățile conceptelor,
- relațiile şi distincțiile între concepte.

O ontologie este rezultatul unei experienții, a unor experiențe trăite, în care sunt evidențiate nişte constante, nişte regularități, care ne îndreptățesc să afirmăm că vor fi regăsite în viitor. în urma investigației făcută pentru a găsi esența regularităților se delimitează entități mentale denumite concepte sau categorii, care pot fi diferențiate de alte categorii. Aceste entități pot intra în combinație cu altele formând noi concepte.

Un aspect deosebit de important în ceea ce priveşte rolul ontologiilor este faptul că ele exprimă o comuniune, (co)existența unei diversități de concepte, de diferențe şi relații între ele. O presupoziție este că există doar un număr limitat de concepte sau categorii, ceea ce înseamnă că se poate face un fel de cuantificare de discretizare a realității. Acestea constituie un punct de sprijin pentru achiziția de noi concepte sau pentru raționamentele făcute de om sau de calculator.

Partajarea unei ontologii este esențială în sistemele bazate pe agenți (pro grame) inteligenți pentru, de exemplu, comerțul electronic, pentru a le asigura autonomia, flexibilitatea şi agilitatea. Ontologiile sunt liantul care integrează sisteme de baze de date, sisteme de obiecte, sisteme bazate pe cunoştințe, diverse aplicații integratoare şi bazate pe colaborare. Ele reduc ambiguități

semantice în partajarea şi reutilizarea cunoştinţelor. "Scopul suprem este dezvoltarea de ontologii reutilizabile care pot fi aplicate pentru mai multe discipline". [OORG]

"O ontologie are drept prim scop facilitarea comunicării între calculatoare, independent de tehnologiile unui anumit sistem individual, arhitectura de prelucrare a informaţiilor şi domeniul aplicaţiei. Ingredienţii cheie care constituie o ontologie sunt un vocabular de termeni de bază şi o specificare precisă a ceea ce înseamnă aceşti termeni." [OORG] O ontologie este însă mai mult decât un vocabular. Ea este punctul de plecare pentru dezvoltarea de structuri de cunoştinţe, nu numai taxonomii sau clasificări de concepte ci şi relaţii complexe. [OORG]

Din punct de vedere al programelor de calculator care folosesc ontologiile, există două tipuri de ontologii. Primul tip este cel al ontologiilor destinate sistemelor bazate pe cunoştinţe, de exemplu, al unui sistem de diagnostic medical. Aceste ontologii sunt caracterizate de un număr relativ redus de concepte, dar legate între ele printr-un număr mare şi variat de relaţii. Conceptele sunt grupate în scheme conceptuale complexe sau scenarii. Pentru fiecare concept pot exista una sau mai multe particularizări.

Spre deosebire de primul tip de ontologii, ontologiile lexicalizate includ un număr foarte mare de concepte, legate printr-un număr redus de tipuri de relaţii (de exemplu, hiperonimică, meronimică etc). Conceptele sunt reprezentate, de exemplu în WordNet [WN], prin mulţimi de cuvinte sinonime. Astfel de ontologii sunt folosite în sistemele de prelucrare a limbajului uman.

Corespondenta ontologiei WordNet (care este concepută pentru limba engleză-americană) pentru limbile europene este EuroWordNet. Aceasta din urmă aduce avantajul că, fiind dezvoltată pentru mai multe limbi (engleză, franceză, germană, italiană, olandeză etc), permite şi dezvoltarea de aplicaţii multilingve. în prezent, în cadrul Institutului de Cercetări pentru Inteligenţă Artificială al Academiei Române este în desfăşurare, în colaborare cu mai multe ţări din regiunea balcanică proiectul BalkanNet pentru integrarea în EuroWordNet a limbilor din zonă, inclusiv a limbii române.

3. Medii hermenofore

Denumim mediu hermenofor o colecţie integrată de instrumente (pe care le vom numi hermenofore) şi aplicaţii informatice direcţionate către facilitarea unor activităţi de tip hermeneutic ale unui utilizator care explorează resurse aflate pe web. Termenul "hermenofor" [Tra01] poate fi parafrazat prin "generator de hermeneutică", pentru a sugera faptul că un mediu hermenofor facilitează activităţi hermeneutice, care acordă un rol important experienţierii şi sunt orientate spre descoperirea unor înţelesuri, a unor structuri profunde, greu detectabile.

Elaborarea de medii hermenofore este absolut necesară în contextul actual al exploziei numărului şi volumului de resurse şi a interconexiunilor între acestea pe web. Sistemele hipertext (hipermedia) aduc noi dimensiuni cum ar fi interactivitatea, posibilităţile cu totul remarcabile de vizualizare, accesul personalizat la volume imense de informaţii. în acelaşi timp, însă, ele introduc şi unele probleme datorate afluxului de informaţie, care poate duce la depăşirea capacităţilor cognitive ale utilizatorului, la dezorientare şi chiar la alienare. Este un fapt că utilizatorul, chiar profesionist în informatică, poate fi dezorientat în "labirintul" de pagini de web şi resurse de tot felul (baze de date, documente, imagini, ontologii, lexicoane etc.) interconectate.

O soluţie la problemele enumerate mai sus este dezvoltarea de instrumente, aplicaţii, medii informatice pentru facilitarea accesului la cunoştinţele dorite pe web. Se poate spune, din această perspectivă, că browserele de web, "motoarele de căutare", agenţii (asistenţii) software sunt rudimente de medii hermenofore. Justificarea necesităţii considerării perspectivei hermenofore este lipsa abilităţilor hermeneutice ale acestor aplicaţii. Un exemplu tipic este faptul că "motoarele de căutare pe web" (de exemplu Google [Goo]) furnizează mii sau chiar zeci de mii de documente ca răspuns la o cerere. Alt exemplu este limita actuală a programelor de calculator în înţelegerea textelor cu scopul traducerii, sumarizării sau extragerii cunoştinţelor. Aceste probleme sunt datorate, în primul rând, problemelor generate de ambiguitatea limbajului natural, a aspectelor legate de semantică, de pragmatică, de interpretare, de considerarea contextului, a metaforelor, a cunoştinţelor de "bun simţ". Toate aceste probleme sunt recunoscute ca fiind "nodul gordian" al aplicaţiilor de inteligenţă artificială. După cum remarca Terry Winograd, programele de inteligenţă artificială nu pot depăşi condiţia unui birocrat, care nu poate să acţioneze când nu are "reguli", care nu se implică [Win87]. Putem spune că, de fapt, problema este că acestor aplicaţii le lipsesc abilităţile hermeneutice. Ideea noastră este de a oferi un cadru în care puterea oferită de tehnologia informaţiei să fie integrată cu capabilităţile specific umane.

Hermeneutica este, după opinia lui P. Ricoeur, o abordare complementară celei structuraliste în analiza limbajului, a înţelesului şi simbolismului cultural. "Hermeneutica bazează înţelegerea textelor pe intenţiile şi istoria autorilor şi relevanţa acestor fapte pentru cititori. în contrast, filosofia analitică identifică de obicei înţelesul cu referenţi externi pentru texte iar structuralismul găsind înţelesul în aranjarea cuvintelor. Hermeneutica priveşte textele ca mijloace pentru a transmite experienţa, crezurile şi judecăţile de la un subiect sau comunitate către alţii. Astfel, determinarea înţelesurilor este o problemă de judecată practică şi raţionament de «bun simţ» şi nu privitor la o teorie a priori sau o demonstraţie ştiinţifică." [MHD].

Hermeneutica este studiul interpretării, inițial ea referindu-se doar la interpretarea textelor [MHD]. în prezent s-a extins accepțiunea termenului hermeneutică, vorbindu-se de o poziție hermeneutică în filosofie, care include pe Heidegger, Gadamer, Habermas şi Ricoeur, deosebită de formalişti (filosofia analitică, neo-pozitivism sau pozitivismul logic), reprezentați prin Descartes, Leibniz şi Russell [Wes97]. Distincția între cele două abordări pleacă de la problema capturării înțelesului. Pe când formaliştii pretind că pot reprezenta înțelesul, semantica, doar prin identificarea unui denotat în lumea reală corespunzător unei expresii formale, adepții hermeneuticii neagă această posibilitate, pentru ei înțelesul implicând şi considerarea experienței, a credințelor subiectului. Se poate spune că, dintr-un punct de vedere se ajunge la aceeaşi dispută dintre Husserl si Heidegger sau dintre Dennett si Chalmers.

Mediile hermenofore furnizează informațiile dorite dintr-o perspectivă particulară, pentru un anumit utilizator, considerând un anumit domeniu şi într-un anumit moment dat. Un mediu hermenofor trebuie conceput deci în scopul personalizării interfațării la resursele web-ului, pentru a facilita înțelegerea. Dacă prezentările făcute într-un mediu hermenofor sunt structurate ca hipermedia, una din preocupările principale ce trebuie avute în vedere este faptul că utilizatorul trebuie să experiențieze parcurgerea unei secvențe de pagini de web, secvență care trebuie să respecte nişte reguli de pragmatică.

In plus față de furnizarea unei interfețe adaptabile, o altă caracteristică a unui mediu hermenofor trebuie să fie facilitarea inițiativei utilizatorului. El trebuie să poată experimenta, să poată investiga resursele web-ului. Instrumentele hermenofore sunt destinate sprijinirii activității hermeneutice umane adică a unei atitudini direcționate către înțelegerea unor cunoştințe sau structuri ascunse în texte (hipiertextelor, hipermedia). Un rol important în procesul înțelegerii îl au modalitățile de a genera experienței, adică experiențe de trăire, fapte de viață (conform teoriei că înțelegerea necesită un proces empatic [Wri95], [Mar97]). Unul dintre cele mai uzitate mijloace de acest gen este folosirea metaforelor [LaJ80], [TraOO]. în acest sens se înscrie preocuparea de a dezvolta instrumente (hermenofore) pentru detectarea, adnotarea şi prelucrarea metaforelor.

O caracteristică pe care o considerăm esențială la un mediu hermenofor, în contextul precizat mai sus, este şi posibilitatea de vizualizare multiplă, din perspective diferite, a aceluiaşi document. Enumerăm aici, drept exemplu, în afara perspectivei conținutului "brut" al unui document, alte perspective, date de concordanțe, adnotări (cu părți de vorbire, de exemplu), extrase, rezumate, arbori de analiză semantică,, structuri care reprezintă conținutul semantic. Remarcăm, în acest context, rolul extraordinar de important al adnotărilor documentelor în limbajul extrem de versatil care este XML [XML].

Vom considera că instrumentele hermonofore au ca scop revelarea şi valorizarea unor cunoştințe sau a unor structuri încorporate în volumele imense de hipetexte şi hipermedia de pe web. Datorită faptului că abordarea p pune pe prim plan rolul experienței umane, un instrument herm pt neapărat considerat în relație cu utilizatorul care îl foloseşte. De ace să aibă asociat modelul utilizatorului, care să conțină cel puțin informații despre utilizator:

- ontologia sa,
- scopurile urmărite,
- profilul psihologic,
- istoricul acțiunilor efectuate,
- preferințele sale (explicite sau implicite, derivate di comportamentului său).

Pe de altă parte, instrumentele hermenofore trebuie să aspectele legate de particularitățile autorilor documentelor:

- ontologiile considerate (de exemplu, ontologiile impus mele sau de practicile domeniilor considerate),
- scopurile presupuse,
- elemente de istoric,
- aspecte psihologice general umane.

Instrumentele hermenofore pot fi împărțite în mai multe clase acțiunile efectuate:

- căutare a documentelor relevante,
- categorizare a documentelor conform unei taxonomii pre
- relevare de regularități (de exemplu, colocații) sa documente,
- segmentarea textelor,
- extragere de informații sau cunoştințe din documente,
- sumarizare,
- relevare de structuri pe web [WSD97],
- instrumente de adnotare (la nivel sintactic, semantic sa documentelor.

Spre deosebire de instrumentele de minerit al textelor instrumentele hermenofore pun, în plus, accentul pe aspectele leg interacțiunii, de experiența utilizatorului.

în secțiunea următoare se va prezenta sistemul GenW asistată a învățării terminologiei financiare într-o limbă străină [TM

care a fost dezvoltat ca un modul într-un proiect mai mare, denumit "Larflast" şi finanţat de Comunitatea Europeană. GenWeb a implementat instrumente hermenofore care identifică şi utilizează metafore pentru a facilita înţelegerea unui anumit concept [TraOO]. în acest scop, el caută metafore în texte considerate relevante. Metaforele sunt identificate printre perechile de cuvinte care corespund la concepte din ontologia domeniului considerat (finanţe) şi din ontologia metaforelor, aceasta din urmă reflectând aspecte psihologice general umane [LaJ80]. Trecerea de la un concept la o mulţime de cuvinte (sinonime sau înrudite) se face pe baza ontologiei WordNet, derivată din investigaţii psiholingvistice [WN]. Metaforele sunt adnotate în XML [XML], unul din atributele folosite în adnotare fiind scopul urmărit de autor [TraOO].

Tot în GenWeb, textele adnotate cu metafore sunt folosite ulterior pentru a genera structuri (bazate pe principii retorice) de pagini de web personalizate conform modelului utilizatorului. Aceste structuri se constituie într-un sit în care cel care învaţă poate experienţia. Tot pe post de instrumente hermenofore, în GenWeb este disponibilă vizualizarea de concordanţe în context.

4. Sistem de instruire asistată cu calculatorul în înţelegerea unor termeni financiari

Există mai multe puncte de vedere asupra modului în care are loc un proces de învăţare. Suntem de partea abordării constructiviste [BIM96, Wil96] în conceperea proceselor educaţionale. Această abordare consideră că fiecare dintre noi ne construim propria realitate, propriul bagaj de cunoştinţe, plecând de la experienţele pe care le-am avut [ErK97]. După cum remarcă [BIM96], "Nucleul studiului este activitatea hermeneutică a construcţiei de interpretări." învăţarea poate fi şi ea văzută constructivist ca un proces hermeneutic, de înţelegere, de transpunere în domeniul studiat, de experimentare, de trăire.

Plecând de la ideile învăţării constructiviste se ajunge la următoarele principii [ErK97]:

- învăţarea este un proces activ în care studenţii experimentează, caută să înţeleagă singuri ceea ce învaţă, profesorul fiind mai mult un îndrumător;
- învăţarea trebuie să fie un proces auto-reglat de către studenţi;
- învăţarea constructivă este un proces situaţional în sensul că studentul trebuie introdus într-un mediu de învăţare care îi permite să experimenteze, în care se pot face simulări;
- învăţarea trebuie să fie socială, trebuie să existe o permanentă colaborare a studentului cu colegii lui.

Dintr-o altă perspectivă, învăţarea poate fi considerată ca un proces de inducere de modele mentale adecvate [JoL83]. înţelegerea poate fi văzută astfel ca momentul în care realitatea supusă comprehensiunii este pusă în corespondenţă cu un model mental complet şi valid. Empatia [Mar 97], identificarea eu-lui cu starea de lucruri considerată poate fi, în acest caz, tocmai sentimentul de "trăire" în lumea modelului mental.

O practică deja răspândită este de a dezvolta sisteme inteligente de asistare cu calculatorul a instruirii ("Intelligent Tutoring Systems") care încearcă să monitorizeze procesul de învăţare prin verificarea asimilării conceptelor din ontologia domeniului considerat [Tra95]. Se consideră că un model adecvat al cunoştinţelor elevului poate fi construit prin raportare la această ontologie. De fapt, această metodă este folosită şi în învăţământul tradiţional: noii termeni sunt introduşi prin genul proxim şi diferenţa specifică. în termenii ontologiilor, noii termeni sunt definiţi prin superconceptele care-i subsumează şi prin particularităţile care-i diferenţiază.

Orice profesor ştie însă că astfel de definiţii sunt necesare dar nu sunt suficiente. Pentru a aprofunda termenii definiţi sunt necesare exemple, imagini cu un grad mai mare sau mai mic de iconicitate, plecând de la poze şi schiţe, diagrame şi grafice, până la imagini sugerate, până la metafore. Acest fapt este prezent nu numai în învăţământ, el apare în orice proces de comunicare (învăţământul fiind, bineînţeles, şi el inclus).

în cele ce urmează nu ne vom referi la utilizarea imaginilor propriu-zise, care facilitează evident învăţarea sau comunicarea. Vom considera un caz particular de imagini, mentale, sugerate, semne iconice lipsite de caracterul vizual dar care comunică o experienţiere (de multe ori chiar mai puternic, printr-un efect care ar putea face să ne gândim la percepţia subliminală). Este cazul metaforelor, care sunt folosite într-o proporţie de cele mai multe ori nebănuit de mare în comunicarea inter-umană.

Pentru a ilustra puterea de expresie a metaforelor şi, bineînţeles, rolul lor în înţelegerea unor termeni, am să exemplific prin metafora "acţiunile la bursă sunt nişte creaturi foarte sensibile" (găsită într-un text pe situl de web al Bursei din New York - http://www.nyse.com). Nu este nevoie să ne imaginăm o anumită creatură concretă pentru a înţelege ce sugerează metafora exemplificată. Succesul unei metafore, puterea ei expresivă, capacitatea de comunicare sunt date de măsura în care "rezonăm" la mesajul transmis. Ori ce este mai percutant pentru un om decât faptul că suntem creaturi extrem de sensibile? Prin urmare, succesul metaforei folosită într-un context foarte pragmatic, al discursului unui specialist în finanţe este determinat de inspiraţia vorbitorului de a se referi la un fapt general uman. Nici o definiţie de tip gen proxim-diferenţă specifică nu poate comunica experienţa referitoare la aspectul foarte fragil al acţiunilor la bursă precum o face metafora de mai sus.

Rolul covârşitor al metaforelor în viaţa noastră a fost remarcat şi de Lucian Blaga ("omul este un animal metaforic" [Bla85]) şi a fost foarte bine evidenţiat de Lakoff şi Johnson într-o lucrare cu un puternic impact ("Metaforele cu care trăim" - "Metaphors we live by" [LaJ80]). Cei doi autori americani consideră că "subcategorizarea şi metaforele sunt două extremităţi ale unei continuum", că metaforele "formează sisteme coerente în care ne conceptualizăm experienţa" [LaJ80]. Putem spune că metaforele oferă alte mijloacd expresive decât cele de categorizare oferite de ontologii. Ele nu ţin de logica lui Ares, care categorizează, ci de logica lui Hermes, propusă de Noica [Noi86].

Dintr-o altă perspectivă, metaforele pot fi considerate instrumente empatice, care determină imersiunea cititorului (receptorului) în lumea experienţelor autorului. Acest fapt era evidenţiat şi de Lakoff şi Johnson: "Esenţa metaforei este înţelegerea şi experienţierea unui lucru prin altul" [LaJ80]. De exemplu, metafora amintită mai sus despre acţiunile la bursă ne comunică o informaţie pe care orice fiinţă vie o înţelege (sensibilitatea, perisabilitatea) dar care nu poate fi exprimată în categorizări.

Importanţa metaforelor a fost revelată şi de studiul preliminar făcut în cadrul proiectului Larflast (care a avut drept scop elaborarea unui sistem de asistare cu calculatorul a învăţării terminologiei financiare într-o limbă străină [Lar], [TMC02], [ABK02]) de o profesoară de limba engleză la o facultate economică din Sofia. Dânsa remarca ca o importantă dificultate "înţelegerea metaforelor. Limbajul economic şi financiar este extrem de metaforic şi, uneori, grupuri de metafore apar în imagini complexe. Deseori cuvinte uzuale sunt folosite în metafore elaborate, ... cum ar fi «a susţine o pierdere»" [Vit99].

Proiectul Larflast a inclus mai multe module tipice pentru sisteme inteligente de instruire, cum ar fi o ontologie, un mecanism de inferenţă, teste (grilă) pentru diagnosticarea cunoştinţelor elevului şi actualizarea modelului acestuia. Sistemul dezvoltat include cinci servere de web, unul la Bucureşti şi altele la Leeds, Manchester, Montpellier şi Sofia. Serverul de la Bucureşti, după ce este lansat, accesează serverul de la Sofia pentru a prelua modelul elevului (ce concepte ştie şi ce concepte nu) şi apoi generează pagini de web personalizate.

Metaforele sunt identificate în texte considerate relevante care au fost obţinute în urma căutării cu o maşină de căutare uzuală (de exemplu, Google [Goo]. Textele găsite sunt grupate într-un corpus care este adnotat cu metaforele identificate. Acest corpus, împreună cu ontologia domeniului şi cu modelul studentului (construit pe baza răspunsurilor date de student la teste) sunt folosite pentru generarea personalizată de pagini de web. în figura următoare este ilustrată arhitectura sistemului GenWeb.

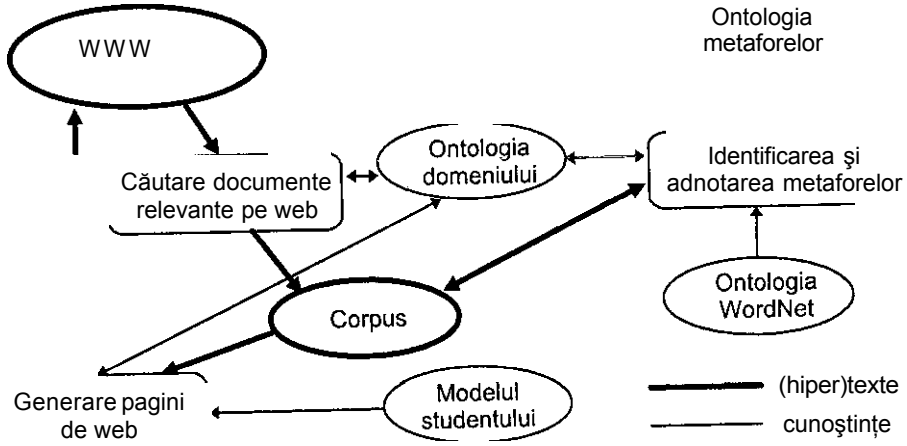


Figura **1**

Pentru identificarea şi adnotarea metaforelor a fost implementat un editor semantic specializat (fig.2) si un editor de concepte (fig.3).

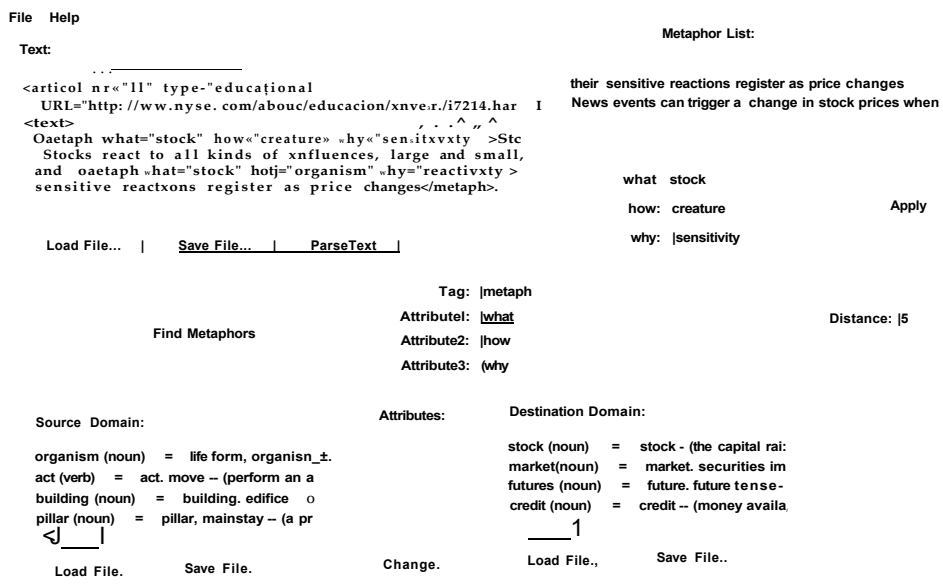


Figura **2**

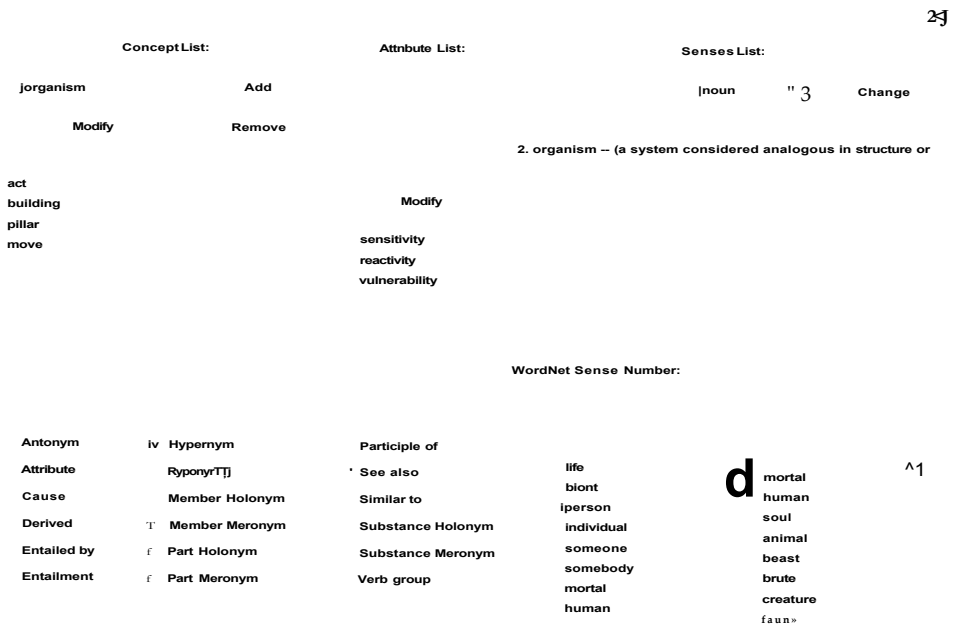


Figura 3

Modelul studentului este creat pe baza răspunsurilor la teste:

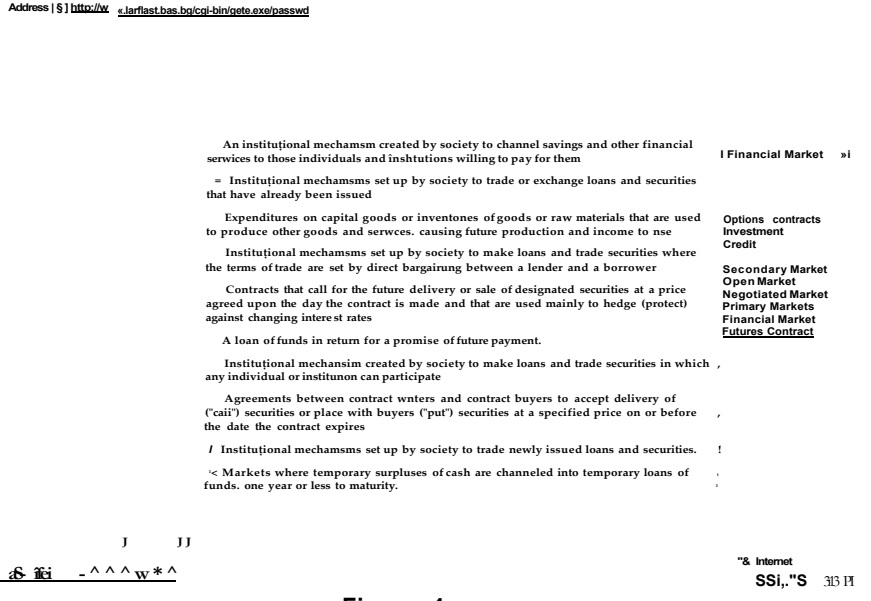


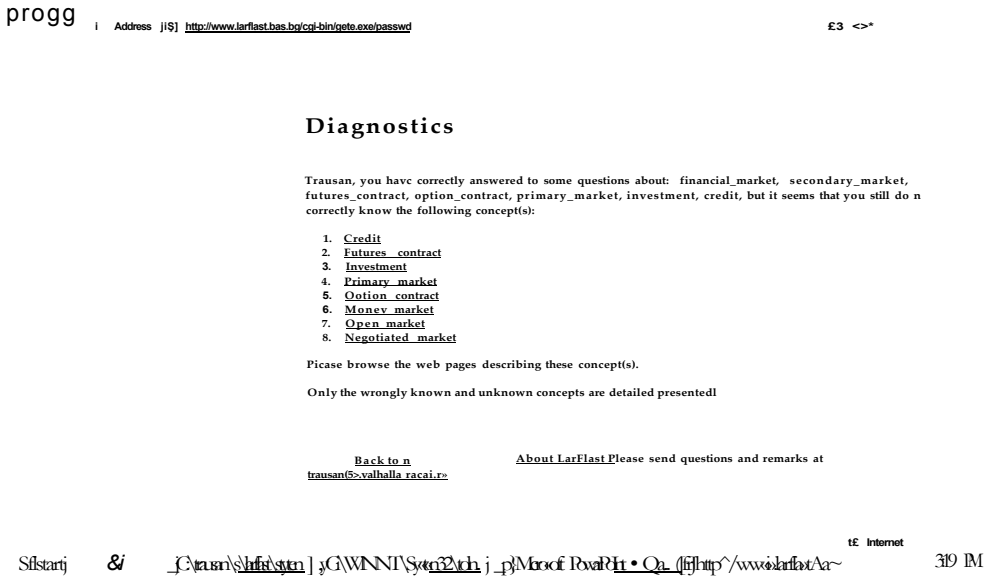
Figura 4

Paginile de web generate dinamic de modulul GenWeb, contribuția românească la proiectul Larflast, se constituie în structuri care reflectă structura conceptuală (ontologia) a domeniului considerat. Parcurgerea acestora poate fi văzută şi în corespondența cu facilitățile oferite de o Artă a memoriei [Cul94]. Din altă perspectivă, structurile trebuie concepute în ideea unei retorici specifice paginilor de web [Cli95], [THH95].

Sunt mai multe tipuri de pagini de web generate:

- pagini de diagnostic,
- pagini care definesc concepte, dau exemple de metafore şi care includ structuri de paggini web care reflectă ontologia domeniului,
- pagini din structurile de mai sus,
- pagini cu concordanțe în context.

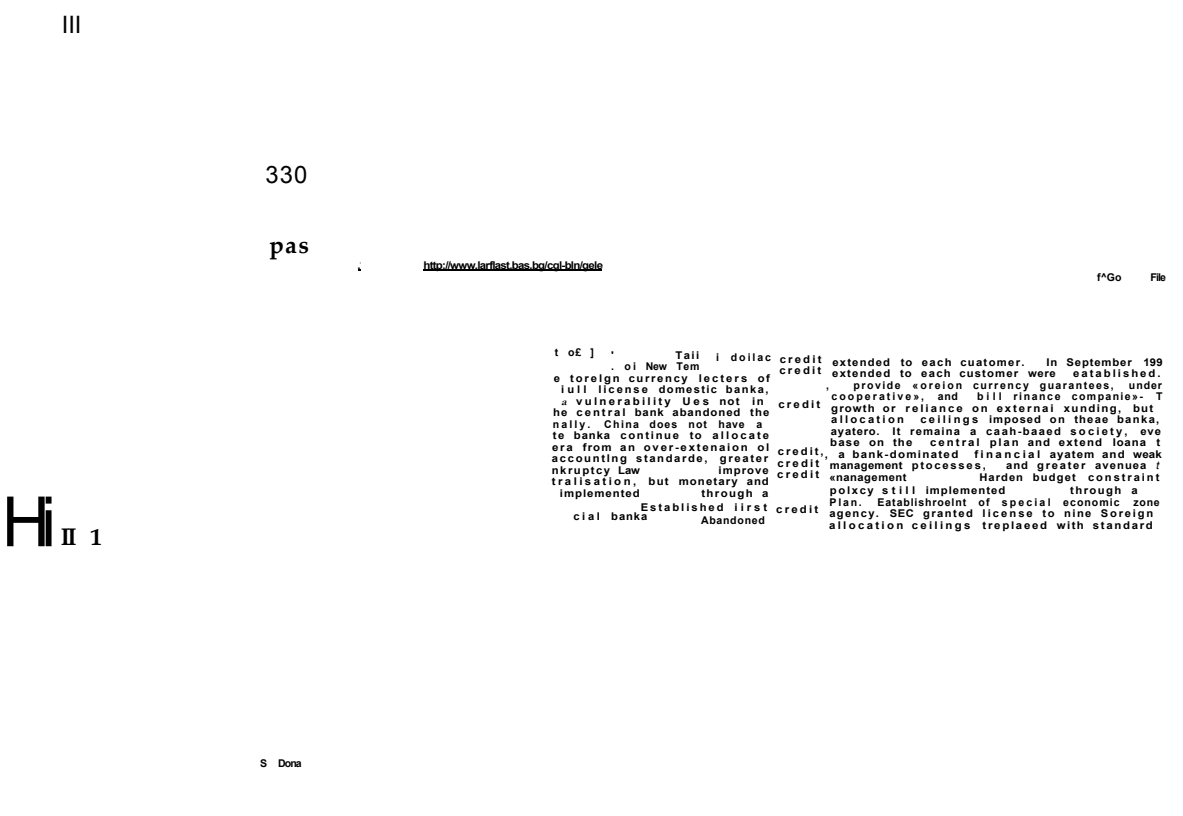
Aceste pagini sunt ilustrate în figurile următoare.



III

330

pas



Concluzii

în contextul dezvoltării explozive a numărului de docum
absolut necesară existența unor medii care să permită utilizato
scopul extragerii cunoştinţelor din texte şi structuri de docume
activitate trebuie sprijinită de ontologii, un rol foarte important a
ontologiilor de mari dimensiuni existente astăzi pe web. Un
integrează instrumente hermenofore cu ontologii într-o arl
utilizatorul trebuie să poată experimenta, să investigheze divers
textelor. Se poate spune că un mediu hermenofor înglobează si
de prelucrare a cunoştinţelor cu instrumente de prelucrare a te
specifice web.

Bibliografie

[ABK02] G. Angelova, S. Boytcheva, O. Kalaydjiev, Şt. Trăuşan-
Strupchanska, Adaptivity in a web-based CALL sy
Harmelen (ed.): ECAI 2002. Proceedings of the 15th Eur
on Artificial Intelligence, IOS Press, Amsterdam, 2002, p

[Bla85] L. Blaga, Trilogia culturii, Ed. Minerva, 1985

[BIM96] Black, J.B., McClintock, An Interpretation Construc
Constructivist Design, in B.G. Wilson (ed.), Const
Environments: Case Studies in Instructional Design, Edu
Publications, 1996.

[Cli95] Clibbon, K., Conceptually Adapted Hypertext For Learnir
CHI'95, http://www.acm.org/sigchi/chi95/Electronic/docu

[CTr01] Constandache, G.G., St. Trăuşan-Matu, Ontologia
calculatoarelor, Editura Tehnică, 2001.

[Cul94] Culianu, I.P., Eros şi magie în Renaştere; 1484, Nemira,

[Eng95] Engelbart, D.G., Toward Augmenting the Human Inte
our Collective IQ, CACM No.8, Vol.38, Aug. 95, pp. 30-3

• [ErK97] Ertl, B., Kraan, A.G., Internet-Based Learning Env
Constructivist point of view, Proceedings of RILW, Ilieni,

[Goo] http://www.google.com

[Gru96] Gruber, T., What is an Ontology, http://www.kr.org/top/d

[JoL83] Johnson-Laird, P.N., Mental Models - Towards a Co
Language, Inference, and Consciousness, Cambridge U

[LaJ80] Lakoff.G., Johnson, M., Metaphors We Live by, The Un
Press, 1980.

[Lar] LarFLaST, http://www-it.fmi.uni-sofia.bg/larflast/

[Mar97] Marcus, S., Empatie şi personalitate, Ed. Atos, 1997.

[MHD] J.C. Mallery, R. Hurwitz, G. Duffy, Hermeneutics, Encyclopedia of Artificial Intelligence, pp. 596-611.

[Noi86] C. Noica, Scrisori despre logica lui Hermes, Ed. Cartea Românească, 1986.

[OORG] http://www.ontology.org/main/papers/faq.html \

[Rad 91] Hypertext from Text to Expertext, McGraw Hill, 1991.

[Sow99] J. Sowa, Knowledge Representation: Logical, Philosophical and Computațional Foundations, Brooke Gole Publishing Co., Pacific Grove, CA, 1999, vezi şi [CTrOI].

[THH95] Thiring, M., Hannemann, J., Haake, J.M., Hypermedia and Cognition: Designing for Comprehension, Communications of the ACM, voi.38, no. 8, pp. 57-66, aug. 1995.

[TMC02] Şt. Trăuşan-Matu, D. Maraschi, S. Cerri, Ontology-Centered Personalized Presentation of Knowledge Extracted From the Web, în S. Cerri, G.Gouarderes (eds.), Intelligent Tutoring Systems 2002, Springer, Lecture Notes in Computer Science number 2363, pp. 259-269.

[Tra95] Şt. Trăuşan-Matu, Programe inteligente pentru asistarea învățării, în Revista Română de Informatică şi Automatică, voi.5, nr.4, 1995, pag. 7-16.

[TraOO] Şt. Trăuşan-Matu, *Metaphor Processing for Learning Terminology on the Web,* in S.A.Cerri (ed.), Artificial Intelligence, Methodology, Systems, Applications 2000, Springer-Verlag, ISBN 3-540-41044-9, 2000, pp.232-241

[Tra01] Şt. Trăuşan-Matu, Interfatarea evoluată om-calculator, Ed. MatrixRom, 2001.

[Vit 99] I. Vitanova, English for Finance, Understanding Money and Markets, http://wwwjt.fmi.uni-Sofia.bg/larflast/

[Wes97] D.West, Hermeneutic Computer Science, CACM, Vol.40, No.4, pp. 115-116, 1997, şi în [CTrOI].

[Wil96] B.G. Wilson (ed.), Constructivist Learning Environments: Case Studies in Instructional Design, Education Technology Publications, 1996

[Win87] T. Windgrad, Thinking machines: Can there be? Are we?, Report No. STAN-CS-87-1161, Stanford, 1987.

[WN] WordNet, http://www.cogsci.princeton.edu/~wn/

[Wri95] von Wright, G.H., Explicație şi înțelegere, Humanitas, 1995.

[WSD97] http://www.research.att.com/-suciu/workshop-papers.html

[XML] www.w3.org/xml

SECȚIUNEA III

TEHNOLOGII ALE LIMBAJULU

Experimente în vederea recunoaşt vorbitorului

Comeliu BURILEANU,
Universitatea „Politehnica" din Bucureşti, Spl.Independenței 3
cburileanu@messnet.pub.ro

Luigi BOJAN,

Graphco Technologies Inc., Newton, PA, USA

1. Introducere

Având în vedere funcția realizată şi concomitent, sarcin tehnologia vorbirii se poate clasifica în mai multe domenii [1, 2]:

- Recunoaşterea automată a vorbirii. Se bazează pe ana semnalului vocal şi are în vedere informația transmisă care "îl ascultă". Din această informație, maşina este extragă acele caracteristici ce îi vor permite să vorbeşte, ce vorbeşte, în ce fel şi în ce condiții.
- Sinteza automată a vorbirii. Se realizează răspunsu maşinilor către operatorul uman.
- Codificare/decodificare (analiză şi sinteză) a vorbiri tehnici de compresie a informației conținută în sem vederea unor prelucrări ulterioare specifice sarcinii de

Un domeniu interdisciplinar important, legat în mod esenț de recunoaştere şi sinteză automată ale vorbirii este cel al dialogulu

Termenul "comunicare om - maşină" pare forțat: maşina nu socială, nu are nici scop nici cultură. Ea nu poate acționa în lumea de a putea să răspundă corect la întrebări de genul: "ai putea să rog?". Ea nu este "conştientă" decât de propria sa "lume". Avem înt de a comunica cu maşinile? Au importanță intențiile lor, chiar dacă poate să-mi comunice sau să mă facă să ştiu o maşină?

Maşina îmi procură "uneltele" pentru a realiza o sarcină, proiectez noi obiecte (eventual, virtuale), ea mă aduce într-un univ permite să utilizez un mediu de programare împreună cu alți u pentru a lucra într-o manieră cooperantă în acelaşi mediu inform prezintă deci ca un *factor de interacțiune.* Ea trebuie să-mi furnize

muncă, unelte şi metode. Dar pentru aceasta, maşina trebuie adaptată sarcinii curente sau unor sarcini noi, să adopte un comportament "comprehensibil", să se arate "prietenoasă" etc. Paradoxul este deci evident: maşina trebuie să fie, dintr-un anumit punct de vedere, *socială* pentru a colabora eficace cu un utilizator în scopul îndeplinirii sarcinilor, din ce în ce mai complexe, care îi sunt încredințate.

Preocupările noastre în domeniul tehnologiei vorbirii au, între altele, scopul de a oferi mijloacele pentru o comunicare între om şi maşină prin mesaje vorbite [3]. Această comunicare este doar un aspect al dialogului. Rămâne în continuare deschisă problema definirii conceptelor şi cea a stabilirii unor strategii de dialog adecvate sarcinii de rezolvat.

Semnalul vocal conține o varietate de informații utile: ce se vorbeşte, cine vorbeşte, în ce fel şi în ce condiții. în cadrul recunoaşterii se pune problema identificării unui anumit tip de informații; de pildă, recunoşterea cuvintelor rostite înseamnă determinarea mesajului (ce se vorbeşte) indiferent (sau ajutându-se) de variabilitățile introduse de vorbitor (cine), maniera de a vorbi (în ce fel) şi zgomotul ambiental (în ce condiții). Putem particulariza afirmând că **recunoaşterea vorbirii** este procesul de transformare a semnalului acustic continuu produs de organul fonator uman într-o reprezentare discretă căreia i se poate ataşa o semnificație şi care, când e înțeleasă, poate fi folosită pentru a determina un răspuns.

Problemele majore pe care le ridică recunoaşterea automată sunt legate de

- discretizarea semnalului vocal care, din punctul nostru de vedere înseamnă *segmentare;*
- caracterul adecvat al răspunsului ce depinde de natura sarcinii de îndeplinit; modalitatea de prelucrare este irelevantă.

Proiectarea unui sistem de recunoaştere presupune câteva opțiuni fundamentale de abordare. Punctul de vedere adoptat poate viza prelucrarea unui semnal acustic ca oricare altul, poate ține seama de mecanismul producerii vorbirii, poate simula recepția senzorială, sau poate folsi modelul uman al percepției vorbirii.

Termenul de **recunoaştere a vorbitorului** desemnează orice aplicație de discriminare a persoanelor pe baza vocii acestora. Procedurile de recunoaştere se desfăşoară în două etape [4]:

- etapa de antrenare: colectarea de material vocal de la persoana care se doreşte a fi recunoscută;
- etapa de testare: compararea unui fragment de vorbire neidentificat cu datele provenite din antrenare şi luarea deciziei de recunoaştere.

Există două subclase de aplicații:

• **verificarea vorbitorului** îşi propune să determine dacă un fragment de semnal vocal aparține sau nu unui anumit vorbitor [5, 6, 7, 8]. Există doi parametri care caracterizează performanțele sistemului: respingerea adevăratului vorbitor şi

acceptarea unui impostor. Considerând un set de N vorbitori, informația (în biți) obținută este

presupunând probabilitatea de verificare a *priori* egală cu 0.5;

• **identificarea vorbitorului** are ca scop punerea în corespondență a unei voci necunoscute cu un vorbitor dintr-un set dat [9,10,11,12]. Pentru N vorbitori, informația (în biți) obținută este

considerând probabilitatea de identificare a *priori* egală pentru toți vorbitorii.

Rezultă că, potențial, un sistem automat de verificarea vorbitorului are performanțe mai bune.

O clasificare suplimentară a automatelor de recunoaştere are în vedere natura sarcinii de îndeplinit şi se reflectă în complexitatea sistemului [13]:

- sisteme de recunoaşterea vorbitorului **dependente de text** - textul utilizat în faza de antrenare este acelaşi cu cel de testare;
- sisteme **independente de text** - indiferent de materialul vocal avut la dispoziție.

Setul de vorbitori vizat poate impune, de asemenea, o clasificare a automatelor:

- *"set închis"* - pentru procesul de identificare descris ca mai sus;
- *"set deschis* - în cazul identificării există posibilitatea ca vocea necunoscută să nu aparțină niciunuia dintre vorbitorii din setul dat, numărul de decizii posibile fiind în acest caz $N+1$. Identificarea pe "set deschis" devine astfel o combinație a proceselor de verificare şi identificare.

2. Reprezentarea parametrică

Variabilitățile pronunțării pentru diverşi vorbitori, sau la un acelaşi vorbitor, la momente de timp diferite, constituie una dintre dificultățile majore ale sarcinii de recunoaştre a vorbitorului. Deosebirile de vorbire depind de dialect, context, stil de exprimare, stare emoțională etc. Mai mult, în opinia noastră, aşa cum vom încerca să argumentăm mai departe, *limba în care se vorbeşte* impune deosebiri de abordare şi diferențe ale performanțelor automatului [14].

Din acest motiv, alegerea judicioasă a **caracteristicilor acustice** care vor fi utilizate în procesul de recunoaştere este deosebit de importantă:

- să diferențieze vorbitori diferiți dar să fie tolerante pentru acelaşi vorbitor;
- să fie uşor măsurabile din semnalul vocal;
- să fie stabile în timp;
- să nu fie susceptibile de a fi contrafăcute de potențiali impostori.

Având în vedere cerințele formulate mai sus, am decis utilizarea *parametrilor cepstrali.*

Anumite abordări ale prelucrării semnalului vocal presupun adoptarea unor decizii fundamentale de dezvoltare a analizei: considerarea unui model de producere a vorbirii având ca prototip aparatul fonator uman, separarea efectelor sursei vorbirii de comportarea tractului vocal propriu-zis, o serie de aproximări care să facă analiza eficientă în condiții normale de procesare [15]. Variația (lentă) în timp a formei tractului vocal este aproximată printr-o serie de secvențe de durată suficient de mică pentru a presupune forma invariantă: este ceea ce se numeşte "analiza în timp scurt". Dacă, în plus, în aceste durate "scurte" de timp se presupune că tractul este caracterizat în mod esențial de frecvențele sale de rezonanță, se ajunge la un model al cărui parametri se pot deduce prin rezolvarea unui sistem de ecuații liniare. Deşi aproximările avute în vedere par destul de restrictive, *analiza prin predicție liniară* (LPC) dă rezultate deosebite pentru că semnalul vocal are o redundanță deosebită; este motivul pentru care metoda ne permite să aproximăm un eşantion de semnal printr-o combinație liniară (deci este liniar predictibil) dintr-un număr de eşantioane precedente. Desigur, principiile în sine ale metodei nu sunt noi; ele au permis însă, în decursul ultimilor ani, evoluția spre metode mai sofisticate [16,17].

Nici principiile *analizei cepstrale* (analiză care, aşa cum vom arăta, se poate baza pe rezultatele analizei LPC) nu sunt noi: se dezvoltă un mecanism care să permită decelarea mai amănunțită a influențelor diverselor elemente ale organului fonator. O serie de presupuneri fundamentale de abordare se păstrează (modelarea producerii vorbirii în maniera aparatului fonator uman, analiza "în timp scurt"); dar separarea efectelor excitației glotale, tractului vocal şi radiației buzelor poate fi făcută într-o modalitate care ține seama mai detaliat de fiecare efect în parte [18, 19].

în concluzie, presupunerile fundamentale care stau la baza parametrizării propuse sunt:

- efectele excitației tractului vocal şi ale tractului propriu-zis pot fi separate;
- tractul vocal este invariant pe durate scurte de timp, ceea ce are drept rezultat obținerea unui model descris de un sistem liniar ai cărui parametri variază lent în timp (constanți "în timp scurt").

Fundamental pentru modul în care concepem abordarea analizei semnalului este asimilarea *analizei* cu *parametrizarea* semnalului şi, în consecință, cu *compresia* sa. Alegerea parametrilor a avut în vedre şi considerente pragmatice:

- complexitatea prelucrării;
- gradul de compresie,
- tipul de aplicație,
- în ce măsură parametrii sunt semnificativi şi robuşti.

O primă variantă a schemei bloc care descrie funcționarea sistemului de recunoaşterea vorbitorului este prezentată în fig. 1. Blocul de *preprocesare* presupune filtrarea şi achiziția semnalului în condiții normale pentru orice sistem de recunoaştere. în această secțiune vom descrie obținerea *cepstrului* pornind de la *analiza LPC,* iar în secțiunea următoare vom descrie principiile *cuantizării vectoriale* şi deci procedura de recunoaştere propriu-zisă.

Fie semnalul vocal presupus a fi convoluția unei excitații şi a funcției de transfer a tractului vocal:

$$s(t)=e(t)*v(t) \tag{3}$$

Analiza homomorfică care duce la obținerea *cepstului* presupune aplicarea unui operator neliniar" *H"*

$$s(n)—^—^s(n) \tag{4}$$

în care *s(n) va* fi numit *cepstrul complex* asociat semnalului *s(n).*

Prin definiție

(5)

Astfel, cepstrul complex asociat semnalului devine

$$S\{n)=e(n)+\quad v(n) \tag{6}$$

ceea ce permite separarea componentelor printr-o "filtrare temporală" aplicată cepstrelor

$$s\{n)- \xrightarrow{} s\{n\} \qquad \begin{matrix} \xrightarrow{H\sim'} e(n) \\ \xrightarrow[H\sim']{} v(n) \end{matrix} \tag{7}$$

Obținerea parametrilor cepstrali se poate realiza ținând seama de câteva proprietăți ale cepstrului.

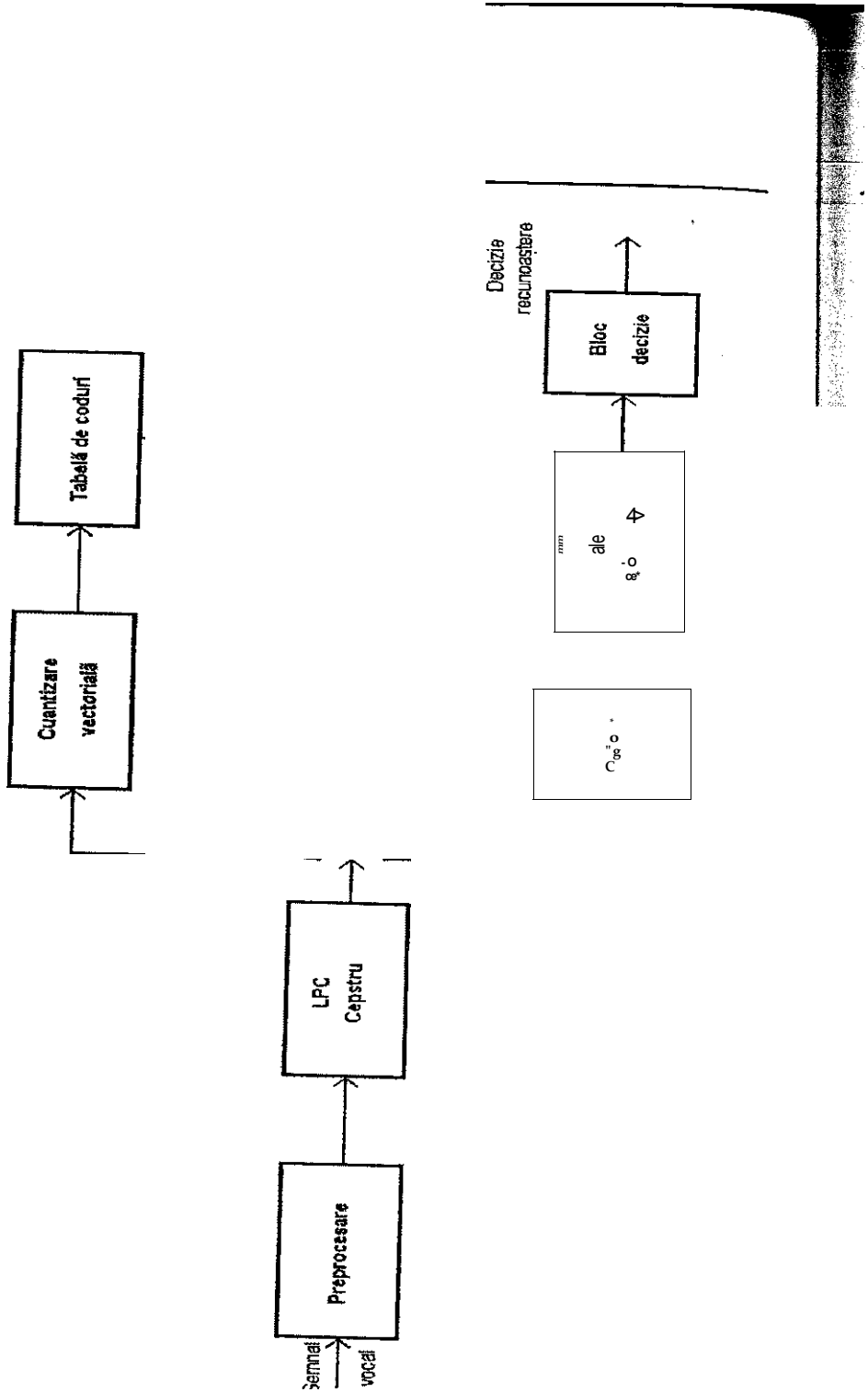


Figura **1.** Un sistem de recunoaştere a vorbitorului - schema de principiu

Fie c(n) partea pară a cepstrului complex al semnalului

$$c(n)=[s(n)+s\{-n)]/2$$

Secvența $c(n)$ se numeşte *cepstrul real* al semnalului ^

este o secvență cauzală - ca şi $s(n);$ rezultă

$$s(n)=c(n)-\begin{cases} 0 & pentru \quad n<0 \\ 1 & pentru \quad n-0 \\ 2 & pentru \quad n>0 \end{cases}$$

Cum transformata "z" a unei secvențe cauzale e determ partea reală a transformatei sale Fourier, rezultă

$$c(n) = \text{^-}lln\backslash_s(e^M\}e^{j\omega} \quad d(H$$

Vom prefera calculul coeficienților cepstrali din coeficie prin predicție liniară (LPC) conform relațiilor recursive:

$$c(l) = -a,$$

$$\underset{n=l}{} \quad \underset{i}{} \bullet a_n \bullet c[i-n)$$

Figura 2 prezintă evoluția coeficienților cepstrali pentru una masculină.

Materialul vocal a fost achiziționat folosind un mi (considerat fără zgomot) şi a fost eşantionat cu frecvența de 8 analiză au lungimea de 240 ms, cu o suprapunere de 160 predicție liniară s-a efectuat cu ordinul de predicție p = *10,* ia coeficienților de predicție liniară s-a folosit algoritmul Levinso observație este aceea că modulul amplitudinii coeficienților es ordinul acestora. Pentru coeficienții de ordinul 5-10, evoluția co tinde să devină uniformă. Amplitudinea redusă a acestora anu dificultăți de estimare în condiții de zgomot.

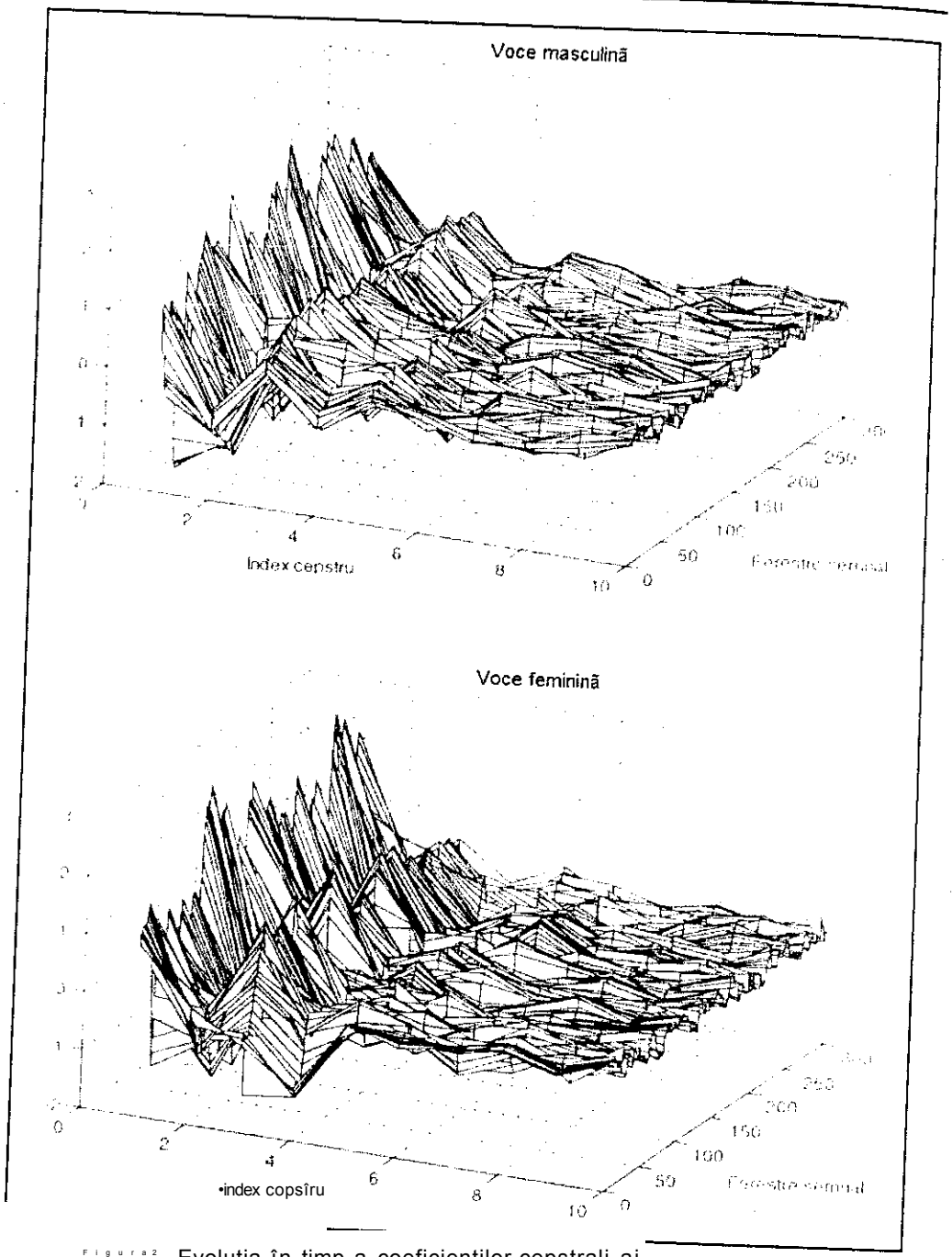


Figura 2 - Evoluția în timp a coeficienților cepstrali ai semnalului vocal

în scopul unei aprecieri calitative, fig. 3 prezintă, di
cepstrali în planul c(1)-c(2), pentru aceiaşi doi vorbitori (masculin
observa distribuția diferită a principalilor coeficienți cepstrali pentr
remarcă o concentrare a coeficienților în anumite zone ale planlu

Voce masculină

o
c(1)

Voce feminina

c(1)

Figura **3.** Reprezentarea coeficienților cepstrali în pla

în fig. 4 este prezentată distribuția parametrilor cepstrali corespunzători unui semnal vocal compus numai din vocalele limbii române. Ordinul analizei cepstrale este *p = 12.* Reprezentarea grafică s-a făcut numai în planul *c(1) - c(2).* Se observă faptul că vocalele sunt relativ uşor separabile în spațiul cepstral, într-o configurație asemănătoare celei din spațiul formantic. Această analiză oferă premize interesante şi pentru recunoaşterea vorbirii în limba română.

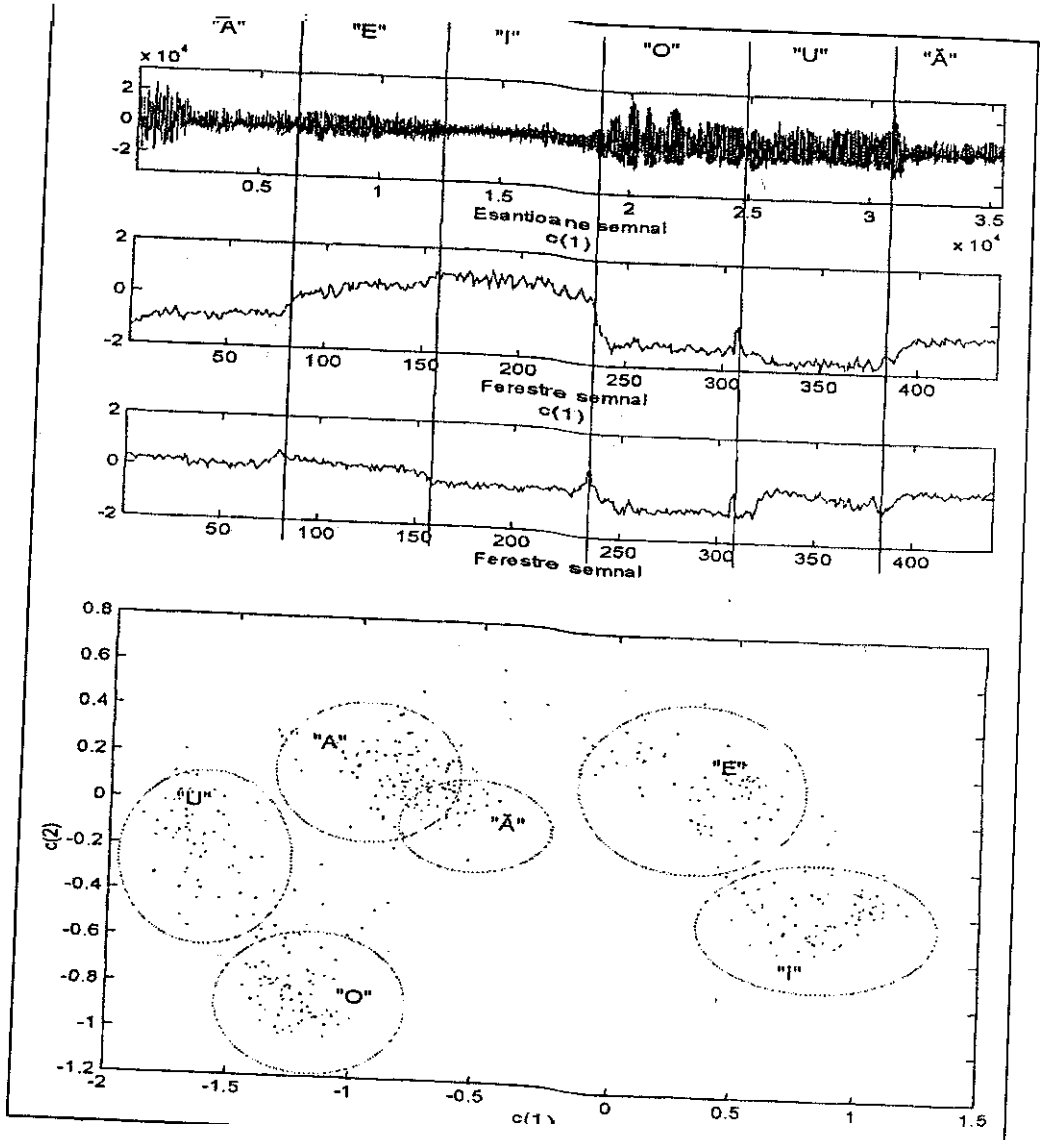


Figura 4 - Semnal compus din vocalele i^ : parametru cepstrali corespunzători

3 Cuantizarea vectorială

Din punctul de vedere al sistemelor de recunoaştere a persoană produce în timpul vorbirii o secvență de vectori de pa caracterizează atât vorbitorul cât şi cuvintele pronunțate. Pentru un suficient de lung, ne aşteptăm ca datele achiziționate să acopere într-un mod care depinde mai mult de caracteristicile vorbitorului ceea ce a pronunțat. Se face pesupunerea că, având ia disp suficient de date, se poate genera un model al vorbitorului care s un proces de recunoaştere [20, 21].

Principiul cuantizării vectoriale este aplicat în sensul volum mare de vectori acustico-fonetici, reprezentând material vo către un vorbitor, într-un set restrâns de vectori denumit *tabelă d centroizi).* în etapa de antrenare, partiționarea spațiului acop spectrali este făcută astfel încât media distanțelor minime dintr cepstral şi cel mai apropiat centroid să fie minimizată. în etapa de de vectori provenind de la un vorbitor necunoscut, este codat uti vectori corespunzătoare vorbitorului vizat. Distorsiunea totală me în decizia de recunoaştere [22].

Fie $\{X_n\}$ ansamblul de N versiuni cunoscute ale vectorulu

Fie {G^} o partiție a acestui ansamblu în K clase; o cla

g_k elemente, astfel ca

Notăm cu $Xf^$ cuvântul prototip ("centroid", "vector-cod") a

Distanța medie între centroizi este

$$— \qquad 1$$

Distanța medie între vectorii dintr-o aceeaşi clasă, p clasele este

Raportul = reprezintă calitatea partiției

Algoritmul utilizat pentru găsirea centroizilor este atunci următorul:

• dacă cei K centroizi sunt aleşi la întâmplare, clasele sunt constituite asociind fiecare vector X centroidului cel mai apropiat:

$$X_n eG_k \quad dacă \quad D(x_{nf}Xt^{p}) < D(x_{nf}X^{\wedge}) \quad Mi \pm k \qquad (15)$$

• se iterează găsirea centroizilor căutând în fiecgre clasă k vectorul $X^{\wedge}$ care are distanța față de vectorul cel mai depărtat al clasei minimă:

$$x\ f\ s\ X\ J\ f\ ' \quad dacă \quad \max_{m} D(xf \backslash x^{\wedge}) \quad eminimă \qquad (16)$$

• această procedură e iterată până când centroizii sunt stabilizați.

Prezentăm în fig. 5 un exemplu de cuantizare vectorială folosind algoritmul Linde-Buzo-Gray (LBG). Vectorii cuantizați sunt coeficienții cepstrali de predicție liniară. Pentru reprezentarea în plan s-a ales sistemul de coordonate *c(1) - c(2).* Dimensiunea tabelei de centroizi aleasă este *M* = 8. Se poate observa cum, în urma operației de optimizare, centroizii tind să "acopere" întregul spațiu ocupat de vectori. în mod evident, eroarea de cuantizare scade pe măsură ce dimensiunea tabelei de centroizi creşte.

Pe parcursul algoritmului se pot utiliza diverse strategii de divizare. De exemplu, dacă după o operație de divizare şi reclasificare, una dintre clase devine subpopulată sau chiar vidă, o alta va fi divizată la pasul următor, pentru a menține constant numărul total de clase. Se pot folosi următoarele criterii de alegere a clasei care va fi divizată: clasa care posedă cel mai mare număr de elemente, clasa care prezintă distorsiunea totală cea mai mare, clasa care prezintă distorsiunea medie cea mai mare. Folosind această structură arborescentă, clasificarea unui vector se poate efectua prin asocieri succesive, printr-o parcurgere a claselor găsite pentru fiecare nivel de divizare. în aplicațiile care necesită o acuratețe de clasificare ridicată, se preferă o metodă de clasificare prin căutare exhaustivă.

Figura 5 TEvi^alg^^

"•" - vectori cepstrali; * - centroizi

4. Rezultate obținute

Un aspect important în proiectarea automatelor de recunoaştere a vorbitorului (eventual independent de text) îl reprezintă posibilitatea de evaluare a performanțelor acestora. Pentru a putea evalua un astfel de automat cu o precizie acceptabilă este nevoie de o bază de date corespunzătoare [23]. O astfel de bază de date trebuie să îndeplinească următoarele cerințe:

- să cuprindă material vocal achiziționat de la cât mai mulți vorbitori (de preferat, de ordinul zecilor sau sutelor);
- să conțină, eventual, dialecte diferite;
- să conțină fraze cât mai variate;
- frazele să fie rostite de mai multe ori, la intervale de timp
- pentru evaluare în condiții reale (de exemplu transmisie telefonică), materialul vocal trebuie să fie achiziționat prin intermediul mai multor aparate telefonice, în decursul mai multor legături, de preferat la distanțe diferite [24, 25].

Proiectarea şi construirea unei astfel de baze de date este o sarcină dificilă.

Am folosit mai multe baze de date: internaționale, oarecum standard pentru procedurile de recunoaştere - TIMIT" şi "YOHO", precum şi o bază de date proprie, în română şi engleză - "DiSPPALL".

Baza de date "TIMIT". conține eşantioane de voce provenind de la 630 de vorbitori, fiecare pronunțând 10 fraze. Experimentele descrise în lucrare au fost efectuate pe secțiunea TEST, care conține 168 vorbitori. Cele 10 fraze sunt: două fraze de calibrare (SA), cinci fraze compacte din punct de vedere fonetic (SX) şi trei fraze variate contextual (SI). în experimente s-au folosit frazele SA şi SX în faza de antrenare şi frazele SI în cea de testare. Pentru evaluarea efectelor zgomotului telefonic în algoritmii de recunoaştere a vorbitorului, s-a folosit o variantă a bazei de date numită "NTIMIT". Aceasta conține aceiaşi material vocal ca şi baza "TIMIT" cu deosebirea că acesta a fost transmis prin intermediul rețelei telefonice. Transmisia s-a făcut folosind un echipament de simulare a tractului vocal uman, în legături telefonice reale, la diferite distanțe.

Baza de date "YOHO" cuprinde fraze rostite de 138 de vorbitori (106 bărbați şi 32 femei), iar vocabularul folosit constă din numere de două cifre rostite în grupuri de câte trei. Pentru fiecare vorbitor am folosit 4 sesiuni de antrenare de câte 24 de enunțuri şi 10 sesiuni de verificare de câte 4 enunțuri.

Baza de date proprie "DiSPPALL" [26] cuprinde materialul vocal de la 26 de vorbitori (23 de bărbați şi 3 femei) cu vârsta ce variază de la 21 la 50 de ani. Fiecare vorbitor în parte s rostit 31 de fraze: 11 fraze echilibrate din punct de vedere fonetic au fost folosite pentru antrenare şi 20 de fraze pentru verificare: 5 enunțuri de bază repetate de câte 4 ori. Frazele de verificare au fost înregistrate în două sesiuni diferite, 5 enunțuri de bază fiind repetate de două ori în fiecare sesiune. Prima sesiune de verificare a fost înregistrată în acelaşi timp cu sesiunea de antrenare, iar sesiunea a doua a fost înregistrată după două-trei săptămâni, înregistrările s-au făcut cu un microfon de tip "head-set" într-o cameră cu zgomot ambiental normal: spre deosebire de baza "YOHO", baza "DiSPPALL" conține material vocal alterat de zgomot pentru a face condițiile de test mai dificile şi mai apropiate de o situație reală de recunoaştere a vorbitorilor

în experimentele de verificare a vorbitorului, o frază de test este comparată cu referința vorbitorului a cărui identitate se dorește verificată, calculându-se o distorsiune totală medie. Dacă aceasta este mai mică decât un prag dat, vorbitorul este considerat acceptat, altfel el este respins. Există două tipuri de erori asociate procesului de verificare: respingerea utilizatorului căruia îi aparține referința (denumită eroare de tip I) şi acceptarea unui impostor (eroare de tip II) [27]. Fiecare frază de test este comparată cu referințele corespunzătoare tuturor vorbitorilor din baza de date aleasă pentru test. Pragurile de decizie nu sunt fixate a *priori* ci se determină distanța medie totală pentru care eroarea de tip I este egală cu cea de tip II *("rata-erorii-egale').* Valoarea corespunzătoare a erorii este considerată rezultatul final al procesului de evaluare. în fig. 6 sunt prezentate rezultatele procesului de verificare a vorbitorului, folosind cuantizarea vectorială, utilizând baza de date "TEST/TIMIT". Ordinul de predicție (şi implicit dimensiunea vectorilor cepstrali) este *P -10* iar dimensiunea tabelei de centroizi, *M =64* . Ca distanță vectorială s-a folosit distanța euclidiană ponderată.

$$d(v_u, v_b) = -lj^{y}£_J(v_{uj}-v_{bj})^2 \quad (17)$$

unde *s]* este varianta componentei *j* calculată pe întreg setul vectorilor de antrenare. Ca metodă de cuantizare vectorială s-a folosit algoritmul LBG modificat.

Sunt evidente tendințele contrare ale erorilor de tip I, respectiv II. Rata erorii-egale pentru evaluarea de mai sus este 6.8%, corezpunzând unui prag de decizie egal cu 2.8. în funcție de aplicația concretă, pragul de decizie se poate stabili a *posteriori* la o altă valoare, adecvată scopului propus. Spre exemplu, dacă se dorește limitarea acceptării impostorilor la 2%, respingerea adevăraților utilizatori va fi de 19.7%. Reciproc, pentru o eroare de respingere a utilizatorilor reali de 2%, acceptarea impostorilor va fi de 12.9%.

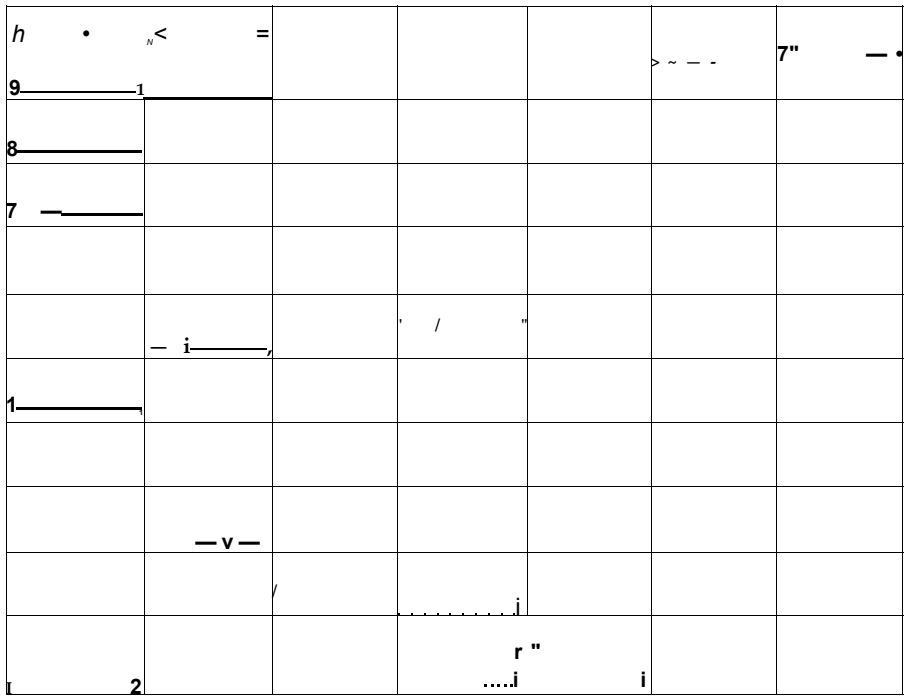


Figura **6.** Eroarea de verificare a unui sistem de recunoaştere a vorbitorului utilizând cuantizarea vectorială

în experimentele de identificare a vorbitorului, fiecare frază de test provenind de la un vorbitor considerat necunoscut este comparată cu referinţele fiecărui vorbitor din baza de date aleasă pentru test. Referinţa asociată cu cea mai mică distorsiune totală medie faţă de fraza de test este considerată ca aparţinând vorbitorului identificat. în funcţie de corespondenţa dintre apartenenţa frazei de test şi a referinţei aceluiaşi vorbitor sau unor vorbitori diferiţi, se decide dacă rezultatul procesului de identificare este adevărat sau fals. Eroarea de identificare este calculată ca raportul dintre numărul de identificări incorecte şi numărul total de identificări [28, 29, 30].

5. Utilizarea frecvenţei fundamentale în recunoaşterea vorbitorului

Frecvenţa fundamentală poate fi utilizată ca parametru discriminato suplimentar în conjuncţie cu algoritmi de cuantizare vectorială a vectorilor cepstrali

Frecvenţa fundamentală F_0 sau *perioada fundamentală* T_0 (cunoscută ş sub numele de "pitch"), constituie un parametru important al vocii umane, care îş găseşte utilizări practice în multe domenii ale procesării vorbirii. încercări d utilizare a frecvenţei fundamentale în procesul de recunoaştere a vorbitorului s cunosc încă de la începutul anilor 70, aceasta fiind pusă în corespondenţă directă cu prozodia. Majoritatea acestor experimente s-au desfăşurat utilizând sisteme d recunoaştere dependente de text şi metode de aliniere temporală. Sistemele d recunoaştere a vorbitorului independente de text bazate exclusiv pe frecvenţ fundamentală nu au dat rezultate satisfăcătoare.

Ideea prezentată în secţiunea de faţă este aceea de a folosi frecvenţ fundamentală ca *un parametru discriminator suplimentar,* în conjuncţie cu algoritm de cuantizare vectorială a vectorilor cepstrali [31]. Justificarea teoretică a aceste abordări rezidă în primul rând în modelul predicţiei liniare aplicat semnalului voca care presupune, aşa cum am arătat o separare clară între sursa de semnal ş tractul vocal. De asemenea, am arătat în secţiunea 2 că analiza cepstrală folosit pentru extragerea vectorilor cepstrali este un proces de deconvoluţie, coeficienţ cepstrali obţinuţi caracterizând în mod exclusiv tractul vocal. Ca atare, utilizarea c date de intrare în acelaşi sistem atât a vectorilor cepstrali cât şi a frecvenţe fundamentale nu reprezintă o abordare redundantă.

Cerinţele de bază ale unui algoritm de extragere a frecvenţei fundamental sunt: acurateţea de estimare (evitarea armonicilor), robusteţea decizie sonor/nesonor, insenzitivitatea la zgomot, volumul de calcule minim. Se cunos numeroşi algoritmi de estimare a frecvenţei fundamentale (AMDF, Dubnowsk Rabiner, SIFT, etc), fiecare prezentând avantaje şi dezavantaje. Trebuie arăta faptul că, din cauza, în principal, comportării nestaţionare a semnalului voca niciunul din algoritmii cunoscuţi nu este considerat perfect. Cu alte cuvinte, s acceptă ideea existenţei erorilor atât în luarea decizei sonor/nesonor cât şi î obţinerea valorilor propriu-zise ale frecvenţei fundamentale. în experimentel prezentate mai jos s-a folosit algoritmul Rabiner, considerat ca fiind unul dintr cele mai robuste.

Ideea introdusă este aceea de a utiliza frecvenţa fundamentală în scopu unei clasificări grosiere a potenţialilor candidaţi, atât pentru sarcina de verificare vorbitorului, cât şi pentru cea de identificare. în cazul verificării, scopul propus est acela de a reduce eroarea de tip II, prin eliminarea vorbitorilor a căror frecvenţ fundamentală nu "corespunde" cu cea a vorbitorului de referinţă. în cazul sarcin

de identificare, se doreşte reducerea numărului de candidaţi posibili, fără a afecta acurateţea de identificare. Aceasta poate conduce la o reducere majoră a volumului de calcule, dat fiind că estimarea frecvenţei fundamentale se face o singură dată pentru fiecare vorbitor şi este mai puţin consumatoare de timp decât clasificarea vectorială.

Având în vedere considerentele de mai sus, schema de principiu a sistemului de recunoaştere a vorbitorului modificat prin introducerea frecvenţei fundamentale ca parametru discriminator este prezentată în fig. 7.

Un aspect important în utilizarea frecvenţei fundamentale în aplicaţiile de recunoaştere a vorbitorului îl reprezintă alegerea formei de prelucrare a datelor furnizate de estimator. "Conturul de pitch", reprezentând evoluţia în timp a parametrului F_0, deşi utilizat în sisteme de recunoaştere a vorbitorului dependente de text, conţine un volum de date dificil de utilizat în operaţii de discriminare. în consecinţă, s-a încercat o reducere a datelor la câţiva parametri statistici. Au fost investigate patru valori statistice derivate din conturul de pitch: valoarea medie, valoarea maximă, valoarea minimă şi dispersia (deviaţia standard). Pentru fiecare vorbitor, s-au Calculat aceste valori pe ansamblul materialului vocal disponibil. Ca parametru de discriminare a fost utilizat raportul valorilor statistice de mai sus

$$P_{medie} = \frac{\text{media } F_0 \text{ pentru antrenare}}{\text{media } F_0 \text{ pentru test}} \qquad (18)$$

tratându-se în mod similar toate celelalte valori statistice (maximă, minimă, dispersie). Pentru a evalua utilitatea acestori parametri în procesul de discriminare, s-a determinat distribuţia fiecăruia atât pentru frazele pronunţate de aceiaşi vorbitori (intra-vorbitor) cât şi pentru toate combinaţiile de fraze pronunţate de vorbitori diferiţi (inter-vorbitor).

Modul de discriminare a vorbitorilor este următorul: fixându-se un prag θ, dacă

(19)

vorbitorul este rejectat şi nu se execută clasificarea vectorială. în caz contrar, vorbitorul este considerat potenţial candidat şi urmează procesul de clasificare prin cuantizare vectorială.

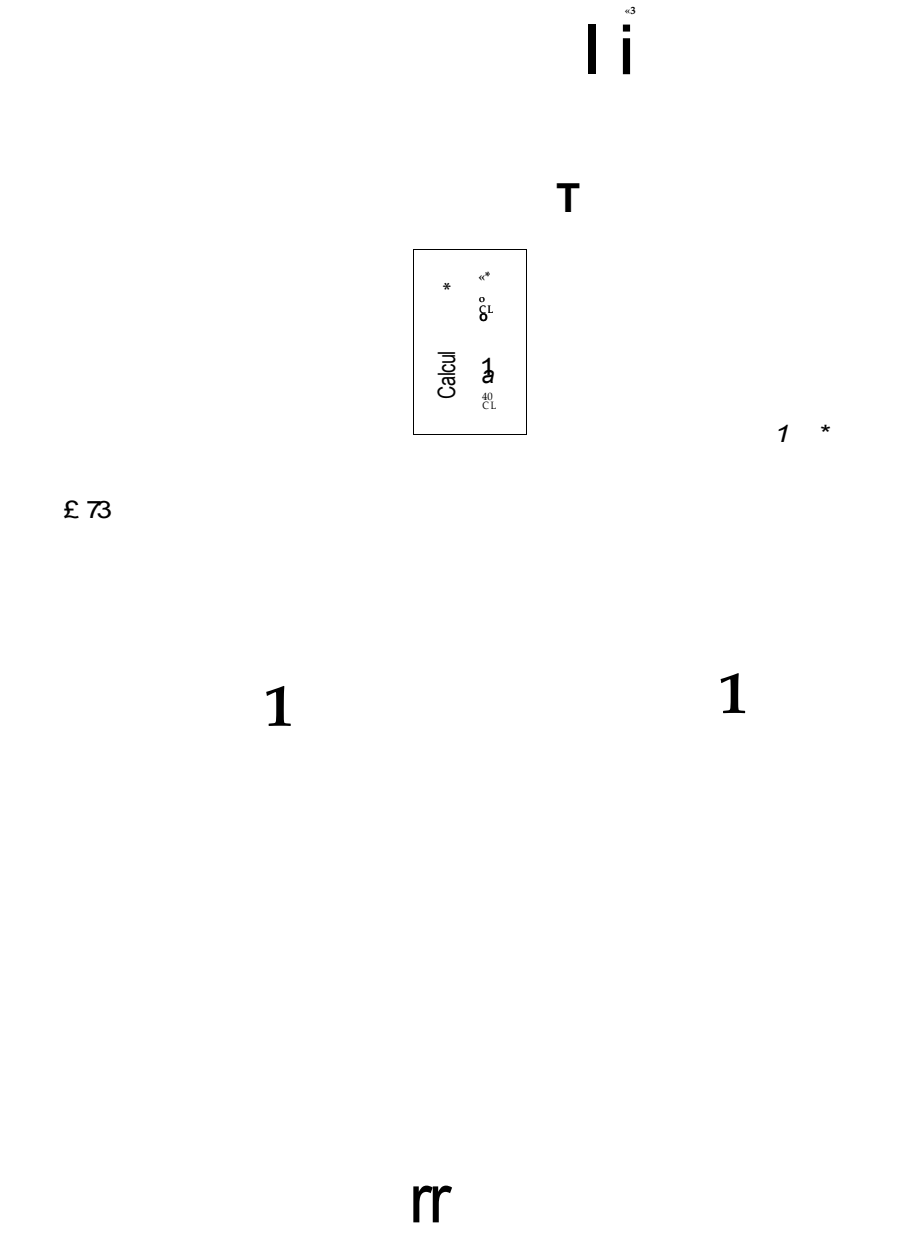


Figura 7. O variantă a sistemului de recunoaştere a vorbitorului - schema de principiu

Utilizând elementul de discriminare descris mai sus s-au obținut îmbunătățiri importante atât în procesul de verificare a vorbitorului cât şi în cel de identificare. Rezultatele obținute pentru 14 coeficienți cepstrali şi 128 centroizi sunt prezentate în tabelul 1.

Tabelul 1

e	Neutilizat	0.30	0.25	0.20	0.15	0.10
EER la verificare (%)	6.3	6.1	5.3	3.9	2.7	6.5
Eroarea de identificare (%)	6.2	6.2	5.9	5.6	5.5	9.4
Candidați identificare (%)	100	57.2	49.1	43.4	32.3	26.5

Cele mai bune rezultate s-au obținut pentru *8 = 0.15*, caz în care eroarea de verificare obținută este de aproape 2.5 ori mai mică decât în cazul folosirii doar a clasificării vectoriale. în cazul identificării, deşi îmbunătățirile de acuratețe nu sunt impresionante, cel mai important rezultat îl reprezintă'reducerea numărului candidaților, cu peste 65%. Pentru valori ale lui *8* mai mici decât 0.10, se observă o degradare abruptă a performanțelor de verificare şi identificare, ceea ce indică faptul ca variația intra-vorbitor a frecvenței fundamentale medii este mai mare decât acest prag.

6. Concluzii

Lucrarea de față se ocupă de un aspect bine delimitat al tehnologiei vorbirii şi anume recunoaşterea vorbitorului ca parte integrantă a recunoaşterii automate şi mai departe a dialogului om-maşină. Tipurile de probleme care apar sunt similare pentru întreg domeniul recunoaşterii automate.

Am precizat presupunerile fundamentale care au stat la baza analizei propuse (în special opțiunea de a aborda proiectarea ținând seama de mecanismul producerii vorbirii); insistăm asupra faptului că aceste abordări nu sunt obligatorii, ci constituie alternative care au avantaje şi dezavantaje.

S-au trecut în revistă etapele esențiale ale procedurilor de recunoaşterea vorbitorului: achiziția semnalului vocal, prelucrarea acustico-fonetică, recu-noaşterea propriu-zisă.

Am subliniat importanța parametrizării semnalului vocal. Analiza cepstrală care a fost aleasă pentru reprezentarea parametrică a semnalului vocal este legată de opțiunile fundamentale de analiză: separarea efectelor sursei de semnal şi ale tractului, separarea efectelor diverselor porțiuni din tractul vocal, analiza "în timp scurt"

Am utilizat cuantizarea vectorială ca metodă de recunoaştere. Sunt prezentate o parte dintre rezultatele experimentelor realizate. Subliniem importanța

utilizării unor baze de date specifice şi, în consecință, am acordat spațiu prezentării acestora.

O contribuție pe care o considerăm interesantă la îmbunătățirea performanțelor recunoaşterii vorbitorului o constituie utilizarea frecvenței fundamentale ca parametru discriminator grosier. Sunt prezentate o serie de rezultate care probează în ce mod anumite performanțe sunt superioare abordării "clasice".

O parte dintre rezultatele obținute sunt susceptibile de a fi generalizate pentru recunoaşterea vorbirii în limba romană [32] (de pildă, coeficienții cepstrali pentru foneme ale limbii române). De asemenea, utilizarea frecvenței fundamentale apropie recunoaşterea vorbitorului de o anumită dependență de limba în care sunt rostite frazele de antrenare şi de test.

Referințe bibliografice

[1] M.Drăgănescu, C.Burileanu, coordonatori (1986). Analiza şi sinteza semnalului vocal - Editura Academiei Române, Bucureşti.

[2] M.Dragănescu, G.Stefan, C.Burileanu (1991). Electronica funcțională - voi. Editura tehnică, Bucureşti, ISBN 973-31-0290-3.

[3] G. Yu and H. Gish (1993). Identification of Speakers Engaged in Dialog, Proc. of IEEE Int. Conf. Acoust, Speech, Signal Processing, Voi.II, p. 383-386.

[4] Sadaoki Furui (1994). An Overview of Speaker Recognition Technology, Proc of ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, p. 1-9.

[5] Y. Bennani, P. Gallinari (1994). Connectionist Approaches for Automat Speaker Verification - ESCA Workshop on Speaker Recognitior Identification and Verification, p. 95-103.

• [6] M. Hanah s.a. (1994). The Role of the Reference Template in Speake Verification - ESCA Workshop on Speaker Recognition, Identification an Verification, p. 181-184.

[7] Chi-Shi Liu; Hsiao-Chuan Wang; Lee, C. (1996) Speaker Verification Using Normalized Log-Likelihood Score, IEEE Tr. on Speech and Audio Processing Voi. 4. Issue 1, p. 56

[8] S. Nakagawa, K. P. Markov (1997). Speaker Verification Using Frame and Utterance Level Likelihood Normalization, Proc. of SPCHL97 ,Vol. 2, p. 1087

[9] K.T. Assaleh, R.J. Mammone (1994). New LP - Derived Features for Speak Identification - IEEE Tr.on SAP, vol.2, no.4, p. 630-638.

[10] H. Gish, M. Schmidt (1994). Text-Independent Speker Identification - IEEE Signal Proc. Mag., vol.11, nr.4, p. 18-32.

[11] Q. Lin s.a. (1994). Microphon Array Speaker Identification - IEEE tr. on ASSP, vol.2. nr.4, p. 622-629.

[12] D. Reynolds (1994). Experimental Evaluation of Feâtures for Robust Speaker Identification - IEEE Tr. on ASP, vol.2, nr.4, p. 639-643.

[13] F. Bimbot, G. Chollet, A. Paoloni (1994). Assessement Methodology for Speaker Identification and Verification - ESCA Workshop on Speaker Recognition, Identification and Verification, p. 75-83.

[14] M. Abe, S. Sagayama (1990). Statistical Study on Voice Individuality Conservation Across Different Languages - Proc. of ICSLP, p. 157-160.

[15] Y. Gong, J.P. Haton (1994). Non-Linear Interpolation Methods for Speaker Recognition - ESCA Workshop on Speaker Recognition, Identification and Verification, p .23-26.

[16] J. He s.a. (1995). On the Use of Feâtures from Prediction Rersidual Signal in Speaker Identification Proc. of EUROSPEECH95, p. 313-316.

[17] D.Naik s.a. (1994). Robust Speaker Identification Using Pole Filtering - ESCA Workshop on Speaker Recognition, Identification and Verification, p. 225-228.

[18] J. Openshaw, J. Masson (1994). Optimal Noise-Masking of Cepstral Feâtures for Robust Speaker Identification - ESCA Workshop on Speaker Recognition, Identification and Verification, p. 231-234.

[19] J. Thompson, J.S. Masson (1993). Within Class Optimization of Cepstra for Speaker Recognition, Proc. of EUROSPEECH, p. 165-168.

[20] K. Sonmez, L. Heck, M. Weintraub (2000). Multiple Speaker Tracking and Detection: Handset Normalization and Duration Scoring, Digital Signal Processing, 10(1/2/3), p. 133-143.

[21] T. Isobe, J. Takahashi (1999). A New Cohort Normalization Using Local Acoustic Information for Speaker Verification, Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing, 26.8, voi. 2, p. 841-844.

[22] X. Zhu s.a (1994). Text-Independent Speaker Recognition Using VQ, Mixture Gaussian VQ and Ergodic HMMs - ESCA Workshop on Speaker Recognition, Identification and Verification, p. 55-58.

[23] L. Boves s.a. (1994). Design and Recording of Large Data-Bases for Use in Speaker Recognition and Identification - ESCA Workshop on Speaker Recognition, Identification and Verification, p. 43-46.

[24] A. Federico (1995). Parametric Speaker Recognition Over Large Population of Telephonic Voices - Proc. of EUROSPEECH95, p. 329-332.

[25] J.L. Gauvain s.a (1995). Experiments with Speker Verification over the Telephone - Proc. of EUROSPEECH95, p. 651-654.

[26] C. Burileanu, D. Burileanu s.a.(2000). Cohort Normalisation Methods for Speaker Verification - Proc. of International Conference "Communications 2000", Bucharest, România, p.118-121.

[27] M. Wagner s.a. (1994). Analysis of Type-II Errors for VQ-Distortion Based Speaker Verification - ESCA Workshop on Speaker Recognition, Identification and Verification, p. 83-86.

[28] J.F. Bonastre (1993). Automaic Spaker Recognition and Analytic Process Proc. of EUROSPEECH93, p. 441-444.

[29] M. Sugiyama s.a. (1993). Speech Segmentation, Clustering Based on Speaker Feâtures - Proc. of ICASSP, p.395-398.

[30] H. Beigi, S. Maes and J. Sorensen (1998.) A Distance Measure Between Collections of Distributions and Its Application to Speaker Recognition, Proc. of ICASSP, Voi. 2, p. 753-756.

[31] L.E. Bojan, C. Burileanu s.a. (1996). Enhancements in Automatic Speaker Verification and Identification for Large Data-bases Using Pitch Contour Analysis - Proc. of ICSPAT96, Boston, SUA, p. 1796-1800

[32] C. Burileanu, L.E. Bojan s.a. (1993). A Representation for Recognition of Isolated Words Spoken in the Romanian Language - Proc. of ICSPAT93 Santa Clara, USA, p. 1478-1484.

Prelucrarea inițială a textului de intrare în cadrul unui sistem de sinteză a vorbirii pornind de la text în limba română

Dragoş BURILEANU
Laboratorul de "Tehnologia vorbirii şi prelucrarea digitală a semnalelor",
Facultatea de Electronică şi Telecomunicații,
Universitatea "POLITEHNICA" Bucureşti
Bdul Iuliu Maniu 1-3, Sector 6, 77202 Bucureşti
bdragos@mESsnet.pub.ro

1. Introducere

Limbajul reprezintă modalitatea de exprimare a ideilor prin intermediul unui ansamblu de semne, fie grafic, fie prin gesturi, sau sunete, un astfel de sistem structurat fiind specific doar oamenilor. Fără îndoiala, *vorbirea* este una din principalele sale componente; ea este cea mai veche modalitate de comunicare între oameni şi este şi astăzi cea mai răspândită. Este deci uşor de înțeles faptul că vorbirea a fost studiată intens şi s-a încercat adesea să fie prelucrată într-un mod automat. Pentru mulți ingineri şi specialişti din domeniu, posibilitatea de a conversa liber cu o maşină reprezintă de fapt o adevărată provocare pentru înțelegerea cât mai deplină a proceselor de producere şi percepție implicate în comunicarea prin voce între oameni. Ceea ce este însa şi mai important este faptul ca *interfețele de comunicare prin voce* devin tot mai mult o necesitate. în viitorul apropiat, sistemele şi rețelele interactive vor oferi un acces simplu si ieftin la cantități mari de informație si servicii, ceea ce va afecta fundamental viața noastră zilnică.

Deşi principiile de bază ale producerii şi recepționării vorbirii au început să fie studiate încă de la sfârşitul secolului al XVIII-lea, când s-au înregistrat primele cercetări în domeniul dezvoltării sintetizoarelor mecanice de sunete asemănătoare vocii umane, *tehnologiile de prelucrare a vorbirii* au obținut rezultate semnificative doar în ultimele decenii (fiind denumite în sens larg *tehnici de analiză şi sinteză a semnalului vocal).* Aceste rezultate au fost posibile datorită progreselor făcute în domeniile acusticii şi lingvisticii, modelării matematice a producerii şi percepției vorbirii, prelucrării semnalelor şi tehnologiilor VLSI. Putem evidenția în acest sens dezvoltarea procesoarelor numerice de semnal pe un singur chip, realizarea de

capsule de memorie mai mari şi mai ieftine, apariţia unor algoritmi îmbunătăţiţi pentru prelucrare de semnal, iar în domeniul comunicaţiilor crearea de standarde globale pentru transmisie, compresie de semnal şi protocoale de comunicaţie.

Prin urmare, putem aprecia că cercetările actuale în domeniul prelucrării vorbirii au ca scop larg îmbunătăţirea calităţii, securităţii şi costului comunicaţiilor şi a accesului uman la informaţii. Pe de o parte, este de aşteptat în viitorul apropiat o extindere importantă a serviciilor integrate de voce, poştă electronică, FAX, paging şi transmisiuni de date pe canale fără fir. Pe de altă parte însă, comunicarea verbală între om şi maşini, în ambele sensuri, tinde deja să devină o realitate, fiind vizibilă tendinţa actuală de a apropia caracteristicile maşinii de cele ale utilizatorului uman.

în acest ultim sens, trebuie observat faptul că tendinţa menţionată anterior este absolut firească. Filozoful grec Aristotel (384 - 322 î.C, fondator al logicii formale), afirma: *"Raţiunea de a fi a oricărui lucru constă în funcţia sa".* Ori este evident faptul că o interfaţă de dialog prin voce reprezintă o modalitate ideală de comunicare cu maşina, vorbirea fiind cea mai naturală, flexibilă, eficientă şi economică modalitate de comunicare utilizată de oameni.

Aceste idei legate de posibilitatea comunicării prin voce între om şi maşina nu sunt noi; totuşi, doar în ultimii ani a început să prindă contur conceptul ce a căpătat denumirea de "dialog om-maşină", iar tehnologia necesară implementării acestui concept a părăsit deja laboratoarele şi a pătruns în lumea reală, într-o gamă largă de aplicaţii.

Pentru a realiza un mod de comunicare cât mai natural şi pentru a permite o utilizare cât mai largă, calculatorul trebuie să înţeleagă şi să producă singur vorbirea; acesta este motivul principal pentru care *recunoaşterea* şi *sinteza vorbirii* au devenit în ultimii ani tehnologii de un interes special şi constituie subiecte pentru cercetări intense şi aprofundate. Ambele tehnologii prelucrează vorbirea în primul rând sub aspectul conţinutului informaţional: recunoaşterea transformă vocea omului în text ce poate fi folosit literal (de exemplu pentru dictare), sau o interpretează sub forma unor comenzi de control pentru diverse aplicaţii, iar sinteza permite generarea limbajului vorbit pornind de la text sau de la anumite concepte.

Cu toate că s-au făcut paşi importanţi în aceste domenii, rezultatele sunt încă departe de aşteptări. Sarcinile enunţate iniţial s-au dovedit în timp a fi deosebit de dificile, în primul rând datorită complexităţii semnalului vocal ca şi a dificultăţilor legate de prelucrarea acestuia, dificultăţi legate fie de recunoaşterea conţinutului său informaţional (semnalul vocal depinzând puternic de vorbitor şi de condiţiile în care acesta rosteşte un mesaj), fie de producerea sa, fie de transmiterea acestui semnal la distanţă [1].

în acest context, producerea vorbirii artificiale şi în special co
voce, care constituie obiectul principal al lucrării de faţă, este astăzi
bază al domeniului prelucrării vorbirii şi subiect al unor cercetări inten
de sinteză pornind de la text (TTS - *"Text-to-Speech")* poate oferi o
de aplicaţii, de la accesul la poşta electronică şi diferite tipuri de baz
pronunţarea unui text pentru persoane cu handicap vizual.

Este important de observat faptul că tehnologia de răspu
prezintă o serie de avantaje fundamentale pentru transmiterea inform

- oricine poate înţelege un mesaj, fără antrenare sau deosebită;
- mesajul poate fi recepţionat chiar dacă cel ce ascultă e alte activităţi, cum ar fi mersul, manipularea unor obiec altor informaţii;
- reţeaua telefonică convenţională poate fi utilizată pentru la distanţă la o bază de informaţii;
- această formă de comunicare este mai economic tradiţională prin mesaje scrise.

Toţi aceşti factori precum şi numeroasele aplicaţii cerute d
creat premisele unor cercetări aprofundate, obţinându-se astfel
sisteme comerciale care pot produce vorbire sintetică pornind de
inteligibilitate acceptabilă.

într-adevăr, scopul principal al celor mai multe sistem
existente este de a produce o vorbire *inteligibilă.* Din acest pun
sinteza pare a fi de mai multă vreme o tehnologie "stabilă", ieft
implementat; se spune chiar, uneori, că acest domeniu este în prez
bine dezvoltat, iar problemele rămase sunt minore din punct de v
Dacă însă scopul este sinteza în timp real, pornind de la un vocabu
cuvinte şi fără restricţii asupra textului, iar vorbirea să fie nu numai i
la fel de *naturală* ca cea umană, atunci se constată că performanţe
departe de a fi satisfăcătoare. Rămân încă multe probleme importan
extinderea vocabularului oferit, înlăturarea restricţiilor impuse text
unor caractere speciale, îmbunătăţirea caracteristicilor de *prozodie,*
modificare a *ritmului* şi *stilului* vorbirii sintetizate, sau elaborarea u
sinteză în mai multe limbi. Aceste sarcini se dovedesc a fi deosel
cer, evident, eforturi interdisciplinare susţinute [2].

2. Sinteza automată a vorbirii

Etimologic, cuvântul "sinteză" provine din limba greacă şi semnifică îmbinarea mai multor elemente diferite într-un tot.

în ceea ce priveşte *sinteza vorbirii,* nu există o definiție precisă şi unanim acceptată de către specialiştii în tehnologia vorbirii. Acest termen a avut în decursul timpului mai multe accepțiuni, majoritatea depinzând de nivelul tehnologic al momentului şi de elementele constitutive ale semnalului vocal care au fost folosite pentru sinteză. De exemplu, primele circuite integrate care permiteau simpla restituire a unui mesaj vocal înregistrat şi stocat digital au purtat denumirea de "sintetizoare vocale", fie că se făcea sau nu o compresie a semnalului. Este evident că în acest caz nu se poate vorbi de sinteză, din moment ce **textul este fix** şi astfel de sisteme nu pot rosti decât mesaje preînregistrate; chiar dacă vocea umană este comprimată cu ajutorul unui algoritm, nu este cu adevărat "sintetică", ci poate fi numită mai curând o "înregistrare cu număr redus de biți".

Aceeaşi situație este în cazul sintezei la recepție a unor mesaje transmise pe canale de comunicație standard (caracteristică sistemelor de tip "vocoder"), care este de obicei considerată ca făcând parte din domeniul *codării vorbirii* şi cuprinde tehnici de reducere a debitului semnalului vocal pentru transmisie; cu alte cuvinte, şi acest tip de sinteză, care reface **acelaşi** mesaj analizat la emisie, deci nu generează fraze **noi,** nu este tratat ca o sinteză automată propriu-zisă.

O categorie distinctă de sinteză vocală este aceea care implică sisteme ce concatenează cuvinte sau fraze preînregistrate, dar generează fraze noi, acestea nefiind niciodată pronunțate ca atare; astfel de sisteme cer utilizarea unor reguli lingvistice mai mult sau mai puțin complicate pentru a funcționa corespunzător.

în sfârşit, o categorie specială o reprezintă sinteza vorbirii pornind de la text; aceasta reprezintă, în esență, transformarea unui text oarecare, scris într-un anumit limbaj, în semnal vocal. Trebuie remarcat faptul că în prezent, în multe lucrări ştiințifice, acest tip de sinteză este sinonim chiar conceptului de sinteză automată a vorbirii.

Analizând exemplele de mai sus, putem defini trei noțiuni generale [3], pe care le vom utiliza pe parcursul lucrării de față:

Definiția 2.1 ***Sinteza automată a vorbirii*** este "tehnologia integrată care simulează procesul uman de generare a vorbirii, mergând de la sisteme simple ce pot genera automat fraze noi şi cuprind un formalism lingvistic minimal şi până la sisteme care transformă în vorbire reprezentări simbolice sau lingvistice ale limbajului".

Definiția 2.2 Un ***sistem de sinteză pornind de la text*** este "un sistem automat care poate produce vorbirea plecând de la un text scris, prin intermediul unei reprezentări fonetice a mesajului".

Definiția 2.3 ***Sintetizorul vocal*** este "etajul unui sistem de sinteză automată vorbirii care realizează conversia finală în semnal vocal, pornin de obicei de la o reprezentare parametrică a unor segment acustice fundamentale".

3. Sinteza vorbirii pornind de la text

Pentru a înțelege mai bine dificultatea sarcinii unui sistem de sintez pornind de la text, considerăm că este util să punem în evidență mai întâi mod (fiziologic) în care o persoană citeşte cu voce tare un text. Imaginea textului es sesizată de neuronii sistemului vizual, transmisă creierului sub forma unor stimu electrici, aici fiind prelucrată pentru a putea permite comanda neuronil responsabili de corecta activare a plămânilor^ coardelor vocale şi organel articulatorii. în acest fel se produce vorbirea, ea fiind permanent monitorizată c creier (în special prin intermediul organelor auditive), în scopul ajustă configurației tractului vocal în timp real.

Desigur, cunoaştem încă prea puțin despre organizarea de ansamblu sistemului nervos uman, care este capabil de această activitate complexă; pute propune totuşi următorul *model funcțional* prin care este prelucrată informat optică şi apoi este dată comanda de generare a vorbirii:

- Atunci când citim un text, efectuăm practic o sarcină de *recunoaşte de caractere,* ignorând, parțial inconştient, anumite erori de redacta a cuvintelor (caractere lipsă sau înlocuite cu altele) şi decodificând m degrabă cuvântul ca un întreg; are loc un proces de inferență informației dintr-un context posibil incomplet. De asemene recunoaştem cu uşurință caractere speciale sau abrevieri.
- Considerând *fonemele* ca fiind cele mai mici elemente sonore ca permit diferențierea între ele a cuvintelor, este evident că secver fonemică corespunzătoare unui cuvânt diferă de şirul de caracte grafice din care este compus cuvântul; creierul trebuie sa facă p urmare o *transcriere fonetică pornind de la litere,* această opera practic instinctivă permițând pronunția unui număr nelimitat de cuvint
- în cele mai multe situații, suntem capabili să începem pronunția u fraze mult înainte de terminarea ei; cu alte cuvinte, putem face *stucturare sintactică,* descompunând fiecare propoziție în grupuri cuvinte şi asociindu-le intonația corespunzătoare. Şi acest proces es practic inconştient, fiind bazat pe educație şi experiență.
- în sfârşit, putem discrimina cu uşurință cuvinte ce se scriu asemănă dar se pronunță diferit, după *înțelesul semantic,* fapt posibil dato aceleiaşi capacități de deducție a creierului de care am vorbit mai su

Concluzia este simplă: pe baza experienței lingvistice căpătate în urma educației, o persoană familiară cu limbajul în care este scris un text depăşeşte imediat paşii descrişi anterior şi poate cu uşurință să citească cu voce tare textul scris, în primul rând pentru că înțelege ceea ce citeşte.

Având în vedere considerațiile expuse anterior, devine evident faptul că o maşină care trebuie să pronunțe un text scris nu va putea adopta o schemă de prelucrare atât de complicată cum este cea care caracterizează acțiunea citirii cu voce tare a unui text de către o persoană. Sunetele vorbirii sunt inerent guvernate de ecuații diferențiale ale mecanicii fluidelor, aplicate într-un context nestaționar, deoarece presiunea aerului ia nivelul plămânilor, tensiunea glotală, ca şi configurațiile tractului vocal şi nazal, evoluează în timp. Toate acestea surit controlate de creierul uman, care beneficiază de avantajul puterii sale de prelucrare paralelă pentru extragerea esenței textului citit: înțelesul. Chiar şi la nivelul la care a ajuns ştiința astăzi (cercetări intense în domeniile sintezei articulatorii, rețelelor neuronale artificiale şi prelucrării limbajului natural), construirea unui sistem de sinteză pornind de la text cu un model atât de complex rămâne practic nerealizabilă; chiar dacă, să spunem, s-ar ajunge foarte aproape de aceste cerințe, sistemul rezultat nu ar fi de loc compatibil cu criteriile economice normale.

Figura 1 introduce o diagramă funcțională foarte generală a unui sistem TTS, bazată pe observațiile anterioare.

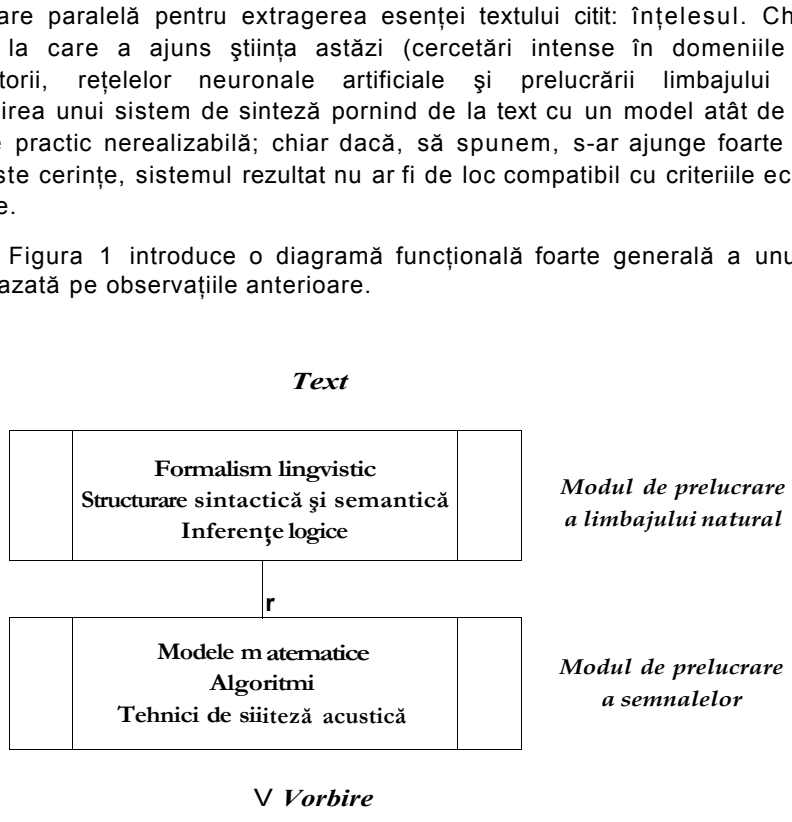


Figura 1. Diagramă funcțională pentru un sistem TTS

Ca şi pentru un cititor uman, schema cuprinde un *modul de prelucrare limbajului natural,* capabil să producă o transcriere fonetică a textului citi împreună cu informații despre intonație, accente, durate şi de asemenea un *modu de prelucrare a semnalelor,* care transformă informația simbolică primită în vorbir sintetică, pe baza unor tehnici de sinteză adecvate şi a unor structuri stocate î urma unei analize preliminare. Etapele de bază ale sintezei pornind de la text pot astfel descrise printr-un număr de transformări succesive ce trebuie aplicat asupra şirului de caractere ce reprezintă textul de intrare; scopul este de a s obține o vorbire **de calitate,** într-o limbă oarecare, fără constrângeri asupra textulu introdus.

Trebuie menționat faptul că formalismul descris poate "sări" uneori pest anumiți paşi, dacă se utilizează în mod adecvat cunoaşterea lingvistică matematică; acest lucru se întâmplă atunci când punem anumite restricții asupr textului ce trebuie pronunțat, sau impunem vorbirii sintetizate o inteligibilitate şi naturalețe moderate. Cu alte cuvinte, proiectarea sistemului TTS se poat simplifica dacă se impun sistemului sarcini precise, corespunzătoare unor aplicat concrete.

Colectivul nostru de cercetare a început acum câțiva ani dezvoltarea un sistem complet TTS în limba română, bazat pe *concatenare de difonem* Arhitectura acestui sistem este prezentată în Figura 2. Sistemul cuprinde o part importantă de prelucrare lingvistică şi un modul de generare a semnalului d vorbire având la bază un algoritm de tip PSOLA [4]. După realizarea unei prim variante a sistemului, se depun în continuare eforturi pentru creşterea naturale vorbirii sintetizate, prin îmbunătățirea performanțelor la diferite nivele d prelucrare.

Modulul de prelucrare a limbajului într-un sistem TTS are ca sarcin transformarea textului de intrare într-o reprezentare fonetică şi prozodică, ca trebuie să descrie cât mai fidel posibil pronunția sa. Acest lucru poate fi realiz parcurgând mai multe etape succesive, puse în evidență şi în figura anterioar Vom discuta în cele ce urmează **modalitățile de rezolvare a părții de prelucra inițială (preprocesare) a textului în cadrul sistemului nostru de sinteză limba română.**

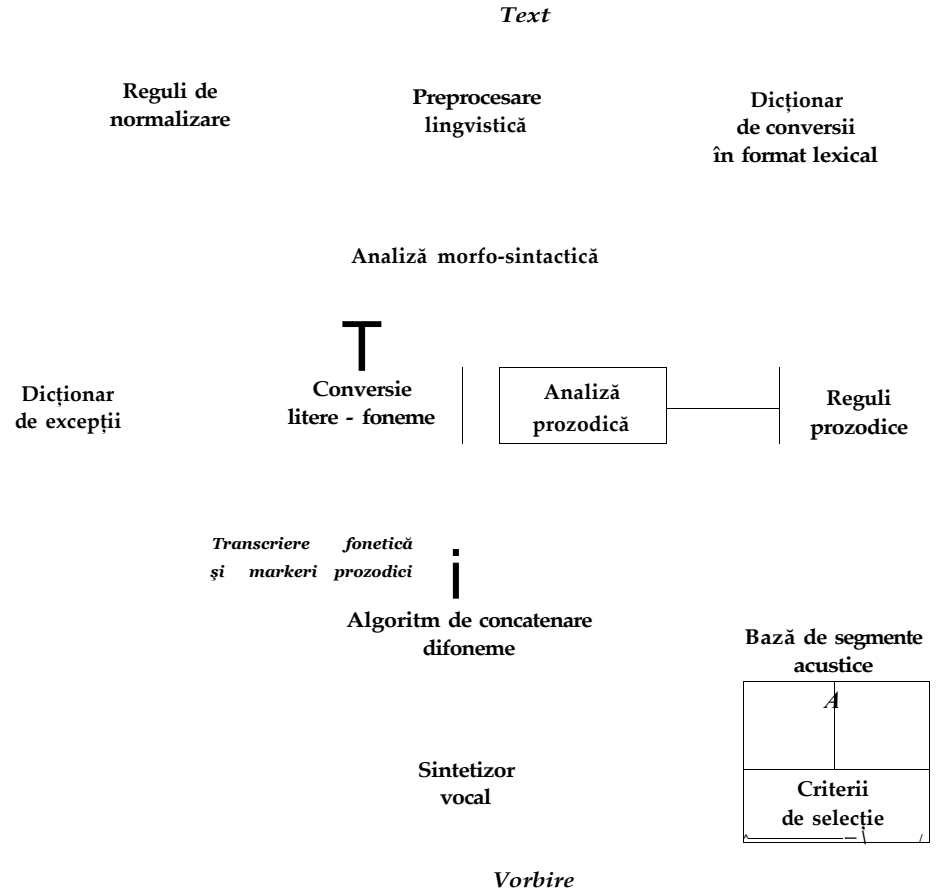


Figura **2.** Arhitectura sistemului TTS în limba română

4. Preprocesarea textului de intrare în cadrul sistemului TTS în limba română

4.1 Probleme generale

«ktpm* dificultățile mai reale sistemelor TTS constă în faptul că aceste sisteme trebuie sa poată prelucra practic orice text, plecând de la propoziții simple izolate şi mergând până la paragrafe complexe, care pot cuprin de propoziții, cu posibile structuri negramaticale şi simboluri sp partea de preprocesare lingvistică a textului are un rol extr deoarece detectarea corectă şi interpretarea şirurilor de ca influențează acuratețea întregului sistem de sinteză şi contribuie text fără restricții în vorbire sintetică.

Uzual, un text scris se prezintă sub forma unei secv ASCII; el este alcătuit din cuvinte compuse cu ajutorul literelor din alte tipuri de caractere: spații albe, semne de punctuație, şiru alte simboluri speciale (de exemplu operatori matematici). Tex numerale *(12, 12.450, 1,245),* abrevieri *(prof., dr, ing.),* sau acr *TTS).* Aceste secvențe sunt de obicei "anormale" din punct de față de majoritatea cuvintelor din text şi trebuie mai întâi tra format ce poate fi recunoscut de partea de analiză lingvistică revine modulului de preprocesare, care trebuie de asemene segmentare a textului de intrare (detectarea cuvintelor şi a sfârş prelucrare a semnelor de punctuație şi a simbolurilor speciale [5,

La prima vedere, preprocesarea unui text pentru un s părea banală; în realitate însă, lucrurile sunt destul de complica nu este totdeauna posibilă determinarea marginilor unei fraze p de punctuație. Astfel, punctul (.) poate apare şi la sfârşitul un multe alte situații, ca de exemplu în abrevieri *(ing.),* acronime (S se omite un anumit fragment de text (...), sau numerale *(12.4 mii patru sute cincizeci),* situații care trebuie diferențiate prin p [9]. De asemenea, cratima creează dificultăți în operația de seg fi folosită pentru despărțirea în silabe, pentru scrierea cuvintelo delimitarea unui nou paragraf, sau în enumerări.

O sarcină dificilă este şi conversia anumitor secvenț cuvinte care să poată fi analizate lingvistic. Dacă unele abre "expandate" imediat, cu ajutorul unui tabel de echivalențe, exis care secvențe de simboluri care nu se pot distinge pe baza c tipuri diferite de conversii; de exemplu, numărul format din ş poate reprezenta un număr întreg sau un număr de telefon şi va cele două situații. în general, prezența şirurilor de numere în tex dificultăți, deoarece ele pot apare în diferite contexte: ore, date, expresii aritmetice etc.

Trebuie observat că aceste ambiguități create de natura semnelor de punctuație sau de modul diferit de citire a acel simboluri, pot avea implicații majore asupra acurateții într prelucrare lingvistică şi în final asupra pronunției corecte a sistemul de sinteză.

Evident, numărul secvențelor de caractere neuzuale dintr-un text ce se doreşte a fi transformat în vorbire depinde mult de tipul şi subiectul textului. Spre exemplu, textele literare dintr-un volum de proză sau comentariile politice dintr-un ziar au mult mai puține situații dificile decât comentariile economice, sportive, sau prezentările de spectacole. în ultimele situații menționate, construcțiile neuzuale, criptice sau chiar negramaticale, abrevierile uneori ambigui, pot fi atât de numeroase, încât se poate spune chiar că astfel de texte nici nu sunt potrivite pentru o sinteză automată pornind de la text; singura soluție rezonabilă este, probabil, o reeditare a lor pentru a le face mai accesibile unui sistem de sinteză.

Problema enunțată anterior este de fapt mult mai generală. Părerea autorului acestei lucrări este că în orice aplicație TTS trebuie făcut un compromis între calitatea vorbirii sintetizate, dimensiunile vocabularului şi complexitatea sistemului de sinteză. Cu alte cuvinte, nu trebuie încercat cu orice preț, prin orice mijloace, obținerea unei vorbiri "perfecte", cel puțin în acest moment.

4.2 Algoritm de preprocesare a textului

Pentru preprocesarea textului de intrare în cadrul sistemului TTS proiectat, am propus un set de definiții, reguli şi proceduri, bazate pe o analiză detaliată a situațiilor cele mai întâlnite în limba română.

Definițiile propuse sunt prezentate în continuare.

Definiția 4.1 Vom denumi *expresii* "secvențele de caractere care cuprind una sau mai multe din următoarele categorii: secvențe de litere dintre care cel puțin una este majusculă, secvențe de cifre, semne de punctuație, alte simboluri speciale".

Definiția 4.2 Vom denumi *caractere extra-textuale* "acele semne de punctuație care îndeplinesc în text o funcție de punctuație propriu-zisă".

Definiția 4.3 Vom denumi *caractere intra-textuale* "acele semne de punctuație care fac parte integrantă din expresii şi ajută la pronunția lor".

Definiția 4.4 Vom denuni *expandare* "procesul de conversie a unor expresii în format lexical (secvențe de litere alcătuind cuvinte uzuale, ce pot fi analizate lingvistic)".

Definiția 4.5 Vom denumi o secvență de caractere *ambiguă* "dacă ea poate fi încadrată, având în vedere forma sa, în mai multe clase lingvistice".

Pornind de la aceste definiții, am proiectat un algoritm de preprocesare a textului, ce constă în principiu din trei etape de bază:

I. Segmentarea textului

Textul se segmentează de la stânga spre dreapta, în *grupuri de caractere.* Se obțin astfel secvențe de caractere ASCII delimitate de spații albe (blanc); semnele de punctuație se includ temporar în aceste grupe.

II. Conversia şirurilor de caractere de tip expresie în caractere ortografice

Se parcurg pe rând grupurile de caractere rezultate în urma segmentării şi se realizează *expandarea* lor (acolo unde este cazul) sub forma unor cuvinte uzuale, pe baza unei analize contextuale simple la nivel de cuvânt sau segment de cuvânt şi a unor dicționare de conversie în format lexical (pentru abrevieri şi unele tipuri de acronime).

III. Interpretarea unor semne de punctuație

Se detectează şi se memorează pozițiile unor *caractere extra-textuale* şi a sfârşitului frazelor, pentru a fi folosite ulterior de modulele de analiză sintactică şi prozodică.

Detaliind etapa I prezentată anterior şi utilizând şi definiția 4.1, putem observa că grupurile de caractere rezultate în urma segmentării textului de intrare pot fi de următoarele tipuri [10, 11, 12]:

a. Secvențe de litere alfabetice, scrise cu minuscule

a1. Cuvinte uzuale;

a2. Abrevieri scrise fără punct (de exemplu unități de măsură: *m, km, ms).*

b. Expresii

b1. Cuvinte scrise cu o singură literă, majusculă: abrevieri (puncte cardinale: *E* - est, *V* - vest; simboluri chimice: C - carbon, O - oxigen; unități de măsură: *A* - amper, *V* -volt); cifre romane: *V* - cinci, / - unu etc.

b2. Abrevieri scrise cu minuscule şi puncte *{tel.* - telefon, a.c. - anu curent)

b3. Secvențe de mai multe litere, scrise cu minuscule şi inițială majusculă

b3.1. Cuvinte la început de frază;

b3.2. Nume proprii;

b3.3. Abrevieri scrise fără punct (de exemplu unități de măsură Hz, *Mw).*

b4. Secvențe de mai multe litere, scrise cu minuscule şi o majusculă pe altă poziție decât prima (unități de măsură: *mA, kV* etc.)

b5. Secvențe de litere scrise cu mai mult de două majuscule, cu sau fără punct

b5.1. Acronime *(NATO, S.R.L.);*

b5.2. Abrevieri (P.S. - post scriptum);

b5.3. Unități de măsură *(MHz, MByte);*

b5.4. Cifre romane *(VI, IX).*

b6. Secvențe de cifre, scrise cu sau fără semne de punctuație

b6.1. Numere întregi;

b6.2. Numere zecimale;

b6.3. Numerale ordinale *(al 2-lea)]*

b6.4. Ore şi date;

b6.5. Numere de telefon.

b7. Semne de punctuație: . ? ! : ; . . . , - / ' " ( ) [ ] { }

bS. Simboluri speciale

b8.1. Simboluri matematice uzuale: + - *(saux) :(sau /) = < > % ~

b8.2. Alte simboluri speciale: @ $ &

Deoarece semnele de punctuație ridică cele mai serioase probleme, vom analiza în primul rând situațiile cele mai uzuale de apariție a lor (pe grupe de importanță), precum şi soluțiile posibile de rezolvare a acestor situații. Vom discuta apoi câteva aspecte fundamentale legate de grupurile de cifre, abrevieri şi acronime.

1. Punctul

Punctul (.) poate apare în abrevieri, acronime, numerale, sau poate semnifica sfârşitul unei fraze. Ambiguitățile create de punct sunt o problemă majoră pentru operația de preprocesare, datorită faptului că el poate reprezenta fie un caracter intra-textual, fie extra-textual, fie ambele în acelaşi timp; de exemplu, punctul după abreviere poate marca în acelaşi timp şi sfârşitul frazei.

Este deosebit de utilă punerea în evidență a câtorva situații de utilizare corectă a punctului în limba română:

- Punctul se foloseşte în abrevierile provenite din cuvinte simple sau compuse în care nu apare litera finală a cuvântului; exemple: *id.* (idem), *etc.* (etcetera), *tel.* (telefon), *a.c.* (anul curent), *a.m.* (ante meridian), *d.a.* (după-amiaza), *P.S.* (post scriptum) - deci categorile **b2, b5.2** puse în evidență anterior.
- Dacă în abreviere apare litera finală a cuvântului, nu se pune punct după abreviere; exemple: *cea* (circa), *dna* (doamna), *dl* (domnul), *dnei* (doamnei), *jr* (junior) - categoria **a2.**
- Nu se pune punct după simbolurile unor termeni de specialitate: C (carbon), *L* (lungime), *V* (vest sau volt), *mA* (miliamperi), *MHz* (mega hertzi) - categoriile **a2, b1, b3.3, b4, b5.3.**
- în acronime (abrevieri provenite din inițialele unor substantive compuse formate din mai mulți termeni), punctul este facultativ; sunt corecte atât formele *O.N.U., S.U.A.,* cât şi *ONU, SUA* (categoria **b5.1).**
- Nu se foloseşte punctul în abrevierile ce reprezintă indicative de state *(RO* - România), sau de județe *(CT* - Constanța) şi în situațiile când abrevierea s-a transformat într-un cuvânt sudat, caracterizat prin lectură cursivă *(TAROM)* - categoria **b5.2.**
- Punctul se foloseşte de asemenea în scrierea unor numere şi a datelor: numere întregi sau zecimale *(1.234, 1.234,567),* date *(15.04.2002)* - categoriile **b6.1, b6.2.**

Considerațiile anterioare sugerează următoarea procedură: atunci când este detectat punctul într-un grup de caractere, se cercetează contextul în care apare şi apoi se ia decizia corespunzătoare, astfel:

- Dacă există cifre la stânga şi la dreapta, el este declarat caracter intra-textual şi:

  dacă mai există un punct în secvența de cifre, secvența reprezintă o dată şi se expandează folosind un set de reguli (de exemplu: *15.04.2002* va deveni *cincisprezece aprilie două mii doi))*

  dacă nu mai există un alt punct, secvența reprezintă un număr şi se expandează folosind de asemenea reguli (de exemplu: 1234 va deveni o *mie două sute treizeci şi patru).*

- Dacă punctul este în poziție finală şi este precedat de alte două puncte (...), această secvență se declară caracter extra-textual, fiind identificată cu semnul de punctuație corespunzător; acest caz îl vom discuta separat.
- Dacă punctul este precedat de o secvență de litere (minuscule sau majuscule) şi eventual de alte puncte, se caută într-un dicționar de abrevieri şi acronime şi:

- dacă grupul de caractere este găsit în dicționar, punctul este declarat caracter intra-textual şi secvența se expandează conform echivalenței din dicționar;

' - dacă grupul de caractere nu este găsit în dicționar, dar conține majuscule, este un acronim - această situație o vom discuta separat;

dacă grupul de caractere nu este găsit în dicționar şi nu conține majuscule şi alte puncte, punctul (care este sigur în poziție finală) este declarat caracter extra-textual şi va reprezenta sfârşitul unei fraze, poziția sa fiind memorată pentru modulele de analiză sintactică şi prozodică.

Ultimele reguli prezentate nu pot însă elimina ambiguitatea situației în care punctul după o abreviere poate reprezenta în acelaşi timp şi sfârşitul frazei (cazul lui *etc.* este tipic, dar există şi numeroase alte exemple).

O soluție ar putea fi cercetarea grupului de caractere ce urmează după blanc, ținând cont de faptul că la începutul unei noi fraze se află de regulă un cuvânt cu inițială majusculă. Această situație nu este însă complet edificatoare, deoarece în limba română majuscula apare ca inițială în multe cazuri: substantive nume proprii de persoană, nume de localități sau denumiri geografice, nume de planete şi constelații, nume de instituții, nume de lucrări, nume de evenimente istorice sau de manifestări artistice şi ştiințifice, nume de sărbători, ca semn de respect etc.

Este clar că această ambiguitate nu va putea fi rezolvată numai de către preprocesor. Soluția pe care o propunem este următoarea:

- Dacă în urma cercetării contextului din dreapta rezultă că punctul din finalul unei abrevieri ar putea fi în acelaşi timp şi sfârşitul frazei, punctul rămâne caracter intra-textual (şi ajută la expandarea abrevierii), dar se adaugă un simbol special pentru marcarea provizorie a sfârşitului frazei, urmând ca acesta să fie validat sau nu de analiza sintactică ulterioară.

2. Semnele de punctuație ? ! : ; ...

Situațiile cele mai frecvente de apariție a lor sunt următoarele:

- Semnul întrebării (?) şi semnul exclamării (/) se folosesc uzual în limba română la sfârşitul frazei. Ele apar foarte rar în interiorul frazelor, când pot reprezenta, de exemplu, considerații personale introduse în text, acestea fiind de obicei puse între paranteze; ca atare, cercetarea caracterului din dreapta lor (blanc sau paranteză) poate diferenția simplu cele două situații.

|- li % £ ;|

*. • t

,, L

~

~

- Semnele : şi ; marchează şi ele, de cele mai multe ori, finalul unui enunț. Deşi nu constituie un sfârşit de frază propriu-zis, pot fi considerate în acest fel în contextul sintezei TTS, deoarece textele din partea stângă şi din partea dreaptă se pot pronunța ca şi cum ar fi izolate, fără să fie afectată naturaleța pronunției.
  Prin urmare, cele patru semne menționate sunt importante în primul rând pentru modulul de analiză prozodică, deci locul lor trebuie detectat şi memorat de către preprocesor, iar poziția în frază (finală sau intermediară) este utilă doar pentru a uşura analiza sintactică ulterioară a textului.
- Semnul ... semnifică faptul că se omite un anumit fragment de text (de exemplu finalul neprecizat al unei enumerări); el apare în mod obişnuit la sfârşitul unei fraze, dar poate apare şi în poziție intermediară. Putem deci aplica aceeaşi regulă ca şi pentru punctul final al unei abrevieri: cercetarea contextului din dreapta şi, dacă este cazul, marcarea provizorie ca final de frază, până la o analiză sintactică mai aprofundată; altfel, el nu modifică prozodia textului,

în toate situațiile menționate, semnele de punctuație vor fi interpretate drept caractere extra-textuale. Există însă şi trei excepții, în care semnele / şi : au altă semnificație decât cea uzuală; aceste situații pot fi descrise de următoarele reguli:

- Dacă simbolul / se găseşte la finalul unei secvențe de numere, el semnifică cu mare probabilitate un "factorial" şi va fi transcris ca atare.
- Dacă simbolul : se găseşte în interiorul unei secvențe de numere, este considerat caracter intra-textual; secvența reprezintă o oră şi se expandează folosind un set de reguli (de exemplu: *14:30* va deveni ora *paisprezece şi treizeci de minute).*
- Dacă simbolul: este înconjurat de blancuri, face parte dintr-o expresie matematică şi va fi transcris conform dicționarului *(împărțit la).*

3. Virgula

Virgula (,) apare în mod uzual într-o frază în poziție intermediară, la finalul unui cuvânt, dar poate apare şi în scrierea numerelor zecimale. Regula aplicată în cadrul algoritmului propus este următoarea:

- Se cercetează contextul în care apare virgula şi:

  dacă este înconjurată de cifre, se consideră caracter intra-textual; secvența reprezintă un număr zecimal şi se expandează folosind un set de reguli (de exemplu: *1,234* va deveni *unu virgulă două sute treizeci şi patru).*

dacă la stânga sa se găseşte o literă sau un alt semn de punctuație (de exemplu punct după o abreviere), se consideră caracter extra-textual şi poziția sa va fi memorată pentru modulul de analiză prozodică.

4. Cratima

Cratima (-) este un semn ortografic ce are în limba română două valori principale:

- *gramaticală,* atunci când serveşte la scrierea unor cuvinte compuse *(bună-cuviință, nord-vest, prim-plan, pare-mi-se, propriu-zis etc.)*\
- *fonetică,* atunci când serveşte la marcarea pronunțării într-o singură silabă a două sunete din două cuvinte diferite, dar care se găsesc alăturate în vorbirea curentă *(de-a).*

în fapt, deoarece simbolurile uzuale folosite de calculator nu cuprind linii mediane de lungimi diferite, cratima devine practic un semn de punctuație şi poate fi folosită atât pentru scrierea cuvintelor compuse sau a unor numerale ordinale, cât şi pentru despărțirea în silabe, pentru delimitarea unui nou paragraf, sau în enumerări.

Determinarea caracterului intra sau extra-textual se poate face prin cercetarea contextului în care apare; ea este mărginită de obicei fie de litere, fie de blancuri, dar această informație este utilă doar pentru analiza sintactică, deoarece în mod uzual nu se citeşte (este suprimată de către preprocesor) şi nu modifică prozodia textului. în numeralele ordinale, expandarea se face simplu, pe bază de reguli *(al 2-lea - al doilea).*

5. Bara oblică

Bara oblică (/) are sensul prepoziției "pe" în abrevierile ştiințifice *(km/h - kilometru pe oră, m/s - metru pe secundă)* şi în exprimarea unei proporț/7 *(2/3 - doi pe trei),* sau sensul conjuncției "sau" în textele uzuale *(c(e/i) - ce* sau *ci)*\ în ambele situații reprezintă un caracter intra-textual. De asemenea, poate semnifica o împărțire în expresiile matematice.

Regulile pe care le propunem pentru simbolul / sunt următoarele:

- Dacă este înconjurat de litere, grupul de caractere din care face parte se caută în dicționarul de abrevieri şi:

  dacă se găseşte în dicționar, se transcrie *pe* şi se foloseşte expresia completă găsită *(metru pe secundă);*

  dacă nu este găsit în dicționar, se transcrie *sau.*

f
ț
1
i

- Dacă este înconjurat de numere izolate, se transcrie *pe.*
- ~ Dacă este înconjurat de secvențe de cifre şi alte caractere matematice (*2x3/4x5*), sau de paranteze şi secvențe de cifre (*(2+3)/(4+5)),* se transcrie *împărțit la.*

6. Apostroful

Apostroful (') este folosit în limba română în mai multe situații:

- pentru a reproduce în scris rostiri în care un sunet sau mai multe nu sunt pronunțate; aceste rostiri sunt însă rare, fiind practic neliterare, populare *(pân'deseară)',*
- în nume proprii străine sau în neologismele neadaptate *(O'Neill, five o'clock)',*
- în scrierea anilor, fără prima sau primele cifre *('907,* '99).

Regulile pe care le propunem pentru simbolul ' sunt următoarele:

- Dacă se găseşte într-o secvență de litere, el este eliminat (nu reprezintă propriu-zis un caracter intra-textual şi nu ajută la pronunția cuvântului).
- Dacă în dreapta se găseşte o secvență de cifre, în funcție de numărul acestor cifre, grupul de caractere se expandează folosind un set de reguli (de exemplu: *'99* va deveni o *mie nouă sute nouăzeci şi nouă).*

7. Alte semne de punctuație: " ( ) [ ] { }

Alte semne de punctuație ce pot fi utilizate în textele obişnuite sunt ghilimelele (sau semnele citării) şi parantezele rotunde; ele semnifică de obicei un citat, reprezintă porțiuni de text cărora li se dă un sens (stilistic) special sau asupra
• cărora autorul vrea să insiste, constituie traducerea ori sensul unui cuvânt, sau delimitează considerații personale introduse în text. Apar de obicei în perechi şi vor fi declarate caractere extra-textuale, servind modulului de analiză prozodică pentru obținerea unei vorbiri sintetizate cât mai naturale.

Parantezele drepte şi acoladele apar extrem de rar în textele româneşti uzuale; ele pot apare însă (ca şi parantezele rotunde) în expresii matematice. Se identifică simplu, deoarece sunt alăturate unor secvențe de cifre şi se expandează de obicei prin utilizarea cuvintelor corespunzătoare semnificației lor, cu ajutorul dicționarului de conversii în format lexical.

8. Secvențele de cifre

Secvențe de cifre pot apare şi în texte obişnuite, dar mai ales în expresii matematice, împreună cu semne de punctuație sau simboluri matematice; evident,

deoarece numărul lor posibil este practic infinit, ele trebuie expandate pe bază de regului de conversie, în funcție de context.

Am propus anterior o serie de regului pentru cazurile cele mai frecvente (numere întregi sau zecimale, numerale ordinale, ore, date). O situație specială (pe care de asemenea am menționat-o anterior), o reprezintă cazul în care o secvență de cifre, scrisă fără semne de punctuație, poate reprezenta fie un număr întreg, fie un număr de telefon. în acest caz, dacă din cercetarea contextului nu se poate elimina ambiguitatea (de exemplu prezența abrevierii *tel.),* această problemă rămâne în sarcina modului de analiză sintactică, care poate realiza o cercetare contextuală mai amplă.

9. Simbolurile matematice uzuale: + - *(sau x) ; (sau /) = <>% ~

Simbolurile matematice au o situație oarecum privilegiată, deoarece ele sunt încadrate de obicei de blancuri în expresiile matematice uzuale şi ca atare pot fi imediat identificate şi expandate pe baza dicționarului de conversii în format lexical (de exemplu *plus, minus, înmulțit cu, împărțit la* etc.) Dacă totuşi în scrierea expresiei nu apar blancuri, contextul secvențelor de cifre şi al celorlalte simboluri duc practic la aceeaşi rezolvare.

10. **Abrevierile**

O serie de considerații privind abrevierile au fost expuse anterior la regulile ce privesc punctul. Situația lor este dificilă datorită faptului că în limba română abrevierile se pot scrie în multe feluri: cu majuscule şi/sau minuscule, cu sau fără semne de punctuație (uzual punct).

Regula principală ce poate fi aplicată este evidentă:

- Dacă în grupul de caractere apare cel puțin un punct şi/sau cel puțin o majusculă, se caută în dicționarul de abrevieri; dacă secvența este găsită, abrevierea se expandează punând-o în corespondență cu cuvântul corespunzător din dicționar.

Pot rămâne însă ambiguități, în special pentru abrevierile scurte (de exemplu *V* - unitatea de măsură "volt", dar şi cifra romană "cinci" şi punctul cardinal "vest"), sau pentru abrevierile scrise cu minuscule şi fără punct *{km, cea, dl),* acestea din urmă nefiind căutate în dicționar (după regula expusă). Singurele soluții practice pentru rezolvarea unor astfel de cazuri ambigui este ca ele să fie preluate mai departe de analiza sintactică sau să fie recunoscute la etapa de conversie fonetică, prin căutarea într-un dicționar limitat de excepții.

11. **Acronimele**

Spre deosebire de abrevieri, cea mai mare parte a acronir stocate în dicționar, deoarece pronunția lor nu necesită inf suplimentare. De obicei, pronunția lor se reduce la citirea caracterelor ce compun acronimul, individual (ca pentru *S.R.L.),* normală a cuvintelor, atunci când pronunția lor s-a generalizat formă compactă *{NATO, TAROM)',* pentru citirea secvențială a ac nesesar doar un set de reguli simple de transcriere a literelor ros exemplu *S.R.L. - serele).*

Regula propusă pentru acronime este deci următoarea:

- Dacă secvența de caractere cuprinde cel puțin două este găsită în dicționarul de abrevieri, se caută în acronime:
  - dacă este găsită aici, secvența se expan echivalenței din dicționar;
  - dacă nu este găsită în dicționarul de acronime puncte, majusculele sunt (eventual) înlocuite c secvența nu va suferi altă prelucrare (se va citi ca dacă nu este găsită în dicționar, dar cuprinde p este expandată secvențial, utilizând un set mini transcriere a literelor rostite separat.

Pentru toate situațiile menționate, preprocesorul va ser modulelor ulterioare, pentru o corectă analiză sintactică şi prozodic

5. Concluzii

Am discutat în această lucrare câteva aspecte fundame sinteza automată a vorbirii, ca şi un număr important de reg generale pe baza cărora a fost proiectat preprocesorul de text pen în limba română. Nu am urmărit totuşi să descriem compl funcționarea şi implementarea acestuia; o serie de considerații totodată modalitatea concretă de implementare (pentru o variantă fost prezentate de autor în [13] şi [14].

în varianta actuală, preprocesorul de text a fost îmbu rezolva unele situații dificile legate de abrevieri, numerale urma măsură etc. De asemenea, un mecanism de automat de preprocesorului să fie "tolerant" cu anumite erori tipice de sinta exemplu fraze ce nu încep cu minuscule, sau un format "uşor" inc sau numerale.

Putem spune, ca o concluzie a celor discutate anterior, că un preprocesor de complexitate medie, cum este şi cel propus pentru sistemul TTS în limba română, poate rezolva cu succes (împreună cu analiza lingvistică ulterioară) o mare parte din problemele întâlnite într-un text obişnuit; el nu poate realiza însă normalizarea completă a **oricărui** text şi nu poate soluționa **toate** ambiguitățile care se pot ivi, datorate în special numărului extrem de mare al abrevierilor, acronimelor - în general a secvențelor neuzuale care pot apare într-un text scris. De asemenea, nu poate face față unor construcții negramaticale (deşi, de exemplu, unele simboluri speciale neaşteptate sunt ignorate).

Desigur că un set mai mare de reguli şi un dicționar de conversii în format lexical mai cuprinzător ar spori eficiența preprocesorului, dar este posibil ca el să devină atât de complicat, încât să fie practic neoperațional pentru un sistem TTS. Singura soluție practică pentru tratarea cazurilor ambigui este folosirea unui set minim de reguli, păstrarea în dicționar a celor mai uzuale situații (cu posibila adaptare **a** dicționarului la **tipul** textului ce se citeşte) şi examinarea cazurilor rămase la un nivel superior, pe baza plauzibilității sintactice, semantice sau pragmatice a frazelor obținute după preprocesare.

Referințe bibliografice

[1] R. Boite, H. Bourlard, T. Dutoit, J. Hancq, H. Leich (2000). Traitement de la parole. Presses Poliyechniques et Universitaires Romandes, 2000.

[2] G. Bailly (1996). Pistes de recherches en synthese de la parole - în "Fondements et perspectives en traitement automatique de la parole" (H. Meloni - Coord.), Aupelf-Uref, pp. 109-120, 1996.

[3] D. Burileanu (1999). Contribuții privind sinteza automată a vorbirii pornind de la text în limba română - Teză de doctorat. Universitatea "POLITEHNICA" Bucureşti, 1999.

[4] D. Burileanu (2002). Basic Research and Implementation Decisions for a Text-to-Speech Synthesis System in Romanian Language - Lucrare în curs de publicare în "International Journal of Speech Technology", Kluwer Academic Publishers, 2002.

[5] G. Fries, A. Wirth (1997). FELIX -A TTS System with Improved Preprocessing and Source Signal Generation - Comunicare la "EUROSPEECH'97", Rodos, pp. 589-592, 1997.

[6] E. Lewis, M. Tatham (1993). A Generic Front-End for Text-to-Speech Synthesis Systems - Comunicare la "EUROSPEECH'93", Berlin, voi. 2, pp. 913-916, 1993.

[7] M.Y. Liberman, K.W. Church (1992). Text Analysis and Word P Text-to-Speech Synthesis - în "Advances in Speech Signal P Furui, M. Sondhi - Coord.), Dekker, pp. 791-832, 1992.

[8] A. Lindstrom, M. Ljungqvist (1994). Text Processing wit Synthesis System - Comunicare la "International Conferen language Processing", Yokohama, pp. 139-142, 1994.

[9] M. McAllister (1989). The Probiems of Punctuation Ambiguity i Text-to-Speech Conversion - Comunicare la "EUROSPEECH 538-541, 1989.

[10] G. Beldescu (1984). Ortografia actuală a limbii române. E Enciclopedică, Bucureşti, 1984.

[11] T. Dutoit (1997). An Introduction to Text-to-Speech Sy Academic Publishers, 1997.

[12] F. Şuteu, E. Şoşa (1993). Dicționar Ortografic al Limbii Româ Bucureşti, 1993.

[13] D. Burileanu (1999). Natural Language Processing for Spee Romanian Language -Comunicare la "The 12th Internationa Control System and Computer Science", Bucureşti, voi. II, pp.

[14] D. Burileanu, C. Dan, M. Sima, C. Burileanu (1999). A Pa Preprocessor for Romanian Language TTS Synthesis "EUROSPEECH'99", Budapesta, voi. 5, pp. 2063-2066, 1999

Utilizarea tehnicilor nuanțate (fuzzy) şi de dinamică neliniară pentru sinteza adaptivă a vorbirii

Horia-Nicolai L. TEODORESCU
Academia Română, Secția Ştiința şi Tehnologia Informației,
Calea Victoriei 125, Bucureşti
E-mail: hteodor@etc.tuiasi.ro

1. Introducere

în timp ce maşina realizează tipic transmisie de date, omul comunică. Diferența constă în participarea intelectuală şi afectivă a persoanei la actul comunicării, participare reflectată atât la nivelul limbajelor neverbale (gestică, mimică etc), cât şi la nivelul vocal. Această participare afectivă dă varietate, coloratură şi sensuri suplimentare, nu neapărat pe plan semantic, semnalului vocal. Sinteza vocii, în prezent, este limitată de lipsa afectului, varietății şi sensurilor suprapuse în planuri multiple. Vocea maşinii rămâne astfel cantonată într-o regiune "moartă" a comunicării, este monotonă şi obositoare pe termen lung.

în această lucrare, reluând unele idei din [1-12], precum şi în contextul unor dezvoltări recente [13-27], în special legate de e-Voice şi VXML, prezentăm şi dezvoltăm unele concepte şi tehnici care ar putea permite maşinii atingerea deziderateIor mai sus menționate. Realizarea unor maşini capabile să mimeze calitățile vocii umane şi să *dialogheze* cu oamenii, sau măcar să comunice într-o manieră similară cu cea în care omul o face, este un deziderat în numeroase domenii, de la dialogul om-calculator, la sistemele auto şi la sistemele de învățare asistată de calculator [13-15]. Rezolvarea acestei probleme are implicații semnificative pentru acceptarea sintezei vocii într-o varietate de aplicații, de la robotică la realitate virtuală, la industria de jocuri electronice şi la protezare.

Prozodia, adică structura acustică ce se extinde pe mai multe segmente de semnal vocal, chiar peste mai multe cuvinte sau propoziții, implică ritm, accent, intonație, timbru, afect şi alte caracteristici ale vocii încă insuficient înțelese, sau vag definite în literatură. Informația paralingvistică ce este conținută de prozodie nu este nicăieri regăsită la nivelul "spus" prin cuvinte, dar - aşa cum am subliniat în [2] - această informație poate fi chiar mai importantă pentru ascultător decât informația lingvistică propriu-zisă. Incapacitatea sistemelor actuale de sinteză

vocală de a reda prozodia naturală este evidențiată chiar de marii producători de aplicații [25] şi este bine cunoscută în mediul cercetătorilor în domeniul sintezei vorbirii: *"One of the most difficult problems in speech to date is prosodic modeling"* [25].

2. Soluții pentru sinteza adaptivă şi varietală

Cele două calități ale vocii naturale, adaptivitatea - în sens larg - şi variabilitatea se pot realiza, cu costuri nu neapărat mari, la nivelul sintetizoarelor actuale, cu adaptări minimale (sau deloc) la nivel hardware şi cu îmbunătățiri ale programelor de control. Sinteza adaptivă se referă la adaptarea la:

- Condițiile sonore ambientale [1, 4];
- Contextul semantic-afectiv al cuvintelor şi frazelor sintetizate [2, 3];
- Interlocutorul sistemului de sinteză automată, atunci când acesta este recunoscut [2].

Sinteza varietală se referă la modificările inter-pronunție, la repetarea unor fraze, chiar şi în cazul în care condițiile ambientale şi contextul (şi interlocutorul) rămân neschimbate. Această variabilitate elimină monotonia şi personalizează vocea (naturală sau sintetizată), în măsura în care variabilitatea se face după reguli adesea proprii individului (cum este cazul în realitate) - şi nu doar aleatoare.

Variabilitatea intrinsecă a vorbirii derivă din mecanismele fizice de producere a semnalului vocal (curgere turbulentă a aerului prin organul fonator), precum şi din mecanismele neurologice de control al producerii semnalului vocal (controlul neuronal este cunoscut ca având o dinamică cu o importantă componentă neliniară). Aceste caracteristici au fost documentate de mai multe grupuri de cercetare, inclusiv de noi şi colaboratorii [5-9].

Adaptabilitatea şi variabilitatea în sensurile de mai sus vor fi prezentate sumar în secțiunile următoare, sintetizând lucrările citate şi unele cercetări mai noi, nepublicate încă.

3. Adaptabilitate la mediu

Una dintre cele mai elementare adaptări ale semnalului vocal generat de om este cea de adaptare la condițiile de mediu. Adaptarea la un mediu real, cu fond de zgomot, se realizează pe patru căi principale: prin modificarea amplitudinii semnalului (mai mare în mediul cu zgomot ridicat), prin modificarea spectrului (creşte contribuția frecvențelor înalte), prin modificarea ritmului (scăderea ritmului, creşterea duratei vocalelor), şi prin creşterea duratei dintre cuvinte, care devin separate, segmentate în timp. Adaptările realizate - instinctiv de un - se operează deci la un nivel relativ elementar, cu modificări minimale.

Realizarea acestei adaptări este esențială în multe aplicații vocii, incluzând sinteza vocală pentru aplicații în medii industriale şi transport, sau sinteza vocală pentru proteze laringiene. Este remarca adaptare se poate realiza, la pretenții reduse, cu foarte puțin har şi/sau cu un soft minimal, aducând însă o îmbunătățire esențială privința hardului, este necesar un canal de culegere a semnalul (semnal sonor ambiental).

Procesarea semnalului de zgomot, în vederea realiză sistemului de sinteză automată, presupune determinarea pute ambiental într-o fereastră temporală şi determinarea componențe semnalului ambiental. Primul parametru de caracterizare a zgomotulu medie aritmetică a pătratului semnalului s, într-o fereastră dată, de eşantioane şi caracterizată de momentul actual de timp, *n:*

$$w$$

$$k=0$$

Caracterizarea spectrală se poate realiza sumar prin rap puterea la frecvențe "înalte" (frecvențele înalte corespunzând, în m frecventă ce include formanții nr. 2, 3, 4 şi 5 din spectrul vocal) frecvențele "joase" (până la aproximativ al doilea formant, deci pâr de cea. 400 -500 Hz, ținând cont şi de vorbitorii feminini):

$$HL = \int_{500}^{10000} s^2(\omega)\,d\omega \Big/ \int_{500}^{} s^2(\omega)\,d\omega$$

Deoarece parametrii respectivi sunt relaționați cu impactul asupra inteligibilității vorbirii, deci sunt dați de calități subiective, abordăm o definire probabilistă sau fuzzy a lor. Dată fiind simplit nuanțat[1] (fuzzy), vom prefera a doua variantă. Un exemplu de defi de apartenență respective este prezentat în Figura 1. Este de presu definire să constituie doar un punct de plecare, îmbunătățirea realizându-se şi prin modificarea funcțiilor de apartenență.

[1] *Deşi nu este larg acceptat şi are o traducere mai dificilă în alte limbi termenul "nuanțat", propus de Grigore* C. *Moisil, în locul englezescului "fu*

[2] *Pentru a nu încărca prezentarea, ecuațiile funcțiilor respective sunt date î*

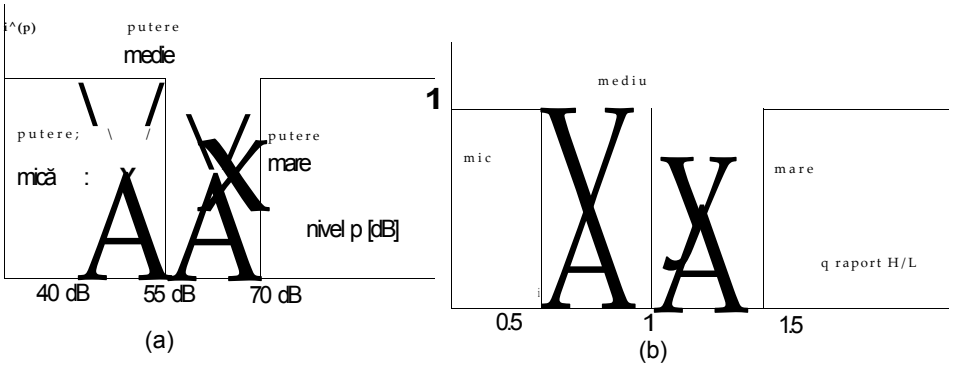


Figura 1. Funcțiile de apartenență ale premiselor regulilor folosite pentru determinarea modificărilor parametrilor de control ai sintetizorului

După cum s-a precizat deja, ca rezultat al aprecierii condițiilor de mediu, se controlează patru parametri ai semnalului sintetizat:

- creşterea amplitudinii (parametru notat AI)
- creşterea conținutului în frecvențe înalte (HFCI)
- creşterea duratei vocalelor (VLI)
- creşterea duratei dintre cuvinte (accentuarea segmentării pe cuvinte a frazei), notat IDBBW.

Controlul se realizează pe bază de reguli şi este rezumat în Tabelele 1-4 de mai jos.

Tabelul 1

Creşterea amplitudinii (AI - *Amplitude Increase)*

HL/P	mic	mediu	mare
mic	0,0	0,1	0,4
mediu	0,1	0,3	0,5
mare	0,4	0,5	0,6

Tabelul 2

Creşterea conținutului de frecvențe înalte
(HFCI - *High Frequency Content Increase - F3 increase)*

HL/P	mic	mediu	mare
mic	0,0	0,1	0,4
mediu	0,1	0,3	0,5
mare	0,4	0,5	0,6

Tabelul 3

Creşterea duratei vocalelor *(Vowel Length Increase* - VLH)

HL/P	mic	mediu	mare
mic	0,0	0,1	0,4
mediu	0,1	0,3	0,5
mare	0,4	0,5	0,6

Tabelul 4

Creşterea duratei dintre cuvinte
***(increase of the Duration of the Break Between Words* - DBBW)**

HL/P	mic	mediu	mare
mic	0,1	0,1	0..4
mediu	0,1	0,3	0,5
mare	0,4	0,5	0,6

Tabelele sunt interpretate în sensul uzual pentru sistemele nuanțate. Preferăm sistemele de tip Sugeno de ordin 0 (vezi Anexa 1), deoarece furnizează ca rezultat, direct, valori numerice, care vor fi interpretate ca şi coeficienți de multiplicare a valorilor nominale ale sintezei. De exemplu, prima linie şi prima coloană din Tabelul 1 spun că:

DACĂ Puterea (zgomotului) este ***medie*** *şi parametrul LH este* ***mediu***

ATUNCI Amplitudinea creşte de ***0,3*** *ori.*

Toate regulile din Tabelul 1 şi toate celelalte tabele se interpretează într-un mod similar.

Rezultatul final se obține prin agregarea rezultatelor parțiale, date de regulile respective. De exemplu, dacă valoarea intensității sonore este de 45 dB, iar raportul HL este de 0,7, prin aplicarea fuzzificării[3] se obține gradul de adevăr al premisei (combinate) din regula respectivă, prin

$$\text{Hlin}(\mu_{putere\text{-}mica}(P_0), \mu_{LH=mic}(LH_0))$$

unde P_0 =45, iar LH_0 =0,7 . Folosind expresiile funcțiilor (v. Anexa 1), se obțin valorile $\mu_{putere\text{-}mica}(P_0) = 0,67$, $\mu_{LH=mic}(LH_0) = 0,6$, deci valoarea minimă este 0,6 şi reprezintă gradul de încredere în faptul că amplitudinea creşte de 1,1 ori. Aceasta este valoarea de adevăr pentru singletonul (de la ieşirea sistemului) ce corespunde regulii respective, oc^. în total, sunt 9 reguli per tabel, deci există 9 valori de singletoni. într-adevăr, în acelaşi timp, valorile de intrare corespund funcțiilor de apartenență „mediu" pentru „putere" şi LH, deci regulei:

[3] *Termenul echivalent românesc ar fi "nuanțare".*

DACA *Puterea (zgomotului) este* ***mică*** *şi parametrul LH este* ***mic***
ATUNCI *Amplitudinea creşte de* ***0,0*** *ori.*

cu gradul de încredere în rezultat:

$$*putere$$

precum şi regulilor:

DACĂ *Puterea (zgomotului) este* ***mică*** *şi parametrul LH este mediu*
ATUNCI *Amplitudinea creşte de* ***0,1*** *ori.*

respectiv:

DACĂ *Puterea (zgomotului) este* ***medie*** *şi parametrul LH este* ***mic***
ATUNCI *Amplitudinea creşte de* ***0,1*** *ori.*

cu gradele de încredere

$$\wedge(\wedge putere \wedge micaMv \qquad LH\text{-}medi\acute{c}(Mo))$$

şi respectiv

$$^{min}\{v_{Pu}tere=medie(^{p}o\backslash V> \qquad LH\text{-}mic$$

Celelalte cinci reguli din Tabelul 1 au gradele de încredere în rezultat nule, deoarece valorile funcţiilor de apartenenţă „mare" ale premiselor („puterea este mare" şi „LH este mare") sunt nule, pentru valorile date, $P_0 = 57$ şi $LH_0 = 0{,}7$.

Prin agregare (defuzzificare), considerată aici conform formulei uzuale:

$$\pounds \; ^{a} \wedge (* o) \qquad (3)$$

5 X W

se obţine valoarea de ieşire (amplitudinea, creşterea conţinutului de frecvenţe înalte, creşterea lungimii vocalelor, respectiv creşterea duratei pauzei dintre cuvinte). în relaţia de mai sus, reprezintă abscisele singletonilor de ieşire din sistemele tip Sugeno respective, ju^Q reprezintă gradele de încredere în concluziile regulilor respective, iar *y* reprezintă valoarea agregată (defuzzificată) de ieşire a sistemului Sugeno. Sumarea se face pentru toţi singletonii de ieşire (notaţi de la 1 la 9). Indicele *"A"* arată că ne referim la parametrul controlat „amplitudine", controlui fiind desigur diferenţiat pentru cei patru parametri discutaţi.

Valorile astfel obţinute sunt folosite, cum s-a precizat, multiplicare a parametrilor nominali[4]. De exemplu, dacă amplitudi este AQ , atunci, prin aplicarea controlului, amplitudinea efectivă a se

$$A = A> \quad {}^{1+k=\backslash}$$

Sistemul de control este instantaneu, în sensul că nu ţine valorile recente (din fereastra prezentă, de lărgime *W)* ale zgomot valorile anterioare. Controlul de amplitudine şi frecvenţă se poate ex sintetizorului propriu-zis, asupra unui amplificator şi a unui filtru plas sintetizorului. Aceste două controale se pot prevedea de altfel şi în precum sisteme de sonorizare mari (eventual distribuite, ca în caz unor spaţii mari, gen pieţe sau stadioane), sau sisteme de sonoriz exemplu, sisteme de interfonie). Controlul pauzelor dintre cuvinte şi al spectrului vocalelor necesită comanda directă a sintetizorului.

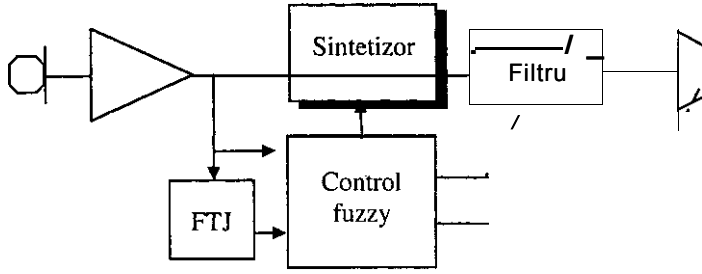


Figura **2.** Schema bloc a unui sistem audio adaptiv la zgomotul

în cazul în care se utilizează doar primele două tipuri de amplitudine şi spectral, adaptarea se poate realiza şi cu mijloace har sintetizorului, putând, de altfel, fi utilizată în orice aplicaţie audio (etc). Schema unui asemenea sistem de adaptare este cea prezenta o variantă fiind iniţial propusă în [4].

[4] *Nominali, în sensul că sunt valorile standard pentru sistemul de sinteză aut şi pentru sunetul respectiv produs în condiţiile contextuale date.*

4. Adaptare şi variabilitate contextual-interpretativă

Interlocutorul uman răspunde cu afect, după cum consideră anormală, nepotrivită, sau oricum în alt fel "departe de aşteptări" întrebarea sau afirmaţia făcută de partenerul la dialog. De asemenea, răspunsul este diferit atunci când vorbitorul uman este nesigur de răspuns, are un interes special în răspuns sau în topica discuţiei, sau, din contra, este dezinteresat. în plus, situarea interlocutorului faţă de partenerul sau partenerii de dialog, în context social sau afectiv, tonalizează discursul verbal şi îi imprimă specificitate relativă. Toate aceste caracteristici participative, precum şi altele asemenea, dau *comportamentul verbal* al omului, sunt traduse în mare măsură la nivelul semnalului vocal prin prozodie, dar în prezent nu se regăsesc la nivelul maşinii. Privitor la elementele de bază privind prozodia, vezi [26].

Pentru a implementa un comportament verbal, maşina trebuie să dispună de o bază de cunoştinţe minimală prin care să genereze acest comportament. De exemplu, este necesar să se interpreteze "departe de normal" într-o aserţiune sau întrebare a interlocutorului uman. Deci, vom presupune că există o bază de cunoştinţe care permite o asemenea interpretare. Construcţia acestei baze de cunoştinţe depinde de domeniul în care se poartă dialogul. în aceste condiţii, accentul va fi mai puternic pe anumite părţi ale frazei, sau răspunsul va depinde de aserţiune sau întrebare. Modul de răspuns va fi dirijat de asemenea de o bază de cunoştinţe, care include regulile necesare modificării sintezei (vezi Figura 3).

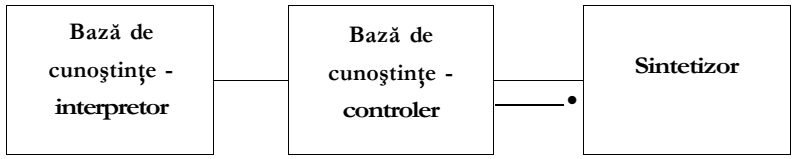


Figura **3.** Schema de principiu a controlului contextual-interpretativ

Baza de cunoştinţe-controler poate de asemenea fi implementată cu reguli Dacă... *Atunci,* de exemplu, de forma:

DACĂ *oferta / răspunsul interlocutorului este neaşteptat (negăsit în baza de cunoştinţe - baza de aşteptare/ baza de cazuri),*

A TUNCI *afectul sintezei este mirare / neîncredere/.../ etc.*

ori

DACĂ *oferta / răspunsul interlocutorului este neaşteptat negativ (conform bazei de cunoştinţe),*

A TUNCI *afectul sintezei este mirare şi/sau furie.*

Folosind rezultatele regulilor de acest fel, se pot seta parametrii ierarhic inferiori, de tonalitate, ai vocii sintetizate, pe baza acestora generându-se parametrii efectivi de control ai sintezei (amplitudine, frecvenţe formanţi etc).

Deşi acest gen de control poate părea complicat, sunt situaţii destul de generale în care el se poate implementa cu un efort relativ redus. De exemplu, atunci când se determină (printr-o măsurătoare relativ simplă, de frecvenţă medie vocal, sau de fundamentală) că interlocutorul este un copil sau o persoană de gen feminin, se poate selecta una sau ambele dintre alternativele:

- sistemul de sinteză automată se setează pe o voce de acelaşi tip (copil/feminin)
- sistemul de sinteză automată se setează pe voce "caldă" şi "vorbire clară".

Utilitatea şi modalitatea de realizare a primei setări nu necesită explicaţii. A doua setare (care poate fi simultană cu prima) se justifică - în cazul interlocutorului copil - prin necesitatea de a îi crea un mediu afectiv propice şi liniştit de dialog (voce "caldă") şi prin necesitatea unei comunicări cât mai informative, uşor de urmărit. Pentru a obţine o voce "caldă", se pot folosi trasee melodice cu variaţii lente precum şi frecvenţe mai joase ale formanţilor şi lărgimi mai mari (în zona spre frecvenţe joase) ale spectrelor formanţilor. "Claritatea" vocii se poate traduce prin segmentarea mai pronunţată pe cuvinte, precum şi vocale mai lungi (cu sau fără accentuări ale spectrelor formanţilor). Utilizarea unor asemenea adaptări - ce rămân în mare măsură să fie concepute în detaliu, implementate şi testate - este neîndoielnic mare la sinteza pentru procese educative [15, 26], în aplicaţii medicale (răspuns sintetic destinat pacienţilor), precum şi în numeroase aplicaţii generale (de exemplu, sintetizoare utilizate în muzee, pentru prezentarea exponatelor).

Alte modalităţi de personalizare afectivă sunt colorarea frecvenţială şi în amplitudine a anumitor părţi din frază sau în cadrul unui cuvânt, aceste modificări locale fiind larg documentate în literatură, de ex. [16-18] şi fiind relativ uşor de implementat.

5. Variabilitate prin metoda modulării de către un sistem dinamic neliniar

Variabilitatea semnalului vocal uman este bine cunoscută [5-9], [19-26]. Variabilitatea de tip natural a semnalului vocal sintetizat se poate obţine prin modularea diverselor controale (al amplitudinii, lungimii vocalelor, accentului, pitch-ului etc) cu semnale lent variabile, generate de sisteme care prezintă dinamică neliniară (haos). Parametrii sistemului haotic respectiv pot modela un anume subiect; considerăm aici că aceşti parametri reprezintă individul vorbitor şi

"personalitatea" lui. Această metodă, propusă de noi inițial în 1992 ([28] ş.a) dar nepublicată în forma extinsă, credem că reprezintă o' metodă promițătoare de "personalizare" a vocii.

Considerăm un sistem dinamic neliniar, dependent de parametri; semnalul în timp generat de acesta este de forma *x(t)= x^X^,...^),* unde ^repre-zintă parametrii sistemului haotic şi permit modelarea specificității vorbitorului Semnalul x poate fi folosit în modularea amplitudinii, frecventei fundamentale sau spectrului. semnalului vocal sintetizat. De exemplu, spectrul poate fi modificat folosind o lege de variație a frecvenței centrale a formanților, de forma-

$$fjtfH + xM\text{-}fjoi') \quad (5)$$

unde *fj(t)* este frecvența formantului numărul *j* la momentul *t, x(t)* este semnalul haotic respectiv (x,(0«l), iar $f_{JQ}(,)$ este frecvența "nominală" a formantului respectiv.

Un exemplu simplu de sistem haotic ce poate fi folosit în acest scop aste dat de ecuațiile:

$$r_{n+l} = X_3 \bullet u \setminus + X_2 \bullet u \setminus \quad (6)$$

unde setul de coeficienți $(x_{o>}A_{1},\ldots_{)}X_{5})e\ \mathbf{R}^{6}$se alege în domeniul de valori ce corespunde unui comportament haotic al sistemului (vezi Anexa 2). Setul de coeficienți $(^{}_{o},X_{„}\ldots,A_{,5})$ se poate seta specific pentru fiecare sistem de sinteză automată, "personalizând" sistemul. Valorile de ieşire ale generatorului se scaleaza corespunzător şi se folosesc la modularea unuia dintre parametrii de sinteza. Pentru exemplul din secțiunea 3, amplitudinea semnalului sonor devine, prin utilizarea modulației haotice:

$$1 + 1=L\frac{}{} \bullet(1 + K\ T\ J \quad (7)$$

unde **K** este un coeficient de scalare a seriei de timp r_{n}. Coeficientul **K** se alege astfel încât contribuția termenului $K\text{-}r_{n}$ să fie de ordinul procentelor $(KT_{„} <0,1\ V/l)$.

Desigur, scara de timp a procesului de generare de eşantioane de semnal vocal diferă de scara de timp a proceselor haotice folosite în modulație, ceasul celui de al doilea proces fiind mult mai lent (de ordinul 1/100) decât al primului proces. Pentru evitarea tranzițiilor bruşte ale parametrului controlat, valorile generate pot fi interpolate şi se poate realiza o variație lentă între două valori succesive. Considerând că un eşantion al seriei haotice r_{n} este generat la fiecare Q eşantioane de semnal vocal, seria r_{n} se poate înlocui cu seria (mai "fină", după ceasul de generare a esantioanelor semnalului vocal):

$$\frac{'/i\text{-}i}{Q} \qquad * = 0,1,\ldots,Q \quad (8)$$

în scopul modulării haotice a mai multor parametri de sinteză (amplitudine, frecvența centrală a formanților, lărgimea formanților, elemente prozodice etc) sunt necesare mai multe generatoare haotice, câte unul pentru fiecare parametru controlat. Alternativ, se poate folosi un sistem nuanțat (fuzzy) haotic, aceste sisteme generând simultan un număr mare de ieşiri necorelate sau slab corelate [28].

6. Concluzii si discuții

Adaptabilitatea şi variabilitatea sistemelor de sinteză a vocii şi ale celor audio, în general, se pot asigura prin modificări relativ simple hard şi soft ale sistemelor actuale. Adaptabilitatea se poate manifesta atât în raport cu mediul sonor, cât şi în raport cu contextul sau cu interlocutorul. Ideea de adaptabilitate şi metodele respective au fost introduse de noi în urmă cu peste 20 de ani şi dezvoltate continuu în lucrările citate, atât pentru aplicații de uz general, cât şi pentru aplicații medicale.

O aplicație de interes medical-educațional este utilizarea unor sisteme de învățare a unei limbi pentru copii de vârste mici (1 lună - 3 ani), care suferă de deficiențe de auz. Utilizarea unor sintetizoare cu spectru şi amplitudine controlate astfel încât să fie optim adaptate auzului (curbei de sensibilitate audiometrică) a fiecărui copil în parte ar ajuta asemenea copii să învețe limba la această vârstă. Este, într-adevăr, demonstrat că învățarea primelor elemente ale unei limbi la aceste vârste asigură o şansă mult mai mare de învățare a limbii ulterior şi de inserare socială [24].

Lucrarea prezentă se situează într-un context mai larg, în cadrul cercetărilor realizate de diverse colective care caută soluții pentru a face vocea sintetică purtătoare de informație emoțională. Astfel, în [31] se descrie o metodă de sinteză a "vocii emoționale", capabilă să transmită trei emoții (supărare-furie, bucurie, tristețe) folosind elemente de prozodie şi segmente de tip vocală

consoană-vocală (specifice limbii japoneze). în [32], starea ("mood") şi personalitatea sunt văzute ca elemente esențiale apărând în subsidiar în voce şi necesar a fi introduse şi în vocea sintetizată. Alți autori [33] vorbesc de "nivelul de plăcere al audiției" (pleasantness) - dincolo de inteligibilitate - şi văd naturalețea vocii sintetizate prin această prismă, a utilizării, la un nivel semnificativ, a prozodiei ("...we need to know more about how prosody could be utilized in human-computer interaction. We believe that we could borrow a lot from professional human speakers. Furthermore, speech applications should be built in a way that makes it possible to use prosodic feâtures efficiently.").

Credem că, în viitor, o metodă comodă de a genera automat prozodia, pentru o voce artificială dată şi pentru o anumită stare, ar putea fi constituită de o procedură inversă celei descrise în [34].

încheiem cu un citat din [35]: "... in spițe of the long history of speech synthesis, no one speech synthesis system available today is able to produce speech that could be characterized as natural or completely pleasant. In order to improve the speech quality of current text-to-speech (TTS) systems in terms of naturalness, three areas must be addressed[5]: 1) improved linguistic analyses, 2) improved prosody modeling, and 3) improved speech synthesis models."

Mulțumiri. Această lucrare a fost realizată cu sprijinul material al Academiei Române - Institutul de Informatică Teoretică laşi - precum şi cu sprijinul material parțial al Societății "Tehnici şi Tehnologii" s.r.l. laşi. Autorul mulțumeşte colegilor Dragoş Burileanu, Bogdan Branzilă şi Oana Geman pentru sugestii şi corecții la o formă preliminară a lucrării.

Referințe bibliografice

[1] Teodorescu H.N., Chelaru M., Sofron E., Adăscăliței A.: Adaptive speech synthesis. în voi. Digitale Sprach-verarbeitung - Prinzipien und Anwendungen. VDE Verlag, Berlin (W), pp. 183-188, 1988

[2] Teodorescu H.N.: Interrelationship, Communication, Semiotics, and Artificial Consciousness. în: Kitamura, T. (Ed.): *What Should be Computed to Understand and Model Brain Functions?* FLSI Book Series, voi. 3, World Scientific, 2000

[3] Teodorescu H.N.: Computer semiotics: understanding meanings and parallel languages (Refereed invited paper), Proc. Int. Conf. IIZUKA'98, Japan, 1998

[4] Teodorescu H.N.: Making speech synthesisers noise-adapta *Engineering* (UK), July 1987, p. 23

[5] Rodriguez, W., Teodorescu H.N., Grigoras Fl., Kandel A., Bun information space approach to speech signal nonlinear a *Intelligent Systems* (Wiley), Dec. 1999

[6] Grigoras Fl., Teodorescu H.N., Apopei V.: Nonlinear Analysis of Speech. *Studies in Informatics and Control,* voi. 7, no. 1, Ma 57-72

[7] Teodorescu H.N., Grigoras Fl., Apopei V.: Nonlinear proces production. *Int. J. Chaos Theory and Applications,* voi. 2, no. 2 52

[8] Teodorescu H.N., Grigoras Fl.: Nonlinear Techniques in Speech Proc. International Conference on Intelligent Technologies in Sciences, ITHURS'96. July 5-7, Leon, Spain. Voi. 2, pp. 293-298,

[9] Grigoras Fl., Teodorescu H.N., Apopei V.: Analysis of nonstationary processes in speech production, IEEE 1997 Applications of Processing to Audio and Acoustics. Mohonk M New Paltz, New York, October 19-22,1997 (IEEE Catalog # 9

[10].Burlui V., Teodorescu H.N., Morarașu C.S.: La fonction p l'edente total. Analyse en frequence. *Les Cahiers de Prothese* 88, Decembre 1994, pp. 63-68 1994

[11] Teodorescu H.N. et al.: Fuzzy models in speech analysi application, în Book of Summaries Int. Conf Modelling and Simu Turkey, July 1988, voi. 1, p. 162 (Summary)

[12] Teodorescu H.N., L. Buchholtzer, Chelaru M., Teodorescu prosthesis based on perilaryngean reflexes, Proc. 9th Int. EM Boston. Voi. 4, IEEE, pp. 2114-2115, 1987

[13] Anonymous Automotive Industry OEM/Supplier: Talking to talking to humans 7/12/2000. http://www-nrd.nhtsa.dot.gov/ 13/driver-distraction/Topics013040293.htm#A293

[14] Anne-Marie Derouault, The Future of Speech Recognition. E recognition technology is driving transparent computing, mak people to interact with computers. http://www.advisor. ID/OA000107.DERO01

[15] House D., Bell L., Gustafson K. & Johansson L. Child- synthesis: evaluation of prosodic variation for an educa program. Proc of Eurospeech'99, pp. 1843-1846, 1999

*Aici, autorul citat face referire la L R. Rabiner, "Applications of Voice Processing to Telecommunications,"*Proc. IEEE, *voi. 82, pp. 199-228, February 1994.*

[16] Heldner M., Strangert E. & Deschamps T.: Focus detection using overall intensity and high frequency emphasis. In: Andersson R, Abelin Â, Allwood J & Lindblad P, eds. Proc of Fonetik 99; pp. 73-76,1999.

[17] Heldner M., Strangert E. & Deschamps T.: A focus detector using overgll intensity and high frequency emphasis. Proc. of ICPhS-99, pp. 1491-1494, 1999.

[18] Heldner M.: On the non-linear lengthening of focally accented Swedish words. In: W. yan Dommelen & T Fretheim, eds. Nordic Prosody: Proc of the VIIIth Conference, Trondheim 2000 . Frankfurt am Main: Peter Lang. 2001

[19] Karlsson I., Banziger T., Dankovicovâ J., Johnstone T., Lindberg J., Melin H., Noian F. & Scherer K.: Within-speaker variability due to speaking mannfcrs. Mannell RH & Robert-Ribes J, eds. Proc. of ICSLP98, 2379-2382.1998

[20] Karlsson I.: Within-speaker variability in the VeriVox database. In: Andersson R, Abelin Â, Allwood J & Lindblad P, eds. Proc. of Fonetik 99, pp. 93-96, 1999.

[21] Karlsson I, Banziger T, Dankovicovâ J, Johnstone T, Lindberg J, Melin H, Noian F, Scherer K (1998), Within speaker variation due to induced stress, Proc Fonetik-98, 150-153. www.ling.su.se/fon/publications/fonetik98/

[22] Gustafson-Gapkova S & Megyesi B.: A Comparative Study of Pauses in Dialogues and Read Speech. Proc. of Eurospeech 2001, pp. 931-935, 2001

[23] Beskow J.: A tool for teaching and development of parametric speech synthesis. In: Branderud P & Traunmuller H (eds). Proc. of Fonetik -98, pp. 162-165.1-98,1998

[24] Rachel I. Mayberry, Elizabeth Lock, Hena Kazmi: Linguistic ability and early language exposure. *NATURE,* Voi. 417, 2 May 2002, p. 38, 2002

[25] Mircrosoft Co.: Platform SDK: Agent. Characters. http://msdn.microsoft. com/library/default.asp?url=/library/en-us/msagent/deschar__8nn6.asp

[26] Mauricio Lumbreras, Gustavo Rossi: Metaphor for the Visually Impaired: Browsing Information in a 3D Auditory Environment. CHT95 Proc, www.acm.org/sigchi/chi95/proceedings/shortppr/mLbdy.htm

[27] Christophe d'Alessandro & Jean-Sylvain Lienard: 5.2 Synthetic Speech Generation. In: Survey of the State of the Art in Human Language Technology. http://cslu.cse.ogi.edU/HLTsurvey/ch5node4.html#SECTION52

[28] Teodorescu H.N.: Chaos in fuzzy systems and signals. Voi. Proceedings of the 2nd Int. Conf. on Fuzzy Logic and Neural Networks. Voi. 1, pp. 21-50 (Jono Printing Co., 1992, Iizuka, Japan)

[29] Teodorescu H.N., Kandel A., Jain L. C. (Eds.), Fuzzy and Neuro-Fuzzy Systems in Medicine (International Series on Computațional Intelligence). CRC Press, Boca Raton, USA, 1998.

[30] Teodorescu H.N., Mlynek D., Kandel A. (Eds.): Intelligent Systems and Interfaces (The Kluwer International Series In Intelligent Systems). Kluwer Publ., Boston, 2000.

[31] Yasuhisa Niimi, Masanori Kasamatu, Takuya Nishimoto and Masahiro Araki: Synthesis of Emoțional Speech Using Prosodically Balanced VCV Segments. http://www.ssw4.org/papers/133.pdf.

[32] Nick Campbell: WHERE IS THE INFORMATION IN SPEECH? (and to what extent can it be modelled in synthesis?) www.slt.atr.co.jp/cocosda/jenolan/ Proc/r82/r82.pdf.

[33] Hakulinen J., Turunen, M.: Prosodic Feâtures for Speech User Interfaces. www.cs.uta.fi/hci/spi/reports/Prosodi^ .pdf.

[34] Ansgar Rinscheid: Voice Conversion Based On Topological Feature Maps and Time-Variant Filtering. www.asel.udel.edu/icslp/cdrom/vol3/235/a235.pdf.

[35] Syrdal A., Stylianou Y., Garrison L.+, Conkie A. Schroeter J.: Td-Psola Vs. Harmonic Plus Noise model in Diphone Based Speech Synthesis. www.research.att.com/projects/tts/papers/1998 ICASSP/paperSYN.ps.

Anexa 1: Sisteme nuanțate de tip Sugeno, de ordin 0. Funcții de apartenență

Reamintim că o mulțime (clasică) A c X, unde X notează universul de discurs, este definită de o funcție caracteristică, de forma:

$$\chi_A(): X \to \{0, 1\}$$

$$\chi_A(x) = \begin{cases} 1 & daca \quad x \in A \\ 0 & daca \quad x \notin A \end{cases}$$

Prin generalizarea conceptelor de mulțime şi de funcție caracteristică, se definesc mulțimile nuanțate (fuzzy) şi funcțiile de apartenență corespunzătoare

astfel: o mulțime nuanțată, notată A, peste universul de discurs X, este caracterizată unic de o funcție de apartenență:

$$\mu_A(\ ): X \to [0,1]$$

în particular, funcția de apartenență poate fi de forma:

$$\begin{cases} 1 & pentru \quad x = a \in X \\ 0 & pentru \quad x \neq a \end{cases}$$

caz în care se numeşte *singleton.*

Un sistem de tip Sugeno, de ordin 0, este descris de reguli de forma:

DACĂ intrarea (premisa) # 1 ŞI premisa #2 ŞI... Şi premisa # n ATUNCI concluzia

unde premisele sunt de forma: x, *este* $\check{A}_y$, iar $\check{A}y$ sunt valoari nuanțate (fuzzy), de exemplu $\check{A}_n$ = "mare", $\check{A}_{n2}$ = "mediu", atributelor lingvistice "mare", "mic" etc. fiindu-le ataşate câte o funcție de apartenență. Specific sistemelor Sugeno este faptul ca în concluzie apar valori numerice si nu valori nuanțate, concluzia fiind deci de forma "y= 0,3" (singleton).

Definițiile funcțiilor de apartenență pentru intensitatea sonoră din Figura 1.a sunt:

$$\mu_{Putere\text{-}medie}(p) = \begin{cases} 1 & pentru \quad p < 40\ dB \\ 1 - \left(\frac{p - 40}{15}\right)^2 & pentru \quad 40 < p < 55 dB \\ 0 & pentru \quad p > 55\ dB \end{cases}$$

$$\mu_{Putere\text{-}medie}(P) = \begin{cases} 3 & pentru \quad p < 40 dB \\ P'' \left(\frac{p - 40}{15}\right)^2 & pentru \quad 40 < p < 55 dB \\ l - P \sim \left(\frac{p - 55}{15}\right)^2 & pentru \quad 55 < p < 10 dB \\ 0 & pentru \quad p > 10 dB \end{cases}$$

$$\mu_{Putere\text{-}mare}(P) = \begin{cases} 0 & pentru \quad p < 55\ dB \\ p \sim \left(\frac{p - 55}{15}\right)^2 & pentru \quad 55 < p < 10 dB \\ 1 & pentru \quad p > 10 dB \end{cases}$$

Definițiile funcțiilor de apartenență pentru raportul HL (Figura 1b) sunt:

$$\mu_{VHL}(q) = \begin{cases} 1 & pentru \quad q \geq 0.5 \\ 0 & pentru \quad p > \backslash. \end{cases}$$

$$\mu_{HL}(q) = \begin{cases} 0 & pentru \quad q < 0.5 \\ \left(\frac{q - 0.5}{0.5}\right)^2 & pentru \quad 0.5 < ? < 1.0 \\ 1 - \left(\frac{q - 0.5}{0.5}\right)^2 & pentru \quad l. < ? < 1.5 \\ 0 & pentru \quad p > l.5 \end{cases}$$

$$\mu_{HL}(q) = \begin{cases} 0 & pentru \quad q < \backslash. \\ \left(\frac{q - 0.5}{0.5}\right)^2 & pentru \quad 1.0 < q < 1.5 \\ 1 & pentru \quad q > l.5 \end{cases}$$

Pentru detalii asupra manipulării funcțiilor de apartenență şi a regulilor în sistemele nunațate, a se vedea orice manual în domeniul sistemelor fuzzy, sau volume precum [29, 30] în care se pot găsi şi aplicații specifice legate de înțelegerea vorbirii, sau alte aplicații medicale.

Anexa 2: Procesul haotic

Procesul reprezentat de ecuațiile (7) are o dinamică haotică doar pentru anumite subintervale relativ înguste din $\mathbf{R}^6$. în restul spațiului, comportamentul este asimptotic instabil (peste tot pentru valori ale coeficienților lui r^3 mai mari ca 1, în modul, dacă şi coeficientul lui u este mai mare ca 1 în modul); comportamentul este stabil sau periodic pentru alte zone, relativ reduse din $\mathbf{R}^6$.

Diagrama de bifurcație a procesului, aşa cum apare în Figura A1, este obținută pentru: valorile coeficienților [Q]={.1, -.17, -.18, .1}; coeff__4 = 1.1; coeff__5= -.15; condiție inițială r[0]= 0.3; număr total de puncte în diagrama de bufurcație: 500 (punctele de la 500 la 1000); regimul tranzitoriu eliminat: primele 500 puncte; precizia tuturor coeficienților şi variabilelor: double.

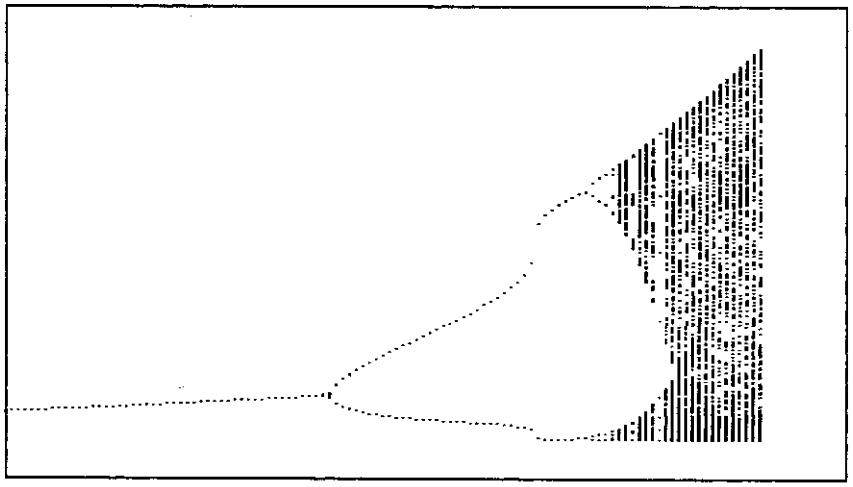


Figura A2-1. Diagrama de bifurcație a procesului

Legile folosite (conform codului, scris în limbajul C) sunt:

u[n]= (coefL4)*r[n]+coefL5 -0.005*(float)k;

x=u[n]; r[n+1]=poly(x, Q, coeff);

(Q este numărul de valori în vectorul coeficienților, Q=4)

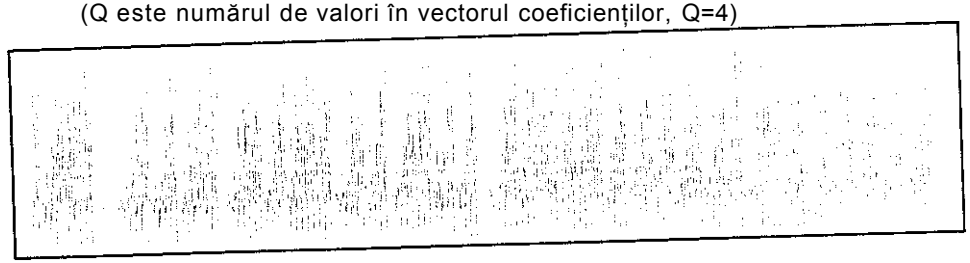


Figura A2-2

Semnalul în domeniul amplitudine-timp din Figura A2-2 a fost obținut pentru ecuațiile (cod C):

u[n]= coeff_4*r[n]+coefL5-0.05*21.;

x=u[n]; r[n+1]=poly(x, Q, coeff);

Semnalul obținut pentru valoarea u[n]= coeff_4*r[n]+coeff_5-0.05*21. (restul programului fiind identic ca pentru cazul anterior) este ilustrat în Figura A2-3.

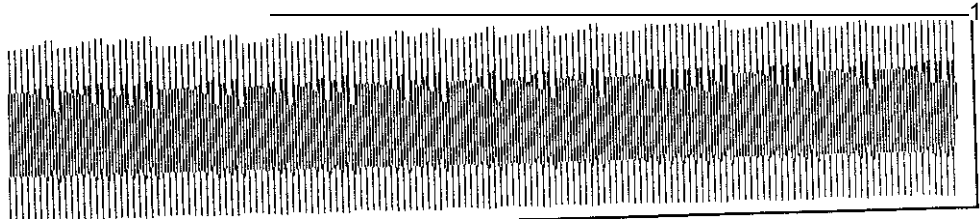


Figura A2-3

Semnalul obținut cu u[n]= coeff_4*r[n]coeff_5-0.05*20.7, precum şi la o scară dublă de timp, este ilustrat în Figura A2-4:

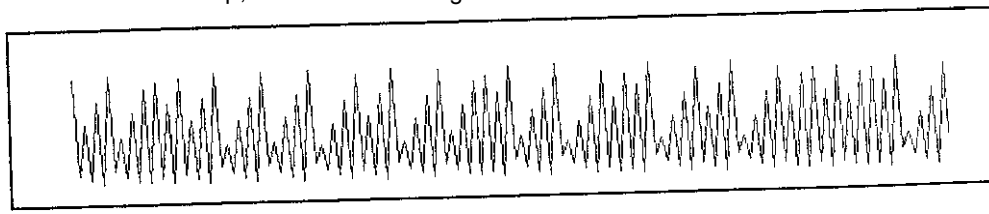


Figura A2-4

Regiunile spațiului parametrilor în care sistemul este stabil, după cum s-a spus deja, sunt relativ înguste. Pentru parametrii coeffJ-coeff_4 fixați si coeficientul coeff_5 variabil între -25.15 şi + 4.85 (600 de pasi, cu pas 0 05) doar zona îngustă din Figura A2-5 este stabilă, oscilantă sau haotică, în rest sistemul fund asimptotic instabil. Pentru uşurința urmăririi scării, linia din partea de jos a figuri, reprezintă intervalul menționat, [-25.15, + 4.85]), în care s-a testat sistemul

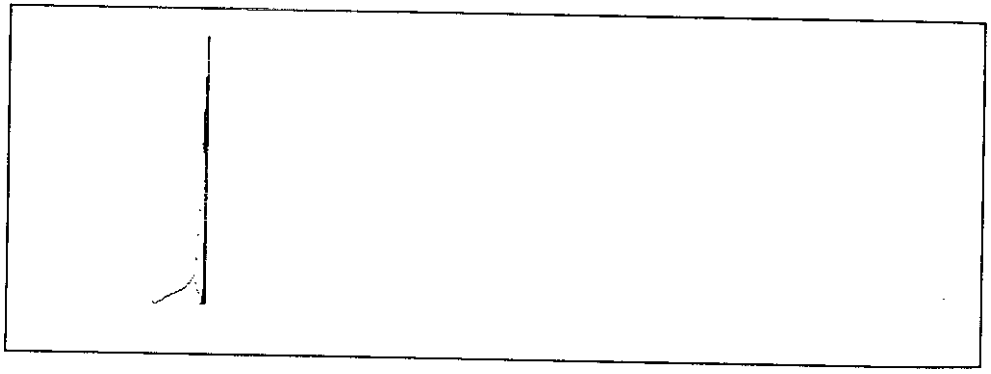


Figura A2-5

în figură, se poate remarca diagrama de bifurcație a sistemului, cu zonele de stabilitate, oscilație şi haos. Pentru restul intervalului, prin program calculele sunt abandonate, deoarece valorile de ieşire ale sistemului depăşesc, în valoare absoluta, 10000.

Dumitru TODOROI, Diana MICUSA, Zinaida TODOROI, Ion LINGĂ, Ion COVALENCO, Nicolae OBJELEANU, Ştefan SPĂTARU, Stela LUNGU, Virginia ȚURCANU, Elana COZLOV, Nadejda AMBROZII, Victor SLOBODEANU, Igor COŞERU, Cătălina SURUCEANU
Academia de studii economice din Moldova, Str. Banulescu-Bodoni, 59-61/503»B», Chişinău MD 2005, Republica Moldova,
E-mail: todoroi@ase.md

Lucrarea actuală realizată în cadrul achiziționării Marelui Dicționar al Limbii Române (MDLR) în format electronic a fost medodologic înfluențată de ideile subliniate în [1-3] şi este o continuare a cercetărilor [4-7,10-11], efectuate în cadrul procesării limbajului natural. Au fost elaborate un şir de proiecte [8-9,12] de informatizare a Limbii Române. Experimentările cu elaborarea sistemelor computerizate de nivelul unu, orientate pe diferite sub-dicționare ale MDLR structurat pe următoarele caracteristici: TEXT, AUDIO, IMAGINI şi VIDEO, au început recent în Academia de studii economice din Moldova (ASEM) în colaborare cu ONG-ul ECO-INFO-MOLD. Unele rezultatele ale cercetării şi experimentării în cadrul platformei alcătuite din aceste 4 subsisteme, sunt expuse în lucrarea de față. Sunt prezentate diferite scenarii [19] şi metodologii de utilizare a sub-dicționarelor informatizate ale limbii române. Clarificarea mijloacelor Hardware-ului şi Software-ului modern, care pot suporta MDLR informatizat [18], constituie o problemă importantă pentru etapa creării Societății Informaționale - Societate a Cunoaşterii [20].

I. Componenta TEXT a dicționarului economic MULTIMEDIA al limbii Române [23].

Scopul acestui compartiment computerizat al MDLR constă în crearea subsistemului TEXT de nivelul unu a unei părți introductive a dicționarului economic al limbii române şi experimentarea cu acest sistem. Acest dicționar economic constă din 35.000 - 40.000 cuvinte. Cuvintele conținute în Dicționarul Enciclopedic Ilustrat (DEI) [21] vor fi definite în-totalitate. Experimentarea cu subsistemul TEXT al MDLR computerizat este efectuată la momentul actual cu circa 200 articole din DEI.

Baza de date TEXT (BDT) a dicționarului economic constă dintr-o culegere de texte-articole, alcătuită din cuvinte, fraze, paragrafe, capitole etc. ale DEI. Documentele în BDT includ nu numai informații textuale (definiții de cuvinte), dar pot conține şi informații de alt tip, de exemplu, prin extindere, imagini. Prin urmare BDT în sistemul computerizat al MDLR conține nu numai materialul textual, dar şi ilustrativ: diagrame, grafice, fotografii etc.

Prin crearea subsistemului TEXT de nivel unu, utilizatorul obține un mijloc important, prin intermediul căruia informația poate fi introdusă şi utilizată în mod complementar pe cale electronică.

1.1. Capacitățile necesare ale unui sistem de gestiune a bazei de date MULTIMEDIA

MDLR este o bază de date MULTIMEDIA. Sistemul de gestiune al MDLR este un sistem de gestiune a bazei de date MULTIMEDIA (SGBDMM) şi constituie un mecanism care operează cu diferite tipuri de date, reprezentate într-o diversitate de formate pe un set larg de mijloace şi surse. Pentru a funcționa efectiv e necesar ca SGBDMM să posede următoarele capacități:

(a) Capacitatea de a interoga uniform datele reprezentate în diferite formate;

(b) Capacitatea de a interoga datele, reprezentate în diferite media;

(c) Capacitatea de a transmite subiectele media din dispozitivele de stocare locale într-un mod efectiv;

(d) Capacitatea de a primi răspunsul la o interogare şi de a realiza prezentarea acestui răspuns pe baza componentei audiovizuale;

(e) Capacitatea de a furniza această prezentare pe o cale adecvată, care ar satisface calitățile diferitor cerințe ale serviciului.

1.2. Structura bazei de date TEXT (BDT) a dicționarului economic MULTIMEDIA

Dicționarul economic, care este pe cale de a fi pus pe calculator, este o BDT cu posibilitatea de a fi extinsă cu diferite componente ale MULTIMEDIA. Subsistemul TEXT a dicționarului economic MULTIMEDIA este un subsistem al SGBDMM, care aprovizionează acesta posibilitate împreună cu utilizarea complimentară a BDT.

Structura BDT e compusă din:

(1) Indice cu caracteristica "număr";

(2) Termen principal (cuvânt, articol) cu caracteristica "text";

(3) Variantă(e), derivate, abreviere (concretizare) cu caracteristica "text";

1

(4) Categorie gramaticală cu caracteristica "text";

(5) Domeniu cu caracteristica "text";

(6) Definiții pentru termenul principal (şi concretizări) cu caracteristica "text";'

(7) Sinonim(e) cu caracteristica "text";

(8) Antonim(e) cu caracteristica "text" şi altele.

De asemenea BDT are posibilitatea de a fi extisă cu următoarele subdiviziuni MULTIMEDIA:

(9) Audio cu caracteristica „OLE";

(10) Imagini cu caracteristica „OLE";

(11) Video cu caracteristica „OLE" şi altele.

1.3. Scenarii de utilizări şi interogări ale subsistemului TEXT al MDLR informatizat

Interogarea este o formă de interacțiune care ajută utilizatorul să prezinte o informație anumită într-o structură anumită, definită de utilizator. Spre exemplu, utilizatorul doreşte să obțină informații din arhive, articole, sau alte documente, care conțin informația despre Uniunea Europeană. Interogarea poate avea următoarea formă: "Găseşte toate dosarele, legate de investițiile străine, făcute de UE în domeniul educației". Un simplu cuvânt-cheie al acestui dosar nu va permite găsirea răspunsului corect, chiar şi dacă indicile acestui document deja există. Totuşi, sistemul ne va prezenta unele cuvinte, legate de această interogare, dar este posibil ca acestea să nu poată fi direct asociate la tema dorită. De aceea textul trebuie să fie indexat nu numai pe cuvintele cheie, dar şi pe conținutul semantic şi/sau pragmatic al cuvintelor (în cazul BDT, de exemplu, concretizarea).

Soluționarea problemei utilizatorului, care doreşte să afle definiția cuvântului "Academie", de exemplu, cere introducerea polisemiei în BDT, care conține concepte ca precizia şi rechemarea. întrebarea, propusă de către utilizator în acest context, este: "Cum să aflu din baza de cunoştințe a MDLR sensul cuvântului "Academie - ca instituție de învățământ economic". Pentru aceasta BDT va fi completată cu o nouă coloană "concretizare", care va preciza şi va face posibilă afişarea pe monitor a acelei definiții a cuvântului, de care utilizatorul este interesat (de exemplu: *Academia de Studii Economice).*

Un fragment de structură schematică a BDT.

1 Indece cuvânt 1	1 Cuvânt	1 Concretizare	1 .... I Definiție	. I Traducere
1 03342	1 Academia	1 de studii economice 1	I Nume dat şcolii de ...I	I
1 14269 I	Banii I	EURO i	I Denumire a princip...!	I
J 14271 I	Banca I	de economii I	| , ,	

SGBDMM, ca o extindere a SGBD Ms ACCESS-2000, în baza căruia este creată componenta TEXT a MDLR informatizat, gestionează BDT, utilizând limbajul SQL. în exemplele următoare utilizatorul este interesat de sfera finanțelor. Accesul BDT a MDLR este efectuat prin intermediul următoarelor interogări din SQL (care, în general, constituie comenzile SUMMARING, JOIN, PROJECTION, DIVISION, SELECT şi altele):

```
Ex.1.  SELECT   Banii                    (termen principal, nume de interes)
       FROM     Ambrozii-Godzina         (nume de fişier)
       WHERE    Concretizare = EURO      (concretizare pentru termenul principal)

Ex.2.  SELECT   Academia
       FROM     Ambrozii-Godzina
       WHERE    Concretizare =  de studii economice

Ex.3.  SELECT   Banca
       FROM     Ambrozii-Godzina
       WHERE    Concretizare = de economii.
```

II. AUDIO-dicționarul explicativ economic al limbii Române [24]

Dicționarul explicativ economic MULTIMADIA al limbii române, ca o parte componentă al DEI, include circa 35000-40000 de cuvinte şi este divizat în compartimentele: Text, Audio, Video şi Imagini. Aceste componente MULTIMEDIA ale MDLR informatizat satisfac cerințele de bază ale unui dicționar informatizat prezintă formele exacte ale cuvintelor, accentul, etimologia, definiția-text, definiția-sunet (audio), definiția-video (film), definiția-imagine (grafic, schema, poza etc.) şi corespunde cerințelor unor categorii foarte largi de utilizatori, nu numai elevi şi studenți, dar şi funcționari şi profesionişti, contribuind la ridicarea nivelului de cultură.

Compartimentul AUDIO al MDLR informatizat furnizează informații necesare ale articolului respectiv (cuvântul, definiția lui) în forma AUDIO. Subsistemul AUDIO de nivelul unu al SGBDMM oferă posibilitatea de AUDIO utilizare a dicționarului. Acest AUDIO-dicționar va contribui din plin la ridicarea pe o treaptă superioară a societății noastre în utilizarea corectă a limbajului economic atât la nivel oral cât şi scris. Conținutul de date AUDIO poate fi caracterizat prin două metode: (a) folosind metadata prin explicarea conținutului unui fişier AUDIO sau (b) prin extragerea tipului potrivit de date AUDIO, folosind procesorul tehnic.

2.1. Componenta AUDIO a metadatelor

Cu un fişier AUDIO se procedează la fel ca în cazul unei date VIDEO acestui fişier i se asociază un set (grup) de segmente, toate referindu-se la o perioadă de timp. Fiecărui segment i se atribuie un set de activități, care au decurs în acea perioadă de timp, subliniate prin aceste segmente. în general, jnetadata utilizează reprezentarea AUDIO, care este sesizată ca un set de obiecte marcate în timp.

Utilizarea componentei AUDIO a metadatei din MDLR informatizat este recomandabilă mai ales ca o modalitate de creare şi de modificare a acestei metadate şi, îndeosebi, la interogarea AUDIO - dicționarului de către utilizatori care necesită această formă de comunicare om-maşină.

Crearea componentei AUDIO a metadatei este mai complexă decât crearea altor forme de dicționare informatizate, deoarece identitatea indivizilor ce vorbesc nu poate fi uşor cunoscută; de asemenea conținutul discursului poate fi neclar.

Conceptul despre conținut este descris în termeni de metadată a procesului. Ca rezultat, data AUDIO este considerată ca un semnal DELTA(x) în timpul x. Trăsăturile de utilizare ale acestui semnal DELTA(x) sunt: (a) extragerea (b) indicarea şi (c) depozitarea.

O undă constă dintr-un set de vârfuri (creste) şi adâncituri (văi). Perioada vibrației T este definită ca timpul în care o parte a undei revine la poziția inițială

Alte caracteristici utilizate de componenta AUDIO în crearea metadatei sunt: (1) frecvența, (2) viteza şi (3) amplitudinea.

Baza de date **AUDIO (BDA)** poate fi interacționată şi gestionată, utilizând sunetul auditiv prin intermediul secvenței de prelucrări: segmentare, memorizare şi extragere a informației.

2.2.2. *Segmentarea*

Segmentarea e o procedură de separare a semnalului audio în câteva ferestre egale. Această procedură poate fi utilizată conform următoarelor două metode:

(a) Utilizatorul specifică dimensiunile ferestrei, presupunând că proprietățile undei şi ale ferestrei se vor obține prin medie;

(b) Utilizatorul segmentează sunetul în acelaşi mod ca şi imaginile, folosind predicatul de omogenitate H.

2.2.2. *Extragerea*

La .extragere cel mai des utilizate sunt facilitățile de indicare a intensității, zgomotului, înălțimii şi clarității.

2.2. Unele sisteme de utilizare a BDA

Din punct de vedere a MULTIMEDIA, AUDIO - baza de date (BDA) poate fi interpretată ca o sursă auditivă, ca un fişier cu o fereastră auditivă şi cu trăsăturile respective, asociate aceastei ferestre.

Scenariile de utilizare a BDA cuprind toate formele MULTIMEDIA, care pot fi utilizate în diferite domenii. în sistemele comerciale, de exemplu, **Bazele de date Informix** includ bazele de date a sistemului managerial, care permit utilizatorului să acceseze baza de date, bazându-se pe nesiguranța conținutului.

Baza de date DB2, un alt exemplu, utilizată cu calculatorul de tip IBM, necesită cuplarea cu un sistem auxiliar, care permite lăsarea mesajelor vocale pe robot. DB2 poate importa şi menține clipurile, care pot fi căutate printr-un nume sau descriere.

Putem reasculta mesajele lăsate pe robot, prin intermediului **Internetului.** Un exemplu în plus îl constituie o utilizare a unui cuvânt din AUDIO-dicționarul economic al limbii române prin intermediul AUDIO-VIDEO-robotului, care este un sistem autorizat şi care acționează pe baza unui program de lucru stabilit sau care reacționează la anumite influențe exteriore.

Un exemplu de interogare a componentei AUDIO a subdicționarului economic al MDLR, prin indermediul limbagului SQL şi al subsitemului AUDIO de nivel unu al SGBDMM, poate avea forma:

```
SELECT      Robot
FROM        Țurcan-Mutruc
WHERE       Attribute IS Definiție AND Attribute IS Audio
```

Ca rezultat al acestei interogări utilizatorul prin intermediul răspunsului prietenos, obține pe ecran definiția TEXT a cuvântului Robot şi, paralel acest subsistem AUDIO al SGBDMM, difuzează acesta definiție cu voce feminină sau masculină (la dorința utilizatorului).

III. Subsistemul IMAGINI de nivel unu al dicționarului economic informatizat al limbii române [25]

Scopul acestui capitol constă în descrierea posibilităților de introducere a imaginilor în baza de date a MDLR informatizat şi de utilizare a acestora în viața cotidiană. Baza de date IMAGINI (BDI) a subdicționarului economic al MDLR informatizat constituie o componentă, care oferă posibilitatea de extindere a procesului de înțelegere a sensului cuvântului dat. Din cele aproximativ 35000-40000 de articole ale dicționarului economic din MDLR doar 50-60%, pot fi prezentate în forma de imagini, după părerea noastră.

Experiența, obținută pe baza câtorva zeci de articole din DEI în cadrul evaluării subsistemului IMAGINI al SGBDMM, ne confirmă întru totul conținutul zicalei: «Mai bine odată să vezi decât să auzi de o sută ori» şi al zicalei «O imagine este mai mult decât o mie de cuvinte». Aceste facilități de utilizare din evoluția procesului de creare şi utilizare a MDLR informatizat sunt confirmate şi de lucrările din [22] precum şi prin intermediul următorului Tabel 3.1, prezentat în original

Table 13.1.

Data rates and storage requirements per hour, day, and lifetime fora person to record all the text they've read, all the speech they've heard, and all the

Data type	data rate (bytes per second)	storage needed per hour and day	storage needed in a lifetime
Read text, few pictures	50	200 KB; 2-10 MB	60-300 GB
speech text @ 120 wpm	12	43 K; 0.5 MB	15GB
speech (compressed)	1,000	3.6 MB; 40 MB	1.2TB
video (compressed)	500,000	2 GB; 20 GB	1 PB

3.1. Baza de date IMAGINI (BDI)

Imaginea poate transmite mai multe informații despre un obiect decât câteva pagini (Vezi Tabelul 3.1) de descrieri textuale. Pentru un chirurg este cu mult mai uşor să-şi găsească un pacient potențial prin investigarea diferitor imagini. Imaginile pot fi combinate cu corpusuri, text-definiții, sunet-definiții, traduceri etc.

în afară de datele IMAGINI ale dicționarului economic MULTIMEDIA în MDLR informatizat vor fi prezente video, audio, document, manuscrise şi altele. Datele VIDEO sunt des folosite în domeniul învățământului. Datele AUDIO sunt importante în domeniul criminalisticii, de exemplu, în identificarea vocilor celor suspectați. Datele documentare diferă de datele TEXT prin aceea că pot conține nu numai informații textuale, dar şi imagini încadrate. Datele manuscrise se presupune că în viitorul apropiat vor prevala înregistrările electronice.

Sunt cunoscute diferite formate electronice, care dau posibilitatea de a vizualiza imaginea (fişierele de tip GIF, TIFF, PCX, de exemplu). Subsistemul IMAGINI a SGBDMM are anumite trăsături specifice necesității de utilizare a imaginilor ca o componentă vitală a MDLR informatizat.

3.2. Subsistemul IMAGINI

Baza de date IMAGINI diferă de bazele de date TEXT şi AUDIO prin complexitatea imaginilor, necesitatea de a diviza, combina şi utiliza diferite părți componente ale imaginii, care deseori la interogare se complică şi prin utilizarea incorectă şi analiza neprecizată a tehnicilor de manipulare a imaginilor. Aceasta se complică şi prin faptul că diferite organizații adună date fotografice, hărți, scheme şi alte imagini de tip universal sau specializat (cum ar fi, de exemplu, NASA). Interogările datelor de tip IMAGINI sunt efectuate în baza datelor de tip TEXT, căutate în baza de date de tip IMAGINI şi vizualizate în formă de text şi imagini. în final imaginile pot fi transferate în baza de date specializate, cum ar fi, de exemplu, încadrarea lor în baza de date MULTIMEDIA comerciale. în subsistemul IMAGINI al SGBDMM este prevăzut un set larg de proceduri cu imaginile.

3.2.1. *Plasa imaginii*

Conținutul imaginii constă din toate obiectele acestei imagini şi caracteristicile lor, care reprezintă interes din punctul de vedere al programului aplicativ. Imaginea poate avea o mulțime de proprietăți, precum descrierea formei, prezentarea vectorului subdiviziunilor, prezentarea vectorului ordinii de descompunere şi compunere a imaginii şi altele. Fiecare imagine "I" are o pereche asociată schimbătoare de numere pozitive (m,n), care se numeşte plasa imaginii. Ea este compusă din m*n celule de măsuri egale.

3.2.2. *Transformări de imagini*

Imaginea se împarte în părți omogene, care se numesc segmente. Schemele de compresare a imaginii sunt invertibile, deoarece unele scheme de compresare pot conduce la pierderea informației sau la pierderea perfecțiunii. Există două abordări a problemei căutării similarii imaginilor: abordarea metrică şi abordarea de transformare.

Abordarea de transformare este mai generală decât abordarea metrică. Această abordare utilizează operațiuni ca: transformarea, transferarea, rotația, scalarea, simetrizarea ş.a.

3.3. Utilizarea imaginii

în prezent multe instituții de învățămînt oferă programe de studii individuale. Unele persoane doresc să se specializeze în diferite domenii independent de o anumită formă instituționalizată de învățământ. Astfel de cursuri pot fi reprezentate sub formă de imagini speciale.

Imaginile pot fi utilizate în industria filmelor. Specialiştii au posibilitatea de a vizualiza imaginile alese de ei, lucrând la calculator.

Imaginile sunt importante şi în industria turismului. Pentru informații despre imaginile necesare se poate de asemenea apela la sistemul de tip IMAGINE a SGBDMM.

Interogările de imagini în dicționarul economic al MDLR informatizat pot fi efectuate la fel ca în subsistemele, de acelaşi nivel unu, de tip TEXT şi AUDIO prin intermediul limbajului SQL al SGBD. Rezultatul în forma textuală a articolului şi imaginea într-o formă complementară este prezentată utilizatorului în formă de Soft-copy sau Hard-copy.

Obținerea imaginii cuvântului «bancă», de exemplu, în subsistemul IMAGINI al SGBDMM al MDLR informatizat se efectuează prin intermediul următoarelor acțiuni. Se deschide baza de date IMAGINI a dicționarului economic al limbii române (în care sunt acumulate la data experimentării cu SGBDMM al MDLR informatizat doar 25 de cuvinte cu imaginile respective). Se alege cuvântul «bancă». în înregistrarea respectivă a băncii în compartimentul imagini se găseşte OLE al imaginii cuvântului ales. Se efectuează clic pe ea şi vizualizăm pe ecran imaginea respectivă. Analog se procedează şi cu alte cuvinte din BDI.

IV. VTDEO-dicționarul economic al limbii române[26]

în ultimii ani a crescut imens necesitatea de a putea chestiona şi procesa cantități mari de date, care nu sunt întotdeauna uşor de reprezentat prin intermediul simbolurilor. Exemple de astfel de date sunt: informația în formă

imagini, informația-video, datele-audio, informația textuală, notițe şi altele. în continuare vor fi examinate unele probleme de realizare a dicționarului economic informatizat cu VIDEO clipuri. A fost inițiată baza de date VIDEO (BDV) a dicționarului economic MULTIMEDIA- o subdiviziune a MDLR informatizat - prin crearea subsistemului VIDEO de nivel unu al SGBDMM. Se va demonstra viabilitatea acestui subsistem.

4.1. Problemele creării subsistemului VIDEO al SGBDMM.

Pentru a opera o bază de date MULTIMEDIA (BDMM), un SGBDMM trebuie să posede următoarele abilități:

(a) Capacitatea de a chestiona uniform datele reprezentate în diferite formate;

(b) Capacitatea de a chestiona uniform datele reprezentate în diferite surse media;

(c) Capacitatea de a aporta unitățile media dintr-o diviziune locală de depozitare, asigurând continuitatea acestui proces;

(d) SGBDMM trebuie să primească răspunsul, generat de o chestionare şi să poată genera o prezentare a acelui răspuns utilizând audiovizualul;

(e) Capacitatea de a oferi prezentarea într-un mod care ar satisface diferite cerințe ale utilizatorului.

Tehnologiile, legate de bazele de date, au dezvoltat în ultimii 40 de ani baza pe care ar trebui să fie creată o BDMM. în prezent sunt create limbaje de chestionare, tehnicile de aranjare, algoritmii de aportare pentru o mulțime de baze de date de tip relațional, spațial, temporal şi altele. Fiecare din aceste mijloace extind posibilitățile limbajelor şi algoritmii precedenți pentru a face față noilor tipuri de date sau pentru a dezvolta paradigmele respective.

în acest capitol se va analiza informația de tip VIDEO. Necesitatea de a accesa o bază de date VIDEO (BDV) poate apărea în numeroase aplicații, şi de obicei modelul de acces variază considerabil de la o aplicație la alta.

în procesul reprezentării conținutului unui film în BDV este necesar să se răspundă la un set de întrebări de tipul:

(A) Ce aspecte posibilele ale filmului pot interesa utilizatorii BDV?

(B) Cum pot fi aceste aspecte ale filmului eficient depozitate, astfel încât să minimalizeze timpul necesar subsistemului VIDEO al SGBDMM pentru a răspunde interogărilor utilizatorilor?

(C) Cum ar trebui să fie limbajul de interogare a datelor VIDEO şi cum ar trebui schimbat modelul relațional pentru a corespunde informației VIDEO?

(D) Poate fi oare automatizat procesul de extragere a informației în baz contextului?

Aceste probleme au fost abordate în procesul creării şi experimentării c BDV şi subsistemul VIDEO de nivel unu al SGBDMM.

4.2. Definițiile datelor de tip VIDEO

De obicei un film este caracterizat prin personajele sale, atribute acestora şi activitățile în care sunt angajate aceste personaje. Principalele surs de interes într-un film includ: (a) oameni, (b) obiecte neînsuflețite, (c) fiin însuflețite şi (d) activități.

De observat că tema generală, care se repetă în fiecare cadru, constă existența un grup de obiecte şi activități asociate. Astfel vom încerca să definim bază de date VIDEO printr-un şir de definiții.

Definiție 1: O *proprietate* VIDEO este o pereche *{pname, Values),* unde *pnar* este numele proprietății şi *Values* este o mulțime. O *instanță* proprietății *(pname, Values),* este o expresie de forma *pname-v,* un vc *Values.*

Definiție 2: O *schemă obiect este* o pereche *(fd, fi),* unde:

fd este o mulțime de proprietăți cadru-dependente,

fi este o mulțime de proprietăți cadru-independente *(fi* şi *fd* sînt mulț disjunctive).

Definiție 3: O *instanță obiect* este un triplet *(oid, os, ip),* unde:

o/of este o frază numită identitatea obiectului,

os= *(fd, fi)* este o schemă obiect şi

ip este o mulțime de afirmații de tip:

(a) pentru fiecare proprietate *(pname, Values),* în *fi, ip* conține cel r o instanță a proprietății *(pname, Values),*

(b) pentru fiecare proprietate *(pname, Values)* în *fd* şi pentru fiec cadru *f* al filmului, *ip* conține cel mult o proprietate insta *(pname, Values).* Această proprietate instanță este notată *pname* = v IN f.

Definiție 4: O schemă activitate *ACT_SCH* este o mulțime finită de proprie astfel încât, dacă *(pname, Valuesl)* şi *(pname, Values2)* apa *ACT_SCH,* atunci *Values^ Valuesl.*

Definiție 5: O activitate este o pereche, care constă din

(a) AcID, indicele schemei activitate *ACT_SCH* şi

(b) pentru fiecare pereche *(pname, Values)* c *ACT_SCH* este valabilă ecuația de forma *pname*= v, unde v c *Values.*

Oricărei activități i se asociază o schemă de activitate şi fiecărei proprietăți i se asociază o valoare din mulțimea valorilor posibile.

Fiind dată o singură dată VIDEO v, putem defini "conținutul" filmului v.

Definiție 6: Fie că *framenum(v)* specifică numărul total de cadre din filmul v. *Conținutul* lui *v* constă dintr-un triplet *(OBJ,AC,Ă),* unde:

1. 08J={oid$_1$,...,oid$_n$} este o mulțime finită de instanțe ale obiectului,

2. i4C={AclDi,...,AclD$_k$} este o mulțime finită de activități/evenimente şi

3. A este o hartă de la {1,...,framename(v)} până la 2^{OBJUAC}-

Intuitiv, conținutul unei date VIDEO v este teoretic descris de tripletul *(OBJ,AC,Ă),* unde:

1. *OBJ* reprezintă mulțimea obiectelor de interes în film,

2. *AC* reprezintă mulțimea activităților de interes din film şi

3. A reprezintă obiectele şi activitățile, care sunt asociate cu fiecare cadru *f* al filmului.

4.3. VIDEO biblioteca

O persoană interesată de obținerea unei lecții imprimate pe o casetă video ar dori să chestioneze o VIDEO bibliotecă, care găzduieşte o colecție de casete video, referitoare la un anumit subiect. De exemplu, Universitatea Maryland oferă cursuri, utilizând contactul prin satelit. în viitor casetele video, create în acest fel, vor putea fi accesate cu ajutorul unui calculator, oferind astfel studenților prelegeri pentru mai multe obiecte de studiu adunate de-a lungul anilor şi ținute de diferiți lectori. Chestionarea bazei de date VIDEO de un student individual ar presupune accesarea unui număr foarte mare de casete video.

O bibliotecă VIDEO este o colecție, care specifică: (a) totalitatea filmelor din bibliotecă, (b) conținutul fiecărui film şi (c) memorizarea fizică a filmelor.

Definiție 7: O *VIDEO bibliotecă VidLib* constă dintr-o mulțime finită de cuvinte de tip *(VidContent, Vidjd, framenum, R, plm),* unde:

(a) *VidContent* este conținutul filmului,

(b) *Vidjd,* este numele filmului,

(c) *Framenum* este numărul de cadre în film,

(d) *Plm* este amplasarea, care specifică adresele diferitor părți ale filmului şi

(e) R este mulțimea relațiilor despre filme în întregime.

4.3.1. Chestionarea bibliotecii VIDEO

Chestionarea unei VIDEO biblioteci conține următoarele tipuri de interogări: (a) *aportarea segmentelor* (Găseşte toate segmentele care corespund unei anumite cerințe), (b) aportarea obiectelor, (c) aportarea activităților şi (d) aportarea proprietăților de bază (Care VIDEO-date sunt în bibliotecă, care este conținutul fiecărei VIDEO-date selectate, unde sunt localizate fizic VIDEO-datele).

4.3.2. Funcțiile VIDEO-datei

Cu bibliotecile VIDEO pot fi definite o serie de funcții:

FindVideoWithObject(o): fiind dat numele obiectului o, această funcție ne oferă tripletul (Videold, StartFrame, EndFrame),

FindVideoWithActivity (a)

FindVideoWithActivityandProp(a,p,z)

FindVideoWithObjectandProp(o,p,z)

FindObjectsInVideo(\/,s,e)

FindActivitiesInVideo *(v,s,e)*

FindActivitiesAndPropsinVideo *(v,s,e)*

FindObjectAndPropsInVideo *(v,s,e)*

O chestionare standard a VIDEO-bibliotecii, utilizînd SQL are forma:

SELECT câmp!,..., câmp$_n$
FROM relația^RV), relația$_2$(f?2),..., relația$_k$ *(Rk)*
WHERE condiție.

4.3.3. Ordonarea datelor VIDEO

O problemă importantă este crearea structurilor informaționale, care organiza bazele de date VIDEO în aşa fel încât să optimizeze procesarea celor o . funcții enumerate mai sus. Este imposibil de a se depozita conținuturi al VIDE datelor cadru cu cadru, deoarece un singur film de 90 minute conține 162,0 cadre. Astfel, este necesar să se creeze reprezentări compacte ale conceptului conținut video. în acest sens vom prezenta două astfel de structuri: (a) arbo segment cadru, şi (b) arborii R-segment.

4.3.4. *Arborii segment cadru*

Ideea de bază a arborelui segment cadru este foarte simplă. La început creează două tabele unidimensionale: OBJECTARRAY şi ACTIVITYARRAY.

în acest context arborele poate fi creat în 2 etape:

La prima etapă presupunem că [s^e!),..., [s_w,e_w) sunt toate intervalele în coloana "Segment" a tabelei segment. Fie qi,...,q_z o enumerație ascendentă a tuturor membrelor {Sj.es 11< i < w}. Dacă z nu este o putere a numărului 2, atunci se procedează astfel: fie r cel mai mic număr întreg astfel încât 2'>z şi *2'>framenum(v).* Se adaugă noi elemente $q_{z+i},...,q_{2^r}$ astfel încât $q_z r$ = *framenum(v)+î* şi q_{z+j}= q_z+j (j>0, z+j<2').

La a doua etapă arborele este unul binar format după cum urmează:

1. în fiecare nod arborele segment cadru reprezintă o secvență de cadru [x,y).
2. Fiecare frunză este la nivelul r. Prima frunză din stânga marchează intervalul [z-i z_2), a doua [z_2,z_3) şi aşa mai departe.
3. Numărul din interiorul fiecărui nod este adresa acelui nod.
4. Mulțimea de numere de lângă nod marchează numărul de identitate al VIDEO-obiectelor şi al VIDEO-activităților, care apar în întreaga secvență de cadru asociată cu nodul dat.

Definiție 8: O *secvență de cadru* este o pereche [i,j), unde 1< i < n şi [ij) reprezintă mulțimea tuturor cadrelor între i (inclusiv) şi j.

Definiție 9: **O** *ordonare* parțială c asupra mulțimii tuturor secvențelor de cadru este definită ca [i-iji) c [$i_2$$j_2$) cu condiția, că ii<ji=$i_2$<$j_2$.

Definiție 10: O mulțime X de secvențe de cadru este *bine aranjată* dacă:
1. X este finită (adică X= {[iiji),..., [ir_2,jr_2)}, pentru oricare r_2)şi
2. [iiji)c[i_2,j_2)c . . . c [ir_2,jr_2)

Definiție 11:0 mulțime X de secvențe de cadru este *solidă* dacă:

1. X este bine ordonată şi
2. Nu există nici o pereche de secvențe de cadru în X de forma [h,i_2) şi P2J3)

4.3.5. *Operații cu arborii segment cadru.*

Fiecare film v este o structură de VIDEO-date, care constă dintr-un arbore segment cadru, un tablou obiect şi un tablou activitate. în particular, dacă biblioteca *VidLib* conține filmele v^..., $v_{n>}$ atunci este suficient să asociem următoarele:

1. O singură tabelă numită INTOBJECTARRAY cu schema (VID.ID, OBJ, PTR),
2. O tabelă numită INACTIVITYARRAY cu schema (VID.ID, ACT, PTR) şi
3. Pentru fiecare arbore segment cadru Vj, *fst(v,) este* asociat cu filmul Vj.

De asemenea pot fi exprimate cele 8 funcții, în SQL, introduse mai sus. De exemplu, una din aceste funcții FindVideoWithObject(o), poate fi implementată cu arborii segment cadru printr-o operație de selecție, efectuată asupra INTOBJECTARRAY DE TIP:

```
SELECT   VIDEOJD
FROM     INTOBJECTARRAY
WHERE OBJ = o.
```

4.3.6. Arborii R-segment (RS-arbori)

Arborii R-segment sunt foarte asemănători cu arborii segment cadru, cu o singură deosebire. Deşi conceptele de OBJECTARRAY şi ACTIVITYARRAY rămân aceleaşi, în locul utilizării unui arbore segment cadru, pentru a reprezenta secvența de cadru, profităm de faptul că o secvență [s,e) este un dreptunghi cu lungimea laturii (e-s) şi lățimea 0. Fiecare nod va avea o structură specială pentru a specifica, pentru fiecare dreptunghi, care obiect sau activitate este asociată acestuia.

4.4. Operații cu VIDEO-clipuri

Un film este creat prin filmarea unor secvențe şi combinarea lor, utilizând un operator de combinare. O secvență este de obicei filmată de mai multe camere, fiecare având o viteză relativă de rotație constantă. în general o secvență poate avea mai multe atribute asociate precum durata filmării, tipul de cameră utilizat şi altele.

Un operator de combinare a filmărilor, deseori numit *edit effect,* este o operație care în baza a două filmări Si şi S_2, şi a unui interval de timp t efectuează o secvență compusă în timpul t. Aşadar un film este creat prin combinarea unei mulțimi de secvențe filmate, utilizând un şir finit de operații de compunere. Exemple de astfel de operații de compunere a filmelor includ:

1. Concatenarea filmărilor,
2. Compoziția spațială şi
3. Compoziția cromatică.

4.5. Standardele video

Deşi în general standardele industriale nu sunt parte componentă a nucleului cadrului MULTIMEDIA, este important să explicăm în linii generale ideea de bază a standardelor MPEG.

Toate standardele de comprimare a informației VIDEO încearcă să comprime filmele prin executarea unei analize intra-cadru: fiecare cadru este

divizat în blocuri, diferite cadre sunt comparate, pentru a vedea dacă informaţia conţinută de acestea nu se repetă în două cadre. Calitatea tehnicii de compresie este măsurată conform următorilor trei parametri de bază:

(a) Fidelitatea hărţii color: cât de multe culori ale filmului original sunt prezente după comprimare?

(b) Rezoluţia pixel pe cadru: câţi pixeli au fost abandonaţi?

(c) Numărul de cadre pe secundă: câte cadre au fost abandonate?

4.6. Scenarii de utilizare a VIDEO-dictionarului

Dicţionarul MULTIMEDIA al limbii române cuprinde peste 70000 de cuvinte din cele mai diverse domenii. Dicţionarul este conceput atât pentru studenţi, cât şi pentru cercul larg al vorbitorilor limbii române, care doresc să cunoască sensul propriu care trebuie conferit cuvintelor. Dicţionarul MULTIMEDIA satisface cerinţele de bază: dă definiţia exactă a cuvântului şi, dacă e cazul, genul, numărul, sinonimele, antonimele, imagini, secvenţe VIDEO şi AUDIO, care exprimă sensul exact şi limpede, deplin accesibil, ceea ce constituie partea cea mai importantă de utilizare. Acest dicţionar este una din pietrele de temelie ale culturii tineretului, care va contribui la opera de culturalizare a maselor prin iniţierea în folosirea limbii române în mod corect, exact şi unitar.

Compartimentul VIDEO al acestui dicţionar MULTIMEDIA al limbii române conţine, după pronosticurile noastre, peste 12000 cuvinte. Acest compartiment furnizează informaţii necesare referitoare la cuvintele căutate, secvenţe video ce oferă posibilitatea de a percepe mai bine esenţa cuvintelor. Diviziunea video face dicţionarul mult mai accesibil şi atractiv pentru utilizatori de toate vârstele şi preocupările.

Necesitatea utilizării VIDEO-dicţionarului poate apărea în cele mai diverse situaţii. Să considerăm situaţia în care un student este nevoit să scrie un referat la merceologia şi tehnologia produselor alimentare. Studentul trebuie să analizeze procesul tehnologic de producere a pâinii. în acest sens, apelarea la VIDEO-dicţionarul limbii române îi va uşura lucrul; acesta îi va furniza secvenţe VIDEO, ce prezintă procesul de fabricare a pâinii, ingredientele utilizate, utilajul necesar.

4.6.2. *Chestionarea Video dicţionarului*

Dicţionarul VIDEO este organizat ca o mini-bibliotecă VIDEO. După cum am subliniat mai sus, în procesul de chestionare cele mai importante aspecte sunt:

(a) Aportarea segmentelor: utilizatorul poate cere bazei de date VIDEO să-i ofere toate secvenţele, care conţin informaţii despre procesul tehnologic de producere a pâinii. O astfel de chestionare ar fi: "Găseşte toate secvenţele unde se combină ingredientele", sau "Găseşte toate secvenţele unde se frământă pâinea".

(b) Aportarea obiectelor: în acest caz, utilizatorul poate solicita toate segmentele, în care este prezent cuptorul, banda rulantă etc. Formularea întrebării ar fi: "Găseşte toate secvenţele, în care apare cuptorul", "Găseşte toate secvenţele, în care apare banda rulantă" etc.

(c) Aportarea activităţilor: se solicită prezentarea tuturor segmentelor, în care pot fi urmărite diferite operaţiuni de producere. întrebarea poate fi "Găseşte toate secvenţele, în care se desfăşoară operaţiunile de producere".

4.6.2. *Utilizarea bazelor de date VIDEO în diferite domenii*

După cum am menţionat anterior, scopul baze de date VIDEO este de a satisface cele mai diverse cerinţe. Astfel, aceste BDV îşi găsesc aplicarea în cele mai diverse domenii.

4.6.2.1. Educaţie. Bazele de datele VIDEO au o aplicare largă în educaţie şi cercetare. Universităţile pot acorda servicii precum studii la distanţă prin satelit sau utilizând Internetul. Acestea pot pune la dispoziţia studenţilor un set de casete VIDEO cu înregistrări ale cursurilor. Dicţionarul VIDEO, fiind şi el o bază de date VIDEO, pune la dispoziţia utilizatorilor secvenţe VIDEO care pot fi utilizate în cadrul comunicărilor, pentru pregătirea unor prezentări, lecţii deschise, rapoarte.

4.6.2.2. Sport. Sălile de Sănătate oferă baze de date, în care sunt înregistrate casete VIDEO ce conţin diferite programe de antrenament, utilizatorului oferindu-i-se posibilitatea de a alege între programe de slăbire, fortificare sau menţinere a condiţiei fizice.

4.6.2.3. Agricultură. Institutele de cercetări ştiinţifice în domeniul agriculturii din ţară ar putea utiliza VIDEO-dicţionarul pentru a studia mai aprofundat procesul de plantare, condiţiile de creştere şi dezvoltare a plantelor, specificul dezvoltării plantelor în diferite regiuni sau ţări, aclimatizarea plantelor în condiţiile ţării în cauză.

4.6.2.4. Economie. VIDEO-dicţionarul poate fi utilizat în foarte multe domenii ale economiei: finanţe, contabilitate, management, marketing, statistică, turism. Vocabularul economic cuprinde destul de mulţi termeni, care pot fi redaţi printr-un limbaj VIDEO mai accesibil, atât specialiştilor, cât şi utilizatorilor obişnuiţi.

V. Concluzii

5.1. Compartimentul TEXT. Dicţionarul economic TEXT al limbii române în forma sa de BDT, ca o subdiviziune a MDLR, are posibilitatea de a fi extins cu caracteristicile respective ale MULTIMEDIA: Imagine, Audio, Video etc. Aceasta BDT va ocupa aproximativ 18 MB memorie. La conferinţa tinerilor savanţi ai ASEM din 4-5 aprilie 2002, pe baza câtorva sute de articole din DEI au fost demonstra

caracteristicele de utilizare prietenoasă a subsistemului TEXT al SGBDMM, utilizând sistemele Ms ACCESS - 2000, Ms WORD - 2000 şi Ms PowerPoint - 2000 în calitate de componente ale Software-ului Ms OFFICE -2000 şi WINDOWS - 2000, exploatate în baza hardware-ului de tip PC Pentiun II, conectat la rețelele Intranet, Externet şi Internet.

5.2. Subsistemul AUDIO. Subsistemul AUDIO interacționează cu celelalte subsisteme de nivel unu (TEXT, IMAGINI, VIDEO) ale SGBDMM, care susține evaluarea Marelui Dicționar al Limbii Române informatizat cu MULTIMEDIA. Acest subsistem AUDIO susține toate definițiile celor 61635 de articole din DEI de comun acord cu subsistemul TEXT al SGBDM. Cele 2320 de ilustrații din DEI sunt susținute de componenta IMAGINI a SGBDMM, dar cu ele poate fi extinsă componenta TEXT şi/sau componenta AUDIO. Exemplele, enumerate mai sus, de utilizare a AUDIO componentei a MDLR informatizat, au un aspect comun, abstract vorbind formează corpul unei date, fiind individual executate în diferite probleme prin intermediul diferitor suporturi ale Software-ului şi Hardware-ului modern. Baza de date BDA a compartimentului AUDIO-dicționarului economic al MDLR informatizat va ocupa un volum de memorie de circa 60 GB memorie.

5.3. Subsistemul IMAGINI. BDI al subsistemului IMAGINI al MDLR informatizat recent a fost expusă pentru analizare şi discuții la Conferința tinerelor cercetători ai ASEM din 4-5 aprilie 2002 în baza câtorva zeci de articole din DEI. Mijloacele Software-ului şi Hardware-ului de tip Ms ACCESS-2000, Ms WORD-2000 şi Ms PowerPoint-2000 cu dispozitivele respective al PC-ului Pentium II au fost suficiente la etapa inițială pentru a demonstra eficiența şi eficacitatea mijloacelor şi metodelor alese pentru realizarea Proiectului "Limba Română - Limba a Comunității Europene" de către grupul de cercetători - autori ai acestei lucrări. Volumul BDI de prezentare în Ms ACCESS-2000 fără comprimare a 50 articole din DEI ocupă circa 550 MB memorie.

5.4. Subsistemul VIDEO. După cele menționate mai sus ținem să subliniem, că subdicționarul VIDEO are o utilitate mare pentru persoanele ce operează în diferite domenii ca: economia, educația, sport, agricultură, industrie etc. Avantajul acestui dicționar este uşurința în folosire şi accesibilitatea. Dicționarul VIDEO este o bază de date, cu care putem opera oricând avem nevoie şi oferă posibilitatea de a percepe o informație în formă de videoclipuri. în acest mod, persoanele ce se folosesc de acest dicționar înțeleg mai uşor sensul cuvântului, care este reprezentat în formă VIDEO, fiindcă se formează o imagine amplă despre cuvântul dat şi este uşor de memorizat.

5.5. Lucrări paralele şi perspective. Paralel cu sistemele de nivel unu, • sunt elaborate sistemele de nivelul doi, care suportă subdiviziunile MDLR în planurile: TEXT&AUDIO, TEXT&IMAGINI şi TEXT&VIDEO.

Elaborarea sistemului, care suportă toate compartimente MULTIMED MDLR informatizat, constituie a treia platformă, mai complexă, de experime implementări ale dicționarelor computerizate în cadrul elaborării MDLR infor **[17].**

Rezultatele evaluării preliminare ale primelor elemente ale aces platforme: sistemele unare TEXT, AUDIO, IMAGINI şi VIDEO au creat posib de a transforma unele concluzii ale evaluării MDLR informatizat ca c componentă a cercetărilor în cadrul Proiectului «Limba română - li Comunității Europene», care se desfăşoară în perioada 2000-2006. Acest a fost inițiat [10-11] de către Forumul Internațional din Chişinău,14-15 apri 2000. Proiectul constituie unul dintre subiectele de cercetare, experimer evaluare, efectuate în cadrul Consorțiului Uniunii Latine «Pentru limba roma Consorțiului «Pentru informatizarea limbii române» şi a Comisiei Aca Române «Pentru informatizarea limbii române».

O serie de aplicații a MDLR computerizat este evidențiată în [13-16]

Referințe bibliografice

[1] V. S. Subrahmanian. Principles of Multimedia Database Systems. // Kaufman Publishers, Inc., San-Francisco, California, USA, 1998, -pp. 44

[2] D. Todoroi, S. Nazem, T. Jucan, D. Micusha. Transition To A Full Info Society: Stage Development. // Working Paper No. 98-2, UNO, Omaha March 1998.-38 p.

[3] D. Todoroi, D. Micuşa, V. Clocotici, I. Lingă, V.Tapcov, N. Drucioc, A. C M. Morari. **Data Bases** and Communications Tools. Ms ACCESS - 20(ASEM, Chişinău 2002, 337 pages. (Eng.)

[4] Dumitru N. Todoroi, Zinaida Todoroi, Diana Micusa. Romanian Compu Language - One of the European Community Languages. // Proceedin 26th Annual Congress of the American Romanian Academy of A Sciences (ARA), Montreal, Quebec, Canada, July 25-29, 2001, pp. 1 (Rom)

[5] Diana D. Micusha, Dumitru Todoroi. Natural language processing transition to a full information society inițial development phase. Part 1. şi cercetări economice. Voi. XXX. Lucrări prezentate la Sesiunea jub comunicări ştiințifice: «Creştere economică, dezvoltare, progres» Napoca, 2001, pp. 1396-1413.'

[6] Diana D. Micusha, Dumitru Todoroi. Natural language processing transition to a full information society inițial development phase. Part 2. şi cercetări economice. Voi. XXX. Lucrări prezentate la Sesiunea jub comunicări ştiințifice: «Creştere economică, dezvoltare, progres» Napoca, 2001, pp. 1414-1427.'

[7] Sabin-Corneliu Buraga, Dumitru Todoroi. Adaptabilitatea informațională şi operațională. // Studii şi cercetări economice. Voi. XXX. Lucrări prezentate la Sesiunea jubiliară de comunicări ştiințifice : «Creştere economică, dezvoltare, progres», Cluj-Napoca, 2001, pp. 1447-1457.

[8] Dumitru TODOROI. 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.

[9] Dumitru N. TODOROI. IEE-2000 PROJECT: Natural Language Processing Initialization. // EUROPEAN EXCELENCE IN BUSINESS STUDIES STUDENTS' EDUCATION. Internetional Symposium. Edited by IOAN ANDONE, Bucureşti, Editure Economica, 2000, pp. 328-334.

[10] Dumitru Todoroi. Project: Romanian Language - One of the European Community Languages. // Proc. of the VI Conf. « Application Sciences», 18-19 May 2000, USAM, Chişinău, pp. 12-15.

[11] Dan Crisrea, Dumitru Todoroi, Dan Tufiş. Computațional Linguistic: Romanian Language - One of the European Community Languages. // Proc. of the Intern. Sc. Seminar "Strategies and Modalities for România and Moldova' European Integration", 28-29 Sept. 2000, V.2, ASEM, Chişinău, pp.276-280.

[12] D. Todoroi, D. Micusa, V. Clocotici, S. Pereteatcu, V. Bordeianu, C. Grigoras, S. Cretu, I. Lingă, S. Spataru. Natural Language Processing: IEE-2000 Project. // Proc. of the Intern. Sc. Seminar "Strategies and Modalities for România and Moldova' European Integration", 28-29 Sept. 2000, V.2, ASEM, Chişinău, pp.281-285.

[13] Ştefan Spataru, Dumitru Todoroi. Distance Education Via Internet, Multimedia and modern System Environment. // Proc. of the Intern. Sc. Seminar "Strategies and Modalities for România and Moldova' European Integration", 28-29 Sept. 2000, V.2, ASEM, Chişinău, pp. 307-312.

[14] Ion LINGĂ. IMPACTUL IMPLEMENTĂRII COMPUTERULUI ASUPRA PROCESULUI DE ASIMILARE A CUNOŞTINȚELOR. // Proc. Of the 27th ARA Congress, May 29 - Lune 2, 2002, Oradea, România.(To be published).

[15] Ion COVALENCO. Metode adaptabile de evaluare a cuniştințelor asistată de calculator. // Proc. Of the 27th ARA Congress, May 29 Lune 2, 2002, Oradea, România. (To be published).

[16] Nicolae OBJELEAN.The Metod for Error Corection in String with Applications in Speach Recognition. // Proc. Of the 27th ARA Congress, May 29 - June 2, 2002, Oradea, Romanja.(To be published).

[17] Dumitru N. TODOROI, ASEM, Chişinău, Nicolae MARGINEANU, L'Ecole Politechnique, Montreal, Canada.THEROMANIANLANGUAGE'MULTIMEDIA - DICTIONARIES IMPLEMENTATION ENVIRONMENT AT THE FULL INFORMATION SOCIETY INIȚIAL DEVELOPMENT PERIOD. // Proc. Of the 27th ARA Congress, May 29 - Lune 2, 2002, Oradea, Romania.(To be published).

[18] Diana MICUSHA. Mijloace adaptabile ale sistemelor de procesare a limbaj natural computerizat. // Proc. Of the 27th ARA Congress, May 29 - Lune 2002, Oradea, România.(To be published).

[19] Zinaida TODOROI, ULIM, Chişinău, Eugenia MARGINEANU, L'Ec Politechnique, Montreal, Canada. MULTIMEDIA - dictionaries for Roma Language. Usage Scenarios on the EAPEC Base. // Proc. Of the 27th A Congress, May 29 - Lune 2, 2002, Oradea, România.(To be published).

[20] Societatea informațională - Societatea cunoaşterii. Concepte, soluț strategii pentru România. // ACADEMIA ROMÂNĂ, Editura EXPE Bucureşti, decembrie 2001. - 541 pages.

[21] Dicționar Enciclopedic Ilustrat (DEI). // Editura CARTIER SRL, Chiş Editura CODEX SRL, Bucureşti, 1999, 1808 pages.

[22] Beyond Calculation : The Next Fifty Years of Computing. // Edited by Pe Denning and Bob Metcalfe, Copernicus, 1997 Springer-Verlad New York, 350 pages.

Comunicări la Conferința tinerilor cercetători ASEM, 4-5 **aprilie** 2002 **Chişinău.**

Coordonator: Dumitru TODOROI, Prof. Univ., doctor habilitatus.

[23]. AMBROZII Nadejda, GODZINA Irina. Componenta Text a Audio Dicțio Economic al Limbii Române.

[24]. TURCANU Virginia, MUTRUC Carolina. AUDIO-DICTIONARUL EXPLI ECONOMIC AL LIMBII ROMÂNE

[25]. COZLOV Elena, BABANU Irina. Subsistemul IMAGE al dicțio economic informatizat al limbii române.

[26]. LUNGU Stela, CIOBANU Diana, GUZUN Oxana. VIDEO-dicționarul ec al limbii române.

Mediu pentru editarea transcrierilor fonetice în limba română. Realizarea atlasului lingvistic român pe regiuni

Silviu BEJINARIU, Vasile APOPEI, Mariana ROMAN
Academia Română, Institutul de Informatică Teoretică, Iaşi, B-dul Carol nr. 8
silviub@academie.is.edu.ro, vapopei@academie.is.edu.ro

Abstract

The goal of our work is to create an Electronic Linguistic Atlas of România. The Electronic Linguistic Atlas has feâtures of a multimedia application allowing the user to consult and/or prinț the linguistic maps and to listen audio recordings or synthesized speech.

In order to show all the spelling variations, the phonetically transcription is used in the linguistic atlases. For the Romanian Language, the graphic symbols have been hand-written.

The editing process is too difficult using a standard text editor as consequence of the great number of fonts used. In this paper we propose an editing interface for the phonetic transcription of the Romanian Language. This interface can be used to edit dictionaries of the Linguistic Atlas and as editing tool for the phonetic transcriptions in stand-alone mode or as server for other text editors.

Keywords: dictionary, phonetically transcription, multimedia, linguistic atlas

1. Clasificarea simbolurilor grafice pentru editarea transcrierilor fonetice

Pentru a putea arăta toate nuanțele de rostire, în lingvistică se recurge (după practica internațională) la transcrierea fonetică. Pe lângă transcrierea fonetică internațională realizată cu Alfabetul Fonetic Internațional (IPA), fiecare țară îşi are propriile simboluri grafice [1], [2]. La realizarea atlaselor lingvistice româneşti, aceste simboluri sunt scrise doar manual. în lucrarea [3] este

prezentată o primă abordare a realizării simbolurilor grafice pentru transcrierea fonetică din perspectiva realizării variantei computerizate a atlaselor lingvistice româneşti.

în această primă parte vom prezenta principiile care au stat la baza modului în care au fost organizate simbolurile grafice folosite în transcrierea fonetică a limbii române.

Pentru claritatea prezentării introducem următoarele noțiuni:

- sunete primare[1]:
- vocale, consoane - existente în alfabetul latin care au corespondent pe tastatură;
- diacritice - vocale, consoane - care nu au corespondent pe tastatură dar pot fi obținute prin combinații de taste;
- sunete marcate cu unul sau mai multe fenomene fonetice.

De aici a rezultat necesitatea realizării unui font de bază *(ALRJBaza)* care să cuprindă simbolurile grafice pentru toate sunetele primare. Poziția în "font" a simbolurilor grafice pentru diacritice, a fost stabilită urmărind păstrarea poziției implicite din familiile de fonturi uzuale *(Arial, Times New Roman).* Pentru realizarea sunetelor marcate cu unul sau mai multe fenomene fonetice am proiectat familii de fonturi ale căror denumiri le-am format folosind denumirea fenomenelor fonetice aplicate (ex. *ALRJSemivocale, ALRJNazalizate, ALRJSeminazalizate, ALRJScurteNazalizate, ALR_* .... etc). Această organizare a fonturilor a fost făcută cu scopul de a permite scrierea textelor cu transcrieri fonetice cu orice editor de text *(Microsoft Word),* iar textul scris cu aceste fonturi să poată fi citit chiar dacă fonturile proiectate de noi nu sunt instalate (în acest context se vor pierde numai fenomenele fonetice aplicate sunetelor primare).

Pentru generarea acestor fonturi am folosit programul *FontLab 3.1* care permite definirea de simboluri grafice compuse, pornind de la o familie de fonturi *TrueType* existentă în sistemul de operare Windows. Pentru familiile de fonturi pe care le-am realizat am convenit să folosim ca model de plecare fontul *ARIAL.*

Facem precizarea că fenomenele fonetice şi modul lor de aplicare este diferit pentru cele două tipuri de sunete: vocale şi consoane.

Formularea "sunete primare", inexactă din punct de vedere fonetic, este folosită cu înțelesul "sunete a căror imagine grafică pe calculator are corespondent pe tastatură, sau este obținută prin combinații de taste ".

1.1. Fenomene fonetice aplicate vocalelor primare

simple			diacritice			
a	â	ă	â	a'	â	
e	e	e				—
i			î			î
o	6	6				
u	o		u			

Cu ajutorul acestor "vocale primare" şi al celor trei variante accentu (a - â â â) ₐ | ₑ fiecăreia dintre ele se obține seria completă de sunete vocal care se regăsesc în fontul de bază *ALRJBaza* (17*4=68 grafeme).

Fenomenele fonetice care pot modifica cele 17 vocale de bază împreună cu variantele lor accentuate), sunt clasificate în următoarele grupe fenomene disjuncte[3]:

Grupe	Poziționare	Notație	Fenomen	Exemp
[Durată	Aşezat cel mai sus	.(a)	[Scurtime	
		(b)	Semilungime	
		(c).	|Lungime	e e e
		(d)	iSeminazalizare	e e l
Nazalizare	Aşezat deasupra vocalei, dar stih fenomenele *(a)-(c)*	(e)	[Nazalizare	e e e
			Coup de glotte	?e 'e,
Ocluzie glotală	aşezat "în umăr", în fața		închidere	e â e
Deschidere	Aşezat imediat sub vocală	(9)	Semideschidere	
		(h)	Deschidere	e ? ?
		(!)	Deschidere mare	f f
		(I)	Semiafonizare	e e e
Afonizare	Aşezat sub vocală dar şi sub fenomenele (g)-Q)	(k)	[Afonizare	e e e
		(l)		

Din punct de vedere lingvistic sunt impuse următoarele reguh.

Regula [1]

- vocalele **a** ă â - deschise prin natura lor (cu cel mai mare grad de apertură) - nu pot contacta fenomenele fonetice *h* (semideschidere), / (deschidere) şi *j* (deschidere mare);
- vocalele i T î u u Q - închise prin natura lor (cu cel mai mic grad de apertură) - nu pot contacta fenomenul fonetic *g* (închidere).

Prin asocierea vocalelor primare cu câte un fenomen *(a)-(l)* rezultă 756 imagini grafice repartizate în 12 fonturi grupate convenţional după criteriul poziţiei semnului faţă **de** vocală.

Regula [2]

Sunt excluse orice combinaţii dintre două nuanţe fonetice din aceiaşi grupă de transformări vocalice. Astfel, o vocală nu poate fi în acelaşi timp „scurtă, semilungă şi lungă" sau „seminazală şi nazală" sau „închisă, semideschisă, deschisă şi foarte deschisă" sau „semiafonizată şi afonizată". în aceste condiţii combinaţiile de câte două sau mai multe fenomene sunt posibile doar cu fenomene din grupe diferite.

în plus, cele *15*4=60 grafeme* excluse ca urmare a restricţiei formulate sub **Regula** 1, nu pot participa la combinaţiile de două, trei, patru fenomene.

1.2. Fenomene fonetice aplicate consoanelor primare

Consoanele primare folosite în transcrierea fonetică sunt:

b, c,c,€A6,c, d.cţdA f, g,Mg,ă, h,h,x>Y> j, K tt
m,m, n,n,n,n, p, r,r,f,p, s,s,s,ş,a, U,ţ, v, w, z,z,z, y

Fenomenele fonetice care pot fi asociate consoanelor primare sunt:

Grupe	Notaţie	Fenomen
Durată	d)'	Semilungime
	(2)	Lungime
Palatalizare	(3)	Semipalatalizare
	(4)	Palatalizare
	(5)	Palatalizare mare
Explozie	(6)	Explozie
Caracter silabic	(7)	Caracter silabic
Afonizare	(8)	Semiafonizare
	(9)	Afonizare

Spre deosebire de vocale, unde s-au putut defini reguli generale pentru realizarea combinaţiilor de fenomene fonetice, în cazul consoanelor primare, transformările fonetice se aplică numai unor consoane specifice. în consoanelor primare le pot fi aplicate numai cel mult două transformări şi n anumite combinaţii. în tabelul următor sunt prezentate combinaţiile posi fenomene şi consoanele pe care acestea le pot însoţi.

1.2.1. Consoane cu un singur fenomen fonetic:

semilungime	**6î** fi5xyjlT*/pnoi}i)f^&sSlşsv»22
lungime	5? hf)xyJîtmmnnnf3fFfpsiâş^vwzz2
semipalatalizare	dVfixKj˅l˅fifşt˅
palatalizare	
palatalizare mare	t"cf'
explozie	c° p° t°
caracter silabic	jmnrşm
semiafonizare	
afonizare	

1.2.2. Consoane cu două fenomene fonetice:

semilungime + semipalatalizare	
semilungime + palatalizare	Rjfînfş
semilungime + caracter silabic	
semilungime + semiafonizare	* r r % * + *• r ·"î:£i-·-·-·-S»-»- i- i- »- v
semilungime + afonizare	
lungime + semipalatalizare	
lungime + palatalizare	njTrtf$
lungime + caracter silabic	jmmnfş
lungime + semiafonizare	5i}yjrtmffinnnnfFrpvWz2fy
lungime + afonizare	
semipalatalizare + semiafonizare	
semipalatalizare + afonizare	

palatalizare + semiafonizare palatalizare + afonizare	
palatalizare mare + semiafonizare	d'
palatalizare mare + afonizare	d'
explozie + semiafonizare explozie + afonizare	b°tfg°
caracter silabic + semiafonizare	1 m m n r i- *l i l v*
caracter silabic + afonizare	1 m m n r i i *t* i i

2. Mediu pentru editarea transcrierilor fonetice

Interfața realizată pentru editarea transcrierilor fonetice poate fi folosită în mai multe moduri:

- editarea dicționarelor Atlasului Lingvistic;
- editor stand-alone sau ca aplicație de tip server pentru inserarea de obiecte de tip "transcriere fonetică" în alte editoare de text.

Funcționalitatea acestei interfețe va fi exemplificată pentru situația Atlasului Lingvistic, ale cărui componente sunt prezentate pe scurt în continuare.

Dicționarele ALR sunt componente care realizează colectarea informațiilor primare despre titlul hărților (cuvinte de bază), punctele de anchetă, speech (colecție audio), transcrieri fonetice şi notele asociate transcrierilor fonetice (Figura 1).

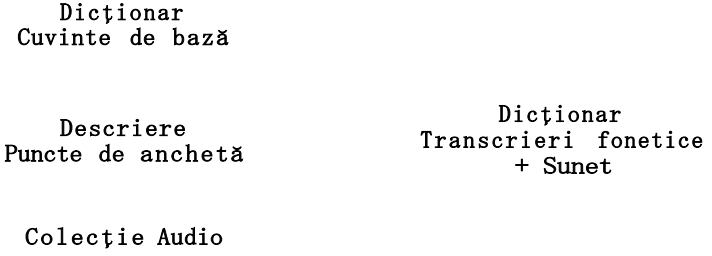


Figura 1. Dicționarele ALR

Dicționarul "Cuvinte de bază" conține fondul de cuvinte (titlul hărților atlasul lingvistic electronic, întrebările care au fost puse la anchetă, observații, şi eventual imagini. Pentru fiecare cuvânt este indicată şi între corespunzătoare care este pusă în momentul interviului.

în momentul completării acestui dicționar, utilizatorul poate vedea completă a cuvintelor de bază introduse, le poate sorta după diferite criterii, modifica articolele introduse anterior, după cum este prezentat în figura 2.

Dicționarul "Puncte de anchetă" conține informații (cod, observații) despre punctele de anchetă prezentate în cadrul atlasului lingvisti fel ca la dicționarul anterior, şi aici, utilizatorul poate vedea lista comp punctelor de anchetă introduse, le poate sorta după diferite criterii, poate m articolele introduse anterior.

2.1. Dicționar transcrieri fonetice

Dicționarul de transcrieri fonetice conține transcrierea fonet răspunsului la întrebarea pusă în etapa de interviu pentru fiecare cuvâ **Dicționarul Cuvinte de bază** în fiecare din **Punctele de anchetă,** iar acolo este posibil şi înregistrarea audio corespunzătoare din **Colecția Audio.**

Pentru claritatea hărților lingvistice, răspunsurile din punctele de a sunt însoțite de note şi comentarii (figura 3).

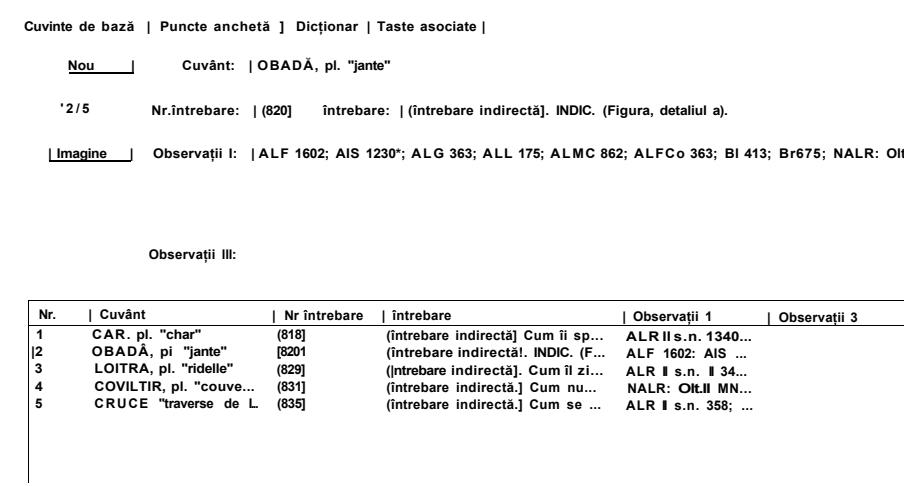


Figura 2. Fereastra de editare a listei cuvintelor de bază

Pentru transcrierea fonetică a cuvintelor din Atlasul Lingvistic Român este folosit un număr mare de fonturi, rezultat din numărul de combinații posibile ale fenomenelor fonetice prezentate în capitolul 1. Aceste fonturi au fost definite astfel încât, toate "variantele fonetice" ale unui anumit caracter să fie obținute prin selectarea caracterului respectiv într-un anumit font.

Deoarece un fişier text normal nu păstrează informații despre fonturile folosite, şi în plus transcrierile fonetice sunt realizate prin diferite poziționări ale caracterelor, s-a folosit un mod propriu de codificare a acestora.

Transcrierile fonetice sunt codificate cu ajutorul unor obiecte de tip CAIrString. Acestea sunt de fapt şiruri de obiecte de tip CAIrChar, care au următoarea descriere:

- caracterul corespunzător sunetului primar (pe 16 biti, codificare UNICODE);
- atribute:
  - poziționare: normal, deasupra sau „în umăr"";
  - mod subliniere: linie sau zigzag;
  - cursiv;
  - aldin;
- fenomene:
  - tip sunet: vocală sau consoană;
  - fenomene specifice aplicate (codificate pe biți).

Fontul folosit pentru desenarea caracterului din transcrierea fonetică este ales dinamic din lista de fonturi a aplicației, în momentul afişării.

în momentul deschiderii dicționarului de transcrieri fonetice, se fac două tipuri de verificări:

- se verifică corespondența dintre fonturile folosite la ultima editare a dicționarului şi lista curentă recunoscută de program.
- se verifică dacă toate fonturile folosite sunt instalate în Windows.

Datorită cantității mari de informație care trebuie stocată pentru Atlasul Lingvistic Român, descrierea fiecărui cuvânt este compresată folosind un algoritm de compresie LZW. 4_a selecția unui cuvânt de bază, descrierea sa este decompresată în memorie. Dacă se fac modificări ale transcrierilor fonetice, aceasta este compresată şi rescrisă în fişier la selectarea unui alt cuvânt, sau la închiderea dicționarului.

Pentru scrierea informațiilor în dicționar am proiectat o interfață utilizator prietenoasă. Operatorul trebuie sa parcurgă următorii paşi:

- selectează cuvântul titlu;
- selectează punctul de anchetă;
- editează transcrierea fonetică, nota şi comentariul asociat cuv pentru punctul de anchetă respectiv.

La editarea transcrierilor fonetice trebuie avute în vedere două aspec

- selectarea sunetului primar;
- selectarea fenomenelor asociate.

Selectarea sunetului primar se face prin apăsarea tastei corespunză dacă sunetul are un corespondent pe tastatură, sau prin apăsarea unei com de taste, dacă sunetul nu are corespondent pe tastatură. Combinațiile de sunt prestabilite în aplicație (la stabilirea combinațiilor de taste au fost p convențiile din Microsoft Word), şi cel puțin deocamdată nu pot fi modific utilizator. Pentru a veni în ajutorul celui care editează dicționarul, aplicația d de o fereastră în care sunt afişate combinațiile prestabilite de taste.

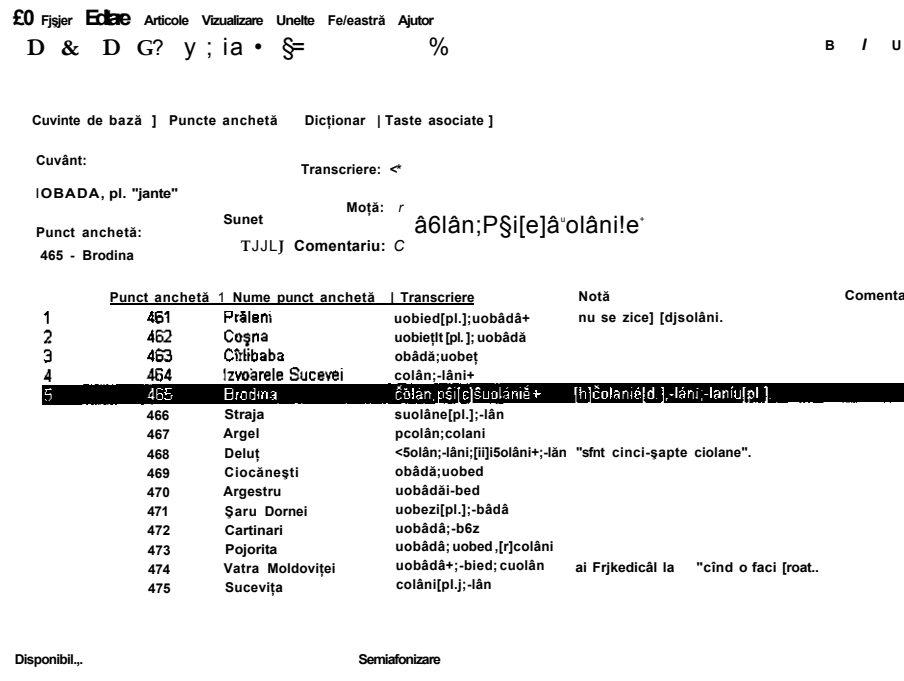


Figura **3.** Editarea Dicționarului de transcrieri fonetice

Pentru selectarea fenomenelor asociate sunetelor, operatorul are la dispoziție 2 grupe de butoane cu imaginile tuturor fenomenelor posibile pentru vocale respectiv consoane. Prin apăsarea pe unul din aceste butoane se va selecta simbolul grafic corespunzător în transcrierea fonetică. Cele 2 grupe de butoane sunt împărțite în subgrupe corespunzătoare grupelor de fenomene (vezi capitolul 1). Pot fi selectate mai multe fenomene, dar, cel mult câte unul din fiecare subgrupă. Selectarea unui fenomen, produce dezactivarea selecției anterioare din subgrupa respectivă.

După selectarea caracterului dorit, utilizatorul va specifica şi poziționarea acestuia (deasupra, în umăr) prin folosirea comenzilor PgUp/PgDown.

Fereastra de editare a transcrierilor fonetice este prezentată în figura 3.

Dicționarul cu transcrieri fonetice permite stocarea înregistrărilor audio (în format WAV) realizate în timpul anchetei.

3. Realizarea Atlasului Lingvistic Român pe Regiuni

Sistemul software care modelează atlasul lingvistic electronic, conține module care realizează gestionarea următoarelor grupe de informații:

- simboluri pentru editarea transcrierilor fonetice;
- dicționarele atlasului lingvistic (cuvinte de bază, puncte de anchetă, transcrieri fonetice).
- informații grafice pentru descrierea hărților, organizate în fişiere DXF,
- hărțile atlasului lingvistic, care pot fi consultate şi/sau tipărite;

Din punct de vedere funcțional, atlasul lingvistic electronic este structurat în două componente principale:

- Proceduri pentru pregătirea datelor primare;
- Interfața multimedia;

Aceste componente sunt prezentate în figura 4.

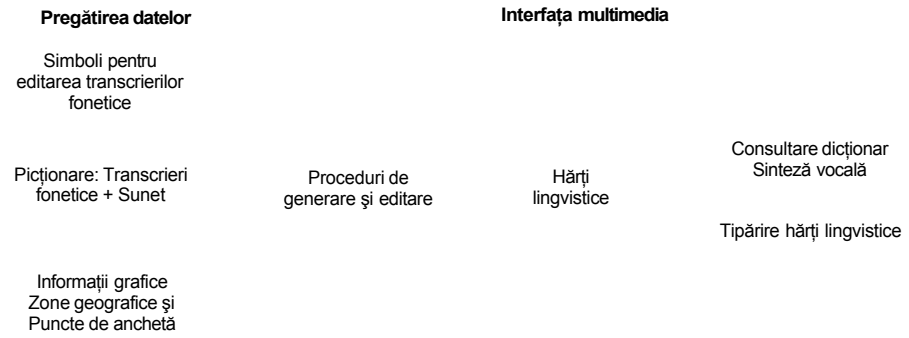


Figura 4. Componentele Atlasului Lingvistic Electronic

în continuare sunt prezentate funcțiile îndeplinite de componenta "Inter multimedia":

- generarea unei hărți noi pe baza informațiilor din dicționarele ALR informațiilor grafice primare cuprinse în fişiere DXF;
- editarea: aranjarea în pagină, selectarea informațiilor care vor fi viz implicit;
- salvarea într-un fişier numit "hartă lingvistica" a selecțiilo modificărilor din faza de editare;
- consultarea atlasului electronic:
  - vizualizarea hărților şi ascultarea înregistrărilor din punctele anchetă;
  - tipărirea hărților lingvistice.

3.1. Modulul pentru generarea şi editarea hărților lingvistice

Acest modul permite crearea descrierilor pentru hărțile lingvistice. în a scop au fost proiectate structuri de date bazate pe obiecte, suficient de flexibile permit dezvoltări ulterioare. Prezentăm în continuare structurile de date fol pentru stocarea informațiilor grafice şi a hărților lingvistice.

3.1.1. Informații grafice primare

La organizarea informațiilor grafice primare, s-a ținut cont de ceri impuse de tehnologia de realizare a atlaselor lingvistice. S-a realizat fiş NALRB.DXF care conține obiectele grafice predefinite organizate pe următoa "straturi DXF":

chenare	limitele paginii şi chenarele harții;
frontiere	conturul zonei studiate (Moldova şi Bucovina);
mijloc	locul de pliere al hărții, la legarea în volum;
municipii	localitățile importante afişate pe hartă;
puncte anchetă	dreptunghiurile în care se scriu codurile punctelor de anchetă;
transcriere fonetică	dreptunghiuri pentru încadrarea transcrierilor fonetice;
note	dreptunghiuri cu pozițiile predefinite pentru Titlu, Nota I, Nota II, Nota III;
zone	delimitări zonale în jurul punctelor de anchetă.

3.1.2. Hărțile lingvistice

Pentru editarea şi salvarea hărților lingvistice din ALR, s-a creat o structură de date care să permită în viitor, extinderea editării asistate de calculator a Atlaselor Lingvistice Româneşti Regionale la nivel național. Astfel, a rezultat o structură de date numită "hartă lingvistică" de forma următoare:

- header fişier;
- lista cu descrieri obiecte;

Descrierile de obiecte au un antet care este comun pentru toate tipurile de obiecte şi un corp obiect specific fiecărui tip în parte. Obiectele pot fi simple sau compuse. Un obiect compus conține la rândul lui alte obiecte simple sau compuse.

Au fost definite următoarele tipuri de obiecte:

- Text;
- AlrString (obiect folosit la editarea dicționarului cu transcrieri fonetice);
- Dreptunghi;
- Hartă cu transcrierile fonetice;
- Hartă sintetică (lingvistică sau fonetică);
- Notă referitoare la continuarea transcrierilor fonetice (vezi Nota II din N.A.L.R. Moldova şi Bucovina);
- Notă sintetică referitoare la cuvântul titlu (vezi Nota III din N.A.L.R. Moldova şi Bucovina);
- Legendă pentru harta sintetică;
- Simbol pe harta sintetică;
- Zonă haşurată pe harta sintetică;
- Imagine de tip bitmap;
- Strat DXF.

La activarea modulului de generare, este prezentată harta regiu punctele de anchetă şi numele localităților pe care acestea le repre Generarea hărților lingvistice se face automat, într-un format predefin momentul în care operatorul selectează un cuvânt de bază. Operatorul modifica formatul hărții, şi poate adăuga informații suplimentare. Modif realizate trebuie salvate în fişier.

3.2. Modulul pentru consultarea atlasului electronic

Componenta pentru **consultarea atlasului,** permite încărcarea une lingvistice generate / editate în etapa anterioară. Sistemul va afişa harta r respective (în situația studiată este vorba de Moldova şi Bucovina), pe ca plasa transcrierea fonetică a răspunsurilor din punctele de anchetă împreu notele şi observațiile introduse anterior (figura 5).

După ce harta lingvistică a fost încărcată, prin selecția unui pu anchetă este posibilă şi redarea înregistrării audio corespunzătoare tran fonetice asociate acestuia (înregistrarea audio sau cuvântul sintetizat).

Tot cu ajutorul acestei componente se realizează tipărirea autom hărților Atlasului Lingvistic Român, în vederea includerii lor în volum (figura 6

Pentru tipărirea hărților au fost prevăzute următoarele facilități:

- posibilitatea de selectare a informațiilor ce se vor tipări;
- tipărirea pe o pagină sau tipărirea pe două pagini cu respe locului de pliere al hărții, indicat prin linia "mijloc".

Dacă utilizatorul doreşte tipărirea transcrierilor fonetice într-un mod s (fără hartă), modulul poate asigura crearea paginilor de tip MN (M Necartografiat). Folosind această opțiune, va fi tipărită numai lista cu trans fonetice, ordonate după criterii de similaritate.

4. Concluzii

Organizarea prezentată pentru simbolurile grafice permite editar texte cu transcrieri fonetice folosind şi alte editoare de text.

Modul de selectare a fonturilor folosit la editarea transcrierilor fo poate fi extins pentru crearea unei aplicații de tip client-server sau la rea unui editor simplu, de tip WordPad.

Realizarea acestui sistem de editare a transcrierilor fonetice este î de testare şi finalizare. în continuare, ne propunem adăugarea de noi op

facilităţi, care să permită transformarea sistemului într-un instrument util cercetătorilor lingvişti.

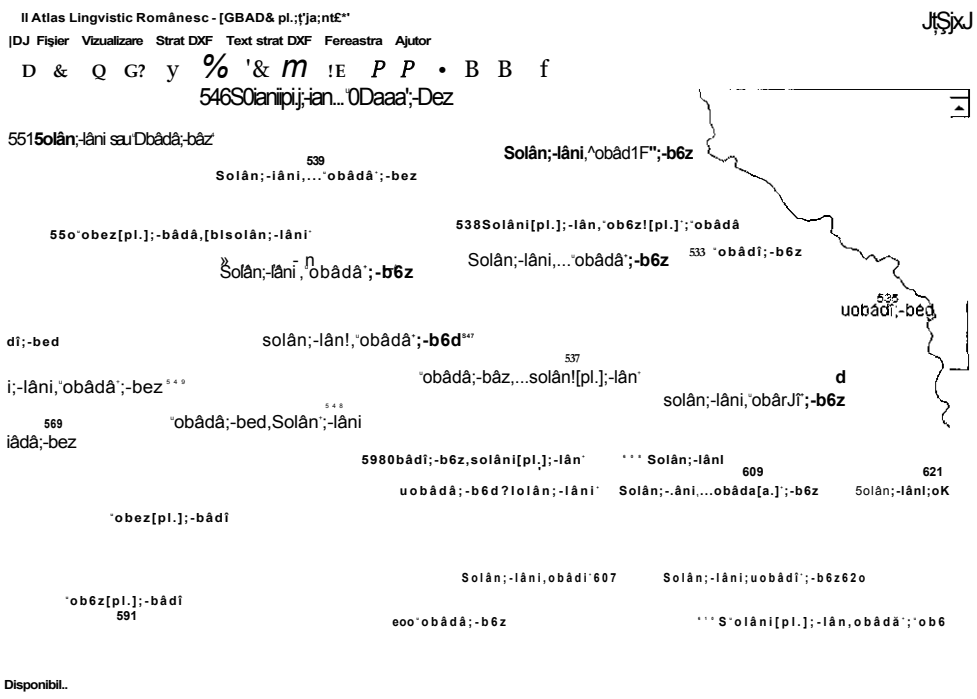


Figura **5.** Fereastra de editare / consultare a Atlasului Lingvistic

Bibliografie

[1] Academia Română, Atlasul Lingvistic Român pe Regiuni, 1987, 1997.
[2] Instituto dell'Atlante Linguistico Italiano, Atlante Linguistico Italiano, Roma, 1995.
[3] S. Bejinariu, M. Roman, V. Apopei, F. Olariu, "Sistem pentru editarea transcrierii fonetice în ALR", Zilele Academice Ieşene, Iaşi, 6 oct. 2000.

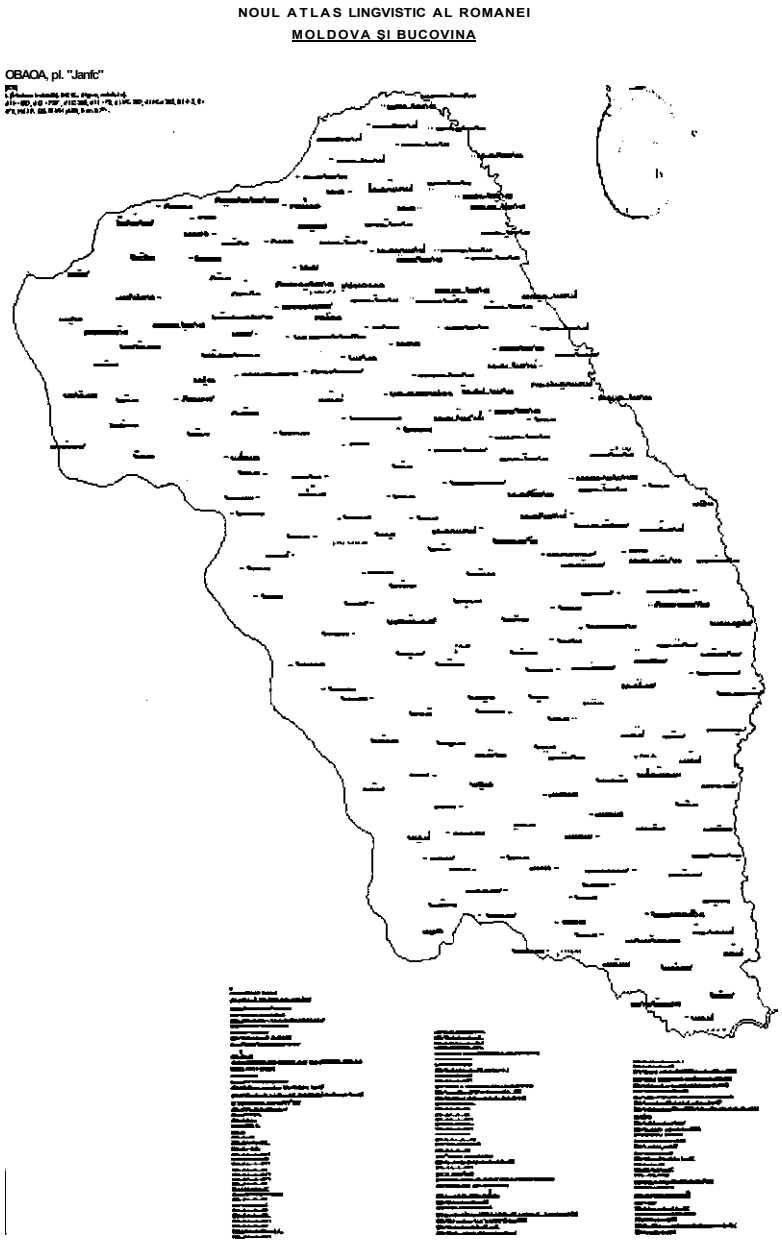


Figura **6.** Imaginea unei pagini tipărite din modulul de consultare

SECȚIUNEA IV

Dezbateri şi discuții

Asupra a doi vectori funcționali ai societății cunoaşterii: managementul cunoaşterii si învățarea electronică. Cultura si societatea cunoaşterii

Mihai DRĂGĂNESCU
Institutul de Inteligență Artificială
Academia Română

Introducere generală

Acest material, care constituie o contribuție la dezbaterea problemelor enunțate în primul volum *Societatea informațională-Societatea cunoaşterii, Concepte, soluții şi strategii pentru România, coord. Filip Gh. Florin, editat de Academia Română-Secția de ştiința şi tehnologia informației (Institutul de inteligență artificială al Academiei Române - denumire prescurtată) şi ICI-INFOSOC, Editura Expert (coordonare editorială, Valeriu Ioan-Franc), Bucureşti 2002,* are următorul cuprins:

I. **Managementul cunoaşterii, vector funcțional al societății cunoaşterii,** comunicare (Mihai Drăgănescu) prezentată la "The Sixth International Conference on Information and Communications Technology in Public Administration, Sinaia, 29 oct.2001".

II. **învățământul electronic şi societatea cunoaşterii,** comunicare (Mihai Drăgănescu) la simpozionul "E-learning (E-învățământ)", Academia Română, 28 martie 2002.

III. **Cultura şi Societatea cunoaşterii** (Mihai Drăgănescu, studiu elaborat în mai 2002).

Societatea cunoaşterii, asupra căreia se insistă cu prioritate în aceste studii şi lucrări, va fi o *perioadă intermediară între Societatea informațională şi Societatea conştiinței* (un studiu privind Societatea conştiinței este în elaborarea autorului). După cum am mai remarcat în alte lucrări, esențială pentru Societatea cunoaşterii va fi inteligența artificială (IA), atât ca vector tehnologic, cât şi prin utilizarea ei în vectorii funcționali ai societății cunoaşterii.

Această primă perioadă interimară va dura până cândva după momentul în care inteligența artificială va egala inteligența naturală (IN) structurală a omului,

respectiv a părții (IN)„„ucturai care nu poate poseda intuiție, creativitate şi spiritualitate. După concepția mea ontologică, nu este posibil pentru orice fel de inteligență artificială (electronică şi în viitor nanoelectronică) să aibă intuiție, creativitate şi spiritualitate fără a recurge şi la alte elemente ale naturii decât cele structurale şi a căror realitate devine din ce în ce mai plauzibilă. Egalitatea IA = (IN)„„„turai se va petrece, după o serie de autori (Moravec, Kurzweil, Buttuzzo, Broderick ş.a.), între 2019-2035. Unii dintre aceştia cred că atunci când se va atinge IA =(IN)„ructurai automat un asemenea creier electronic va avea şi proprietățile fenomenologice aie intuiției, creativității şi spiritualității. Ceea ce nu credem.

Din momentul" în care IA > (IN)„ructurai este evident însă că se intră într-o nouă etapă, care va produce multe consecințe pe plan social, datorită relațiilor omului cu asemenea inteligențe, unele software, altele sub forma de specii de roboți inteligenți. Aceasta va fi a doua perioadă intermediară între Societatea cunoaşterii şi Societatea conştiinței, până în momentul în care va apărea o inteligență artificială cu conştiință veritabilă, adică o conştiință artificială (CA). Din momentul în care CA > IN, se va intra în zona societății conştiinței, urmând ca societatea să fie bazată pe relațiile dintre IN (care şi ea este de presupus că va fi amplificată prin auto-transformări ale codului genetic şi probabil prin cuplaje cu sisteme informatice microelectronice şi nanoelectronice, chiar şi cu rețele internet) şi CA software sau robotice. Va trebui cu siguranță să gândim de pe acum şi asupra societății conştiinței pentru a pregăti societatea pentru o asemenea perspectivă, care nu mai apare, surprinzător, atât de îndepărtată, deoarece se poate manifesta chiar în acest secol. Societatea cunoaşterii trebuie să înceapă să fie gândită şi dezvoltată şi cu gândul la această viitoare societate.

I. MANAGEMENTUL CUNOAŞTERII

L1 Introduction

In the past XXth century a new era began in the history of humanity: *the information era* [1]. This era comprises the *information society* that will be followed naturally by the *knowledge society* and finally, somewhere more or less later in this century, the *society of consciousness.* Knowledge is a form of information [2], and consciousness is another form of information [3]. All the forms of information are intermingled with the physical and energetic realities; still they have a relative independence and can influence these realities.

To pass from the first form of information society (based essentially on Internet and Internet economy) to the second stage, the knowledge society, I considered in a previous work [1], two types of vectors: *technological* and *funcțional.*

Technological vectors are the extended Internet, the e-book a document technology, artificial intelligence (with intelligent agents and Networked Systems of Embedded Computers), nanotechnology and others.

Among the funcțional vectors of the Knowledge Society, a group of v is related to knowledge management:

- *knowledge management for corporations and enterprises, organizati and institutions,* ***local and național administrations;***
- *the management of the moral use of scientific knowledge at the glob level;*
- *e-learning management;*
- *development of a culture of knowledge andinnovation;*
- *management of the scientific and technological knowledge for every domain ofactivity as health care, sustainable society and others.*

1.2 Knowledge management

The problem of management with respect to knowledge is regarded ways:

I. As the management of the organization busy with the us integration of various types of knowledge;

II. As the management of knowledge itself, for generation o knowledge, for discovering existing knowledge (tacit or very lo externai to the organization), for combining available knowledge.

Perhaps, what is really needed is a general vision, in a unity, management of the organization and the management of knowledge.

Knowledge management for enterprises, organizations, institu local and național administrations

In the western literature, in the last years, were elaborated a se works dedicated to the problems of enterprises and knowledge. In România not have yet specialists in knowledge management in the context of the know society. We do not have either a knowledge society, but we need exp knowledge management for building the future knowledge society. A gr members of the Romanian Academy and other wellknown speciali information technology from România and colleagues from USA decid constitute a Romanian-American Foundation for the Knowledge Society, one main aims being to educate in USA a number of young Romanian specialists new domain of knowledge management. All is ready for such a Foundatio contributions of the individual founders are also ready, but no institut

organization sponsored such an exotic objective for The Knowledge Society, with some amount of money asked by the Romanian laws for a Foundation to begin its activity. But let us return to the theory of knowledge management. One definition [4] is the following:

"Knowledge Management is the conceptualizing of an organization as an integrated knowledge system, and the management of the organization for effective use of that knowledge. Where *knowledge* refers to human cognitive and innovative processes and the artifacts that support them."

This definition insists on the management of the organization, even if it recognizes the knowledge system of the organization. This definition, as it is recognized by its authors, disguise knowledge management because of the delicate problem of knowledge measurement [4]:

The recent attractiveness of the term knowledge management appears to have been prompted by three major forces:

1. Increasing dominance of knowledge as a basis for organizational effectiveness.
2. The failure of financial models to represent the dynamics of knowledge.
3. The failure of information technology by itself to achieve substanțial benefits for organizations.'

The second point, of the above quotation, is answered by many studies and books concerning the characteristics of the new economy based on knowledge (see for instance section 6 of [1]: The Economy of the knowledge society. The new economy. About the role of information in the new economy. The intangible goods).

The rapidity of the transformation of the information society into a knowledge society determines a reasoning on the new economy that takes into account not only:

a) the Internet market and the effects of Internet information on all economical and administrative agents, but also

b) the effect of knowledge as an economical and *organizational* factor that imposes the recognition of the intangible goods, in general, in the creation of economical value and organizational efficiency, and

c) the necessity of a sustainable society, an important objective for național arid even local administrations, that predictably is possible only in the frame of the knowledge society, that will demand new industries, challenges the classical economical thinking (for instance, productivity of the resources, of the energy, of materials to be more important than work productivity [8]).

The third point of the above quotation concerns the importance of contents of information, especially of knowledge, but these would not be effi without information technology. The technological vectors of the Knowle Society are equally important as the funcțional vectors.

1.3 Points of view for practicai knowledge management

Knowledge management is both the management of the organiza to use knowledge and the management of all knowledge possible, f inside and outside the organization, to attain the objectives of organization. Because knowledge is a special form of information, informa technology has to play an essential role in knowledge management. Knowle and IT are, without any doubt, going hand in hand and have a synergetic effec the efficiency of organizations.

Lucy Marshall [6] considers that knowledge management refers to control and utilization of the intellectual capital in an organization. For l Marshall, not the information, but knowledge is the most important asset o institution. This author recommends a **Chief Knowledge Officer** for an institu who based on the Intranet of the institution, has to assure the discovery creation of knowledge in the institution.

Rooney and Mandeville consider, the knowledge management at național level. The abstract of their paper is quoted [7] below:

'As the global economy becomes more knowledge intensive and wealth of nations more dependent on their knowledge assets being harnessed, essential for policy makers of having frameworks for the development and utilization of național knowledge assets. This article argues that a policy framew can be developed through which policy initiatives in a range of policy areas ca filtered in order to meet the challenges of the knowledge economy. We h developed an approach that has previously been applied to managing intelle capital in firms and adapted it to the public policy arena. In doing so we ques policy orthodoxies such at the assumption that free trade automatically facilit internațional knowledge flows, that participation in a global knowledge econ necessarily challenges național sovereignty, and that online delivery of educa is necessarily a progressive strategy'.

Peter Drucker (a wellknown professor of social science at Claren Graduate School and the author of more than thirty books, his most recent boo Management Challenges for the 21 st Century, 1999) writes [8] about **knowledge worker:**

'I am convinced that a drastic change in the social mind-set is requir just as leadership in the industrial economy after the railroad required the dr change from "tradesman" to "technologist" or "engineer."

What we call the Information Revolution is actually a Knowledge Revolution. What has made it possible to routinize processes is not machinery; the computer is only the trigger. Software is the reorganization of tradițional work, based on centuries of experience, through the application of knowledge and especially of systematic, logical analysis. The key is not electronics; it is cognitive science. This means that the key to maintaining leadership in the economy and the technology that are about to emerge is likely to be the social position of knowledge professionals and social acceptance of their values. For them to remain tradițional "employees" and be treated as such would be tantamount to England's treating its technologists as tradesmen - and likely to have similar consequences.'

1.4 Cognitive Science

The knowledege of organizations is a form of knowledge that is more and more recognized. The ways and forms of this knowledge have to be carefully studied. Cognitive science might be, indeed, the tool for this study. The cognitive science is today understood in two ways [2]:

1. As a science of human mind cognition, even if it uses models of electronic computers and electronic neural networks.
2. As a general science of cognition, that has to study cognition processes not only of the human mind, but also of animals, of artificial-intelligence systems, of the ensembles man-computer-Internet, of social organizations at the levels of institutions, enterprises, corporations, local and național administrative bodies, even at the global level.

The second way of dealing with the processes of cognition presents today the greatest interest. Such a science does not yet exist. Perhaps it is on the way. The most complex realities are the social organizations because they combine all sorts of cognitive elements, natural and artificial, but they have something more, a social body, with its own social intelligence, cognition and knowledge. To obtain new theories for such large and difficult problems, it is necessary to have talented and interested specialists in knowledge management. But it is also necessary some practice of those charged with knowledge management and knowledge work in organizations. The idea of Lucy Marshall, mentioned before, about a Chief Knowledge Officer seems to be very useful.

1.5 Final remarks

The Knowledge society is paving the way for a Consciousness society. For this we need more fundamental knowledge [9] on physical reality down to the frontier of the quantum world with the deepest reality of existence, on life, mind and consciousness, on cognition, but also on self-organization and organization of social bodies and their behavior. We need also more technological knowledg science and society, knowledge management will become the most im administration.

References

[1] Mihai Drăgănescu, *Societatea Informațională si a Cunoaşterii. Vectorii S Cunoaşterii* (Information Society and Knowledge Society.Vectors Knowledge Society), Romanian Academy, July 2001. On the Web, http. academiaromana.ro/pro__pri/

[2] Mihai Drăgănescu, Cunoaşterea în secolul XXI (Knowledge in the XXI ce communication at the Annual Conference of the Romanian Committee History and Philosophy of Science, Romanian Academy, Bucharest, 15 C 2001, to be published.

[3] Mihai Drăgănescu, *The Interdisciplinary Science of Consciousness* (Cha pp. 46-59, in *Science and the Primacy of Consciousness, Intimation 21 stCentury Revolution,* Richard L. Amoroso a.o, (eds.), Orinda, Californi Noetic Press, 2000.

[4] See http://www.uts.edu.aU/fac/hss/Departments/DIS/km/introduct.htm#Ch

[5] Ernst Ulrich von Weiszäcker, Amory B. Lovins, L.Hunter Lovins, *Facto Dublarea prosperității prin înjumătățirea consumului de resurse,* Raport Clubul de la Roma, traducere din limba germană (FAKTOR VIER.Do Wohlstand - halbierter Verbrauch, Munchen, 1995), Bucureşti, Editura te 1998.

[6] Lucy Marshall, *Facilitating knowledge management and knowledge New opportunities for information professionals,* Online. 21(5): 92-98. Sep/Oct.

[7] David Rooney and Thomas Mandeville, *The Knowing Nation: A Framew Public Policy in a Post-industrial Knowledge Economy,* Prometheus 16 453-467, 1998.

[8] Peter F. Drucker, *Beyond the Information Revolution,* The Atlantic M Digital Edition, 1999, http://www.theatlantic.com/issues/99oct/9910drucke

[9] Menas Kafatos, Mihai Drăgănescu, *Preliminaries to the philoso integrative science,* e-book, MSReader format, Academy of Scien România, Bucharest, 2001, (available free by e-mail: dragam@racai.ro).

II. ÎNVĂȚĂMÂNTUL ELECTRONIC ŞI SOCIETATEA CUNOAŞTERII

ILI Introducere. Sintagma Societății cunoaşterii.

în *societatea cunoaşterii* doi vectori, strâns legați între ei, unul tehnologic - *cartea electronică* - şi altul funcțional - *învățământul electronic* - sunt chemați să joace un rol important în desfăşurarea acesteia.

Problematica societății cunoaşterii a fost abordată în țara noastră începând din anul 2001 la Academia Română [1], la Academia de studii economice [2] şi de revista Diplomat-Club [3]. Primul politician român care a folosit sintagma societății cunoaşterii (din anul 2001) a fost preşedintele României şi protectorul de fapt al Academiei Române, Ion Iliescu.

Este poate interesant de amintit că în anul 1986, în lucrarea Tendencies of becoming' [4] (Tendințele devenirii, republicată în volumul [5]) se justifică şi foloseşte sintagma 'societatea cunoaşterii':

"Cine nu face legătura dintre revoluția microelectronică şi informațională şi tendința devenirii istorice nu înțelege vremurile. Cine se opune acestei revoluții părăseşte linia devenirii istorice. Şi totuşi nici această revoluție nu trebuie absolutizată întrucât trebuie să fie însoțită şi de alte schimbări. Atunci nu ne putem fixa numai asupra ei, ci asupra unui context mai larg în cadrul căruia ea poate juca rolul principal o anumită perioadă istorică. ***Tendința devenirii istorice se conturează a fi tendința către o societate a cunoaşterii,*** *a creației şi a civilizației, către o societate globală şi către o societate interastrală în univers, apoi către un act cosmic în conformitate cu tendința existențială a universului. Mai aproape de noi, ca urmare a revoluției microelectronice şi informatice, a unei noi revoluții industriale, se deschid perspectivele unei societăți orientate informațional...".*

Era o viziune, în **acel** moment, legată de o anumită filosofie pe care am dezvoltat-o în anii 1980, viziune ancorată şi în realitatea electronică şi informatică a **ceea** ce **se** va numi era informației.

IL2 Cartea electronică

Cartea electronică este un vector tehnologic. La Academia Română în anul 2001 s-a desfăşurat un simpozion referitor la cartea electronică şi s-a publicat un volum de referință sub coordonarea prof. Doina Banciu [7]. Atunci am descoperit firma de software SOFTWIN condusă de Florin Talpeş care lucrase în domeniu şi avea un prestigiu internațional în producerea de cărți electronice. Softwin este participantă la elaborarea specificațiilor internaționale OPEN E-BOOK care au st formatul ediției de cărți electronice de interes public.Ca urmare a simpozionu fost înființată şi o librărie de software, cărți şi documente electronice la Inst Național de Cercetare-Dezvoltare în Informatică (http:// www.e-librarie.ro).

Despre cartea electronică şi rolul ei pentru societatea cunoaşter România, am expus considerațiile mele în lucrări anterioare [1b], [7] şi nu reveni asupra lor. în schimb, voi cita doi autori, unul care a exprimat opinii în de apariția cărții electronice propriu-zise, altul care a participat la lansarea electronice. Primul este Paul Saffo, directorul unui elevat Institut al Viitorulu California, care lucrează, foarte scump, numai pentru marile companii america care în anul 1988 prevedea că o carte electronică va fi mai mult decât o tipărită datorită posibilităților de a introduce elemente audio, video, conexiu informații pe rețea. El scria [8]:

'The term "electronic book" is misleading because these products ar books at all, but something new. We are living in a moment betwee revolutions: one of prinț, four centuries old and not quite spent and anoth electronics, two decades young, and just getting underway. Today's pro amount to a bridge between these two revolutions.

Al doilea este Dick Brass, Vicepreşedinte Microsoft pentru dezv tehnologică, care în anul 2000, an în care cartea electronică propriu-zisă d scria [9]:

'If you don't think eBooks will take off, remember that elec encyclopedias have already outsold all paper encyclopedias. [...] They cos than $100, instead of the $2,000 or more for fine paper encyclopedia Similarly, after the triumph of eBooks, paper books will no longer be the pr means of distributing information. But, like horses they will continue to ex pleasure...[...] Like ăll transitions, the move from pBooks to eBooks will be painful and tentative at first. Then, in less than 20 years, eBooks will pervasive that we won't be able to remember living without them. [...] We the verge of the most exciting change to the printed word since movable type

Cartea electronică a decolat. Firme precum Amazon şi Barnes and din SUA sunt cunoscute în întreaga lume pentru modul în care au promovat sunt o adevărată şcoală pentru toți cei care conduc şi vor conduce librării electronice şi software, şcoală accesibilă gratuit prin simpla experimenta Internet pe web-site-urile acestor firme.

IL3 Procesul de învățare

în anul 1988 scriam despre procesul de învățare [10]:

•înțelegerea profundă a procesului de învățare depinde de exp funcționării creierului şi a minții omului, în ultimă instanță de înțelegerea

materiei vii. Cu alte cuvinte, natura intimă a procesului de învățare nu va putea fi elucidată într-o măsură într-adevăr mulțumitoare decât atunci când ştiința va face un nou mare pas în cunoaşterea materiei. Cercetările din domeniile fizicii şi biologiei, esențiale pentru elucidarea naturii materiei vii, se vor îmbina cu cele din domeniul ştiinței informației. Activitatea creierului este în principal o activitate informațională, iar *procesul de învățare este un proces informațional.'*

în acea perioadă ştiința cognitivă se găsea, este adevărat, în perioada post-behavioristă şi se baza pe modelarea simbolică de tip calculator electronic, ceea ce s-a dovedit insuficient pentru înțelegerea mulțumitoare a proceselor cognitive mentale [11]. De aceea procesul de învățare nu era de fapt explicat şi înțeles din punct de vedere ştiințific. în anii 1990, modelarea proceselor cognitive a cunoscut aportul adus de utilizarea modelelor bazate pe rețele neuronice (de tip natural ca în creierul omului) şi neural (artificiale, electronice), dar nici acestea nu au dus încă la o ştiință cognitivă bine constituită [11]:

COGNITIA:

Anii 1970 şi 1980	***Anii 1990***	***Anii 2000. Ce va urma?***
(modelare simbolică tip calculator)	(efectul conectivismului, rețele neuronice şi neurale)	Efectul ştiinței integrative

O speranță este aceea ca în sec. XXI ştiința cognitivă să fie consolidată prin luarea în considerare atât a proceselor fenomenologice (qualia, experiențiale) ale minții, cât şi a rolului proceselor sociale în procesele cognitive (socialul referindu-se nu numai la persoane umane, ci şi la grupuri de inteligențe artificiale sau la grupuri mixte). Un asemenea mod de abordare se încadrează în viziunea unei ştiințe numite integrative [12]. Tot în anul 1988 remarcam [10]:

'Un interes deosebit prezintă cercetările din domeniul inteligenței artificiale, domeniu care este studiat în ultimii ani şi din punctul de vedere al capacității de a învăța. Studiul procesului de învățare de către inteligența artificială ar putea oferi multe elemente utile pentru înțelegerea procesului de învățare al inteligenței naturale a omului. [...] Inteligența presupune şi capacitatea de a învăța. [...] Gh. Tecuci, într-o lucrare originală în care se prezintă un sistem expert la care asociază un sistem de învățare automată [13], deşi constată că "învățarea este un proces cognitiv în cea mai mare măsură necunoscut" [13], arată şi demonstrează prin sistemul său că "forme efective de învățare automată sunt posibile".

Dintre aceste forme de învățare automată pot fi amintite [13]:

- *învățarea pe de rost şi implantare directă de noi cunoştințe* (când este mai eficient să se regăsească o cunoştință în memorie decât să se producă acea cunoştință).
- *învățare prin instruire* (sistemul primeşte cunoştințe de la un p le integrează cu cunoştințele anterioare).
- *învățarea prin analogie.*
- *învățarea din exemple prin detecție de similarități,* proces mente inductiv (fără a exclude şi procese deductive) prin gene exemplelor pozitive, generalizare care evită exemplele negati
- *învățarea prin observare şi descoperire* (spre exemplu a unc rități în structurări de date).

Gh. Tecuci înclină către o *îmbinare de metode de învățare.* Fără câteva lucruri credem că se susțin pentru procesul uman de învățare:

- *Necesitatea unei varietăți de metode,* şi nu o monometo deosebit de important când ar putea apare tendința de a ne viitor, mai mult pe tehnologie în procesul educațional.
- *Obținerea unui sistem de cunoştințe* sub forma unui model bază (unor modele interne) la care să se poată racorda uş tințe de detaliu provenite din exterior eventual prin metode inf
- *Obținerea sensurilor cunoaşterii,* a sensului 'fizic', al intuiției l chiar a unui răspuns creativ în procesul învățării, lucru de matele nu sunt capabile, adică a pune umanul în starea lui fir
- O *deschidere firească spre creativitate şi creație,* spre inovare rea rezolvării de probleme care nu sunt structurate după t mului de cunoştințe existent în modelul intern disponibil la u dat.'

Odată cu apariția e-învățării se deschid perspective noi şi pentru experimentală a procesului de învățare şi confruntarea acestui tip im proces cognitiv cu teoriile ştiinței cognitive care se vor baza pe progrese le va realiza, ceea ce numim, ştiința integrativă. Considerații privind pr învățare şi sisteme de e-educație, inclusiv prin folosirea metodelor i artificiale sunt prezentate într-un grup de trei lucrări recente ale unor ştiințifici de la Centrul pentru Cercetări Avansate în învățarea Prelucrarea Limbajului Natural şi Modelare Conceptuală şi Institutul de 'Mihai Ralea' al Academiei Române [14], [15], [16].

National Research Council de pe lângă Academia Națională de SUA a prezentat, în februarie 2002, un raport [17] privind cercetarea educației în care despre studiul ştiințific al procesului de învățare se arată

'Much of the controversy about education research relates to its lack of quality. [...] Is scientific education research the same as researc and behavioral science generally or the same as research in th sciences? [...] A key finding of this NRC committee is that at a fundame

scientific inquiry in education is no different from scientific inquiry in other fields and disciplines. A set of basic principles is common to all scientific endeavors: these principles include concepts like linking empirical data to theoretical models, using appropriate methods, applying rigorous reasoning, striving toward generalization.'

Considerațiile de mai înainte, inclusiv ale informaticienilor români, arată cât de deschis este în continuare câmpul cercetărilor privind procesul de învățare, în special al omului.

IL4 învățământul electronic (e-learning)

E-learning este un vector funcțional al societății cunoaşterii. învățarea electronică înseamnă a învăța folosind mijloace electronice, ceea ce se poate face în mai multe moduri:

- Individual - folosind resursele existente pe Internet şi CD-uri.
- Instituționalizat - în şcoli şi universități sau organizat în întreprinderi sau de către fundații. Cursurile prin televiziune vor ceda locul cursurilor prin Internet, dar acest procedeu se va desfăşura sub supravegherea şi îndrumarea cadrelor didactice calificate.
- în cursul activității practice, din orice domeniu, care se va desfăşura şi într-un mediu informațional şi de cunoaştere.

Cei care învață sunt persoane, dar şi agenți inteligenți. în viitorul imediat, agenții inteligenți vor deveni nu numai studenți, ci şi profesori, dar rolul lor cel mai promițător este acela de colaborator cu persoane. învățarea implicând agenții inteligenți va deveni o etapă esențială în societatea cunoaşterii, deoarece ***în regim de croazieră societatea cunoaşterii se va baza în cele mai multe activități pe agenți inteligenți. Inteligența artificială va fi esența tehnologică a societății cunoaşterii.*** Ea va antrena internetul, nanotehnologiile, dar şi vectorii funcționali ai societății cunoaşterii [1b]. Inteligența Artificială în primii 20 de ani ai sec. XXI va depăşi inteligența omului (numai pentru aspectele structurale, fără intuiție şi creativitate).

E-învățământul se găseşte astăzi în plină dezvoltare [18], [19], [20], [21], [22]. Din experiența relatată în asemenea studii rezultă:

- Studenții găsesc, chiar în cazul lipsei unei interacțiuni față în față între profesor şi student, că descărcarea notelor de curs prin Internet, corespondență prin e-mail cu profesori şi instructori, examene prin răspunsuri date pe calculator, acasă sau la şcoală, acest e-învățământ este foarte agreabil. lar performanțele studenților şi elevilor sunt similare (evaluare pentru anul 2000) cu cele ale învățămâ clase de elevi şi studenți.
- Corporațiile industriale recurg masiv la e-educație, iar această nu mai poate fi ignorată. Unele corporații au lansat e-universită personalul propriu, de ex. Dell Computer Corp. şi Sun Microsys
- Universitățile au început să introducă nu un e-învățământ co constituirea treptată a acestuia prin unele e-cursuri. Spre University of California, Berkeley, în domeniul ştiinței şi te informației a început (anii 1999-2000) cu patru e-cursuri: informatice, telecomunicații digitale, e-comerț, sisteme inform geografice.
- O serie de firme şi-au dedicat activitatea sau o parte din producerii unor 'e-learning software packages'. Se cons segment al pieții software specializat în e-learning. (Astfel se ş prezența firmelor SOFTWIN şi SIVECO la acest simpozion principalele firme româneşti de software educațional). Dar as pachete e-software pentru învățământ sunt de aşteptat şi di Programului e-şcoală al Ministeului Educației şi Cercet urmăreşte o reformă educațională în România.
- Nu se constată deosebiri între rezultatele învățării on-line şi într-un campus universitar sau o şcoală. învățarea electronică multă disciplină şi maturitate decât învățarea convențională [1
- Pentru experimente de laborator şi pentru viață socială est totuşi de perioade de lucru în instituțiile de învățământ.
- Odată cu creşterea utilizării metodelor de e-învățământ, cons clădiri pentru învățământ se va diminua. în schimb apar pentru noua infrastructură a e-învățământului.
- Modul asincron de acces la cursuri permite e-educația în orice şi în orice loc.
- E-învățământul încurajează studenții să-şi asume o m responsabilitate pentru definirea şi organizarea a ceea ce urm învețe. Studenții sunt mai bine serviți având un acces asistat on-line la cei mai buni instructori decât un contact față în instructori mediocri [19]. In orice caz, nu se neagă rolul instruc
- Discipline ca filosofia şi istoria presupun discuții, iar disciplin presupun proiecte. în aceste cazuri trebuie încă să se găseas mixte de învățământ clasic şi electronic.
- E-învățământul oferă cele mai bune perspective pentru înv întreaga viață (învățarea continuă).

- 'Educația bazată pe Internet resuscită probleme fundamentale ale educației care sunt importante pentru conceperea activităților educaționale'. [19]
- Gradul în care instructorii vii pot fi înlocuiți cu agenți inteligenți specializați nu este încă clarificat.
- în mod diferit se pun problemele e-învățământului în şcoli elementare şi licee în raport cu învățământul superior. Pentru şcoli şi chiar licee, într-o primă etapă se dezvoltă clase conectate la Internet, cu calculatoare personale, dotate cu e-books, e-learning books, discuri compacte şi acces la rețele specializate, eventual servere de clasă sau şcoală.
- Şcolile, ca şi companiile, ca şi guvernul, trebuie să se regândească în lumina noilor tehnologii ale societății cunoaşterii.
- Se preconizează şi se experimentează atât pentru şcoli, cât şi pentru alte forme de învățământ, utilizarea Internetului prin comunicații fără fir (wireless Internet) care oferă posibilități şi opțiuni noi.

Acestea sunt principalele considerații şi constatări la începutul anului 2002. Valabilitatea unora dintre ele se va confirma, alte constatări vor fi, poate, infirmate, dar vor apare cu siguranță multe alte aspecte noi.

II.5 Viata intelectuală

în timp sunt prevăzute multe schimbări datorită învățământului electronic [23]. în primul rând, apariția unor colegii şi universități nelocalizate, extinse uneori la scară globală. Siturile acestora pot fi mari sau mici, structurarea socială având loc sub forma unor comunități (villages) având facilități comune pentru cercetare, proiecte de grup, dar şi pentru activități comunitare culturale, sportive etc. O persoană admisă într-o asemenea universitate îi va rămâne ataşată pentru toată viața, deoarece educația se va extinde pe întreaga viață prin perioade discrete (adică necontinue) şi intensive de învățare. Viața intelectuală se va schimba foarte mult, reflectând modificările în cunoaştere:

'An epistemic change is the abandonment of the notion that any single human mind can bear any significant fraction of what is knowable...Even the renaissance notion of an 'educated person' has been discarded - there is no longer a canonica! body of basic knowledge that defines this notion' [23].

Agenții inteligenți de căutare a informației, bibliotecile electronice, vizualizarea informației, pătrunderea în medii virtuale, toate acestea vor constitui un software care devine literatură [23]. 'Tehnologia va fi văzută ca cea mai bogată dezvoltare în cultura umană' [23]. Rădăcinile intelectuale se vor baza pe inginerie şi tehnologie: Difuzia umanităților în tehnologie şi invers, vor duce la o reorganizare radicală a disciplinelor intelectuale [23].

IL6 Perspective

Cum vor evolua lucrurile în viitor? Ray Kurzweil [24] previziuni privind educația pentru anii 2009, 2019 şi 2029:

Pentru anul 2009 [24, p. 191-192]:

'...most effective learning from computers taking place The profound importance of the computer as a knowle recognized. Computers play a central role in all facet they do in other spheres of life. The majority of reading is although the 'installed base' of paper documents is st generation of paper documents is dwindling, however, other papers of largely twentieh century vintage are beir and stored. Documents circa 2009 rotinely include e images and sounds. Students of all ages typically have own, which a thin tabletlike device weighing under a high resolution display suitably for reading. Students computers primarily by voice and by pointing with a devi pencil. Keyboards still exist, but most textual langua speaking. [...]

Intelligent courseware has emerged as a common mea The tradițional mode of a human teacher instructing a g still prevalent, but schools are increasingly relying on sof leaving human teachers to attend primarily to issu psychological well-being, and socialization.'

Pentru anul 2019 [24, p.204):

'Paper books and documents are rarely used or accesse century papers of interest have been scanned and are wireless network. Most learning is accomplished using i based simulated teachers.[...] The teachers are viewed and counselors than as sources of learning and kno continue to gather together to exchange ideas and to s even this gathering is often physically and geographicall adult human workers spend the majority of their time ac and knowledge.'

Pentru anul 2029 [24, p. 221]:

'Human learning is primarily accomplished using virtua enhanced by the widely available neural implants. The memory and perception, but it is not possible to dow directly. Although enhanced through virtual experi interactive instruction, and neural implants, learning s

consuming human experience and study. This activity comprises the primary focus of the human species.

Automated agents are learning on their own without human spoon-feeding of information and knowledge. Computers have read all available human and machine generated-literature and multimedia material ...Significantly new knowledge is created by machines with little or no human intervention. Unlike humans, machines easily share knowledge structures with one another.'

Dacă în societatea cunoaşterii previziunile de mai înainte se bazează pe o continuare a ştiinţei structurale, ce se va întâmpla dacă ştiinţa, cu bazele ei noi, integrative, va conduce şi la apariţia inteligenţei artificiale conştiente, adică a conştiinţei artificiale? Acest lucru nu se va întâmpla probabil în primii 30 de ani ai acestui secol, dar dacă se va întâmpla cum vom privi şi acţiona în activitatea educaţională?

II.7 încheiere. Propuneri

Roger Bohn defineşte, într-un mod specific pentru societatea cunoaşterii, învăţarea drept evoluţia cunoaşterii în timp [25].

Studiul procesului de învăţare scoate în relief importanţa ştiinţei cognitive şi a învăţării ca proces cognitiv fundamental. Această ştiinţă trebuie nu numai cunoscută, atât cât este ea astăzi, ci mai ales dezvoltată de către psihologi, neurobiologi, sociologi şi specialişti în inteligenţa artificială.

Este necesară o direcţie de cercetare bine susţinută pentru a stimula contribuţii româneşti în acest domeniu. Am propus şi propun în continuare ca în cadrul programului INFOSOC (Programul naţional de cercetare-dezvoltare pentru societatea informaţională) să se stimuleze cercetări în domeniul ştiinţei cognitive care să contribuie la depăşirea limitelor actuale ale acestui domeniu.

Este, de asemenea, necesară o dinamizare nu numai a cercetărilor, dar mai ales a dezvoltărilor şi realizărilor concrete în domeniul inteligenţei artificiale. Există un sistem românesc, sistemul DISCIPOL, creat de acad. Gh. Tecuci [13], [26] la ICI şi apoi la George Mason University din SUA. Ar trebui examinat şi utilizat şi la noi. Ar trebui să cunoaştem ce posibilităţi şi ce potenţial avem în domeniul utilizării agenţilor inteligenţi şi să existe o coordonare şi autocoordonare a eforturilor. Utilizarea agenţilor inteligenţi pentru toţi vectorii societăţii cunoaştşrii, inclusiv pentru e-învăţământ va deveni determinantă pentru calitatea şi eficienţa acestei societăţi. Recenta propunere pentru transformarea Centrului pentru Cercetări Avansate în învăţarea Automată, Prelucrarea Limbajului Natural şi Modelare Conceptuală al Academiei Române într-un Centru de cercetări pentru Inteligenţa Artificială şi Societatea Cunoaşterii sprijinită de Directorul general ICI, Doina Banciu şi de Ministrul Comunicaţiilor şi Tehnologiei Informaţiei, Dan putea să satisfacă aceste cerinţe actuale şi de viitor. Sperăm ca şi Ac Română să sprijine această solicitare pentru a putea fi înaintată Gu României spre a fi aprobată.

Tot la Academie, Comitetul Român pentru Istoria şi Filosofia Ş Tehnicii va acorda o anumită importanţă muzeelor virtuale, nu numai pentr ştiinţei şi tehnicii, dar şi pentru cunoaştere şi învăţare. Ar trebui realizat un de sinteză a tuturor muzeelor virtuale din lume, inclusiv ai web-site-urilor u muzee de mare tradiţie şi importanţă, cu adresele lor pe Internet. Acest we trebui să fie cunoscut şi accesibil tuturor în România.

Apreciez în mod deosebit eforturile care se fac pentru inform învăţământului românesc de către Guvernul României, Ministerul Edu Cercetării, firmele SIVECO şi SOFTWIN, ca şi de toate instituţiile reprez acest simpozion dedicat învăţământului electronic.

Doresc să mulţumesc tuturor celor care au prezentat comunicări simpozion şi celor care au participat la organizarea lui.

Referinţe bibliografice

[1] Mihai Drăgănescu, *Cunoaşterea şi societatea cunoaşterii,* comu seziunea de lansare a programului strategic SI-SC, Academia Ro aprilie 2001; 1b. Mihai Drăgănescu, *Societatea informaţională şi a c Vectorii societăţii* ct//?oaşfer/7,studiu, Academia Română, 7 iulie 2001 pe Internet şi în voi. coord. Florin Gh. Filip, *Societatea info Societatea cunoaşterii. Concepte, soluţii şi strategii pentru Români,* Română, 2002, p.43 -112.

[2] Gabriela S. Sabău, *Societatea cunoaşterii. O perspectivă românească* economică, Bucureşti, 2001; Ion Gh. Roşea, Viorel Petrescu, Cotigaru, Gabriela Sabău, Vasilica Ciucă, Oscar Hoffman, Wilhei *Cercetarea pentru dezvoltarea în reconstrucţia durabilă a eco perspectiva societăţii cunoaşterii,* Economistul, 4 februarie 2002, nr.27

[3] Mihai Drăgănescu, *Societatea cunoaşterii,* Diplomat Club, 2001, N Mihai Drăgănescu , *Knowledge management, a funcţional ve knowledge society,* Diplomat Club, Nr. 10-11, 2001, p.4; Mihai Dr *Factori noi în viaţa cultural-ştiinţifică-politică globală: ter antiterorismul,* Diplomat Club, 2002, Nr.1, p.7.

[4] Mihai Drăgănescu, *Tendencies of becoming,* Romanian Review, 19 p.55-59.

[5] Mihai Drăgănescu, *Spiritualitate, Informație, Materie,* p.23-28, Ed. Academiei R.S.R., 1988.

[6] coord. Doina Banciu, *Cartea Electronică,* Editura AGER, Bucureşti, 2001.

[7] Mihai Drăgănescu, *Societatea cunoaşterii şi cartea electronică,* în voi. coord. Doina Banciu, Cartea Electronică, Editura AGER, Bucureşti, 2001, p. 26-42.

[8] Paul Saffo, Institute for the Future, *Electronic books,* http://www.saffo.org/sflibrarv.html. 1988.

[9] Dick Brass, Vicepreşedinte Microsoft pentru dezvoltare tehnologică, *E-books,* în voi. Inside/Out, Microsoft- in our own words, Penguin Books, New York 2000, p.262-263.

[10] Mihai Drăgănescu, *Microelectronica şi învățământul în domeniul electronicii (I),* Forum, anul XXX, noiembrie 1988, p. 36-48.

[11] Mihai Drăgănescu, *Ştiința cognitivă, ştiință structurală sau ştiință integra-*f/Vă?Comunicare la seziunea ştiințifică de toamnă AOS-R, Bucureşti, 9 noiembrie 2001, E-PREPRINT, MSReader format, november 2001.

[12] Menas Kafatos, Mihai Drăgănescu, *Preliminaries to the Philosophy of IntegrativeScience,* MSReader e-book, Editura ICI, Bucureşti, 2001, ISBN 973-10-02510-X.

[13] Gheorghe Tecuci, *Mediu de dezvoltare a sistemelor expert instruibile pentru proiectarea asistată de calculator,* Teză de doctorat, Institutul Politehnic, Bucureşti, 1988.

[14] Ştefan Trăuşan-Matu, *Achiziția, gestiunea, partajarea şi prelucrarea cunoş-tințelor pe web: elemente esențiale în societatea cunoaşterii,* în voi. coord. Florin Gh. Filip, *Societatea informațională-Societatea cunoaşterii. Concepte, soluții şi strategii pentru Români,* Academia Română, 2002, p.'l 95-207.

[15] Cristina V. Niculescu, *Noi tipuri de sisteme educaționale pentru SI-SC,* în voi. coord. Florin Gh. Filip, *Societatea informațională-Societatea cunoaşterii. Concepte, soluții şi strategii pentru Români,* Academia Română, 2002, p.209-223.

[16] Gheorghe Iosif, Ana Măria Marhan, Ion Juvină, *Strategii de creştere a utizabilității şi de dezvoltare a competențelor de bază ale populației României pentru utilizarea tehnologiei informației,* în voi. coord. Florin Gh. Filip, *Societatea informațională-Societatea cunoaşterii. Concepte, soluții si strategii pentru Români,* Academia Română, 2002, p. 225-235.

[17] Lisa Towne, Study Director, Committee on Scientific Principles in Education Research National Research Council/National Academy of Sciences, *Statement before the Subcommittee on Education Reform Committee on Education and the Workforce United States House of Repre* February 28, 2002.

[18] Robert Ubell, *Engineers turn to e-learning,* IEEE Spectrum, Octo p.59-63.

[19] Peter Wiesner, *Distance Education: Rebottling or a New Brew?* Pr of the IEEE, July 2000, p.1124-1130.

[20] Ralph B. Ginsberg, Kenneth R. Foster, *The Wired Classroom,* IEEE August 1998, p.44-51.

[21] Paul G. Shotsberger, Ron Vetter, *Teaching and Learning in th Classroom,* Computer, march 2001, p.110-111.

[22] http://www.microsoft.com-education

[23] Edward A. Lee, David G. Messerschmitt, *A higher education in the* Proceedings I.E.E.E., September 1999, p.1685-1691.

[24] Raz Kurzweil, *The Age of Spiritual Machines,* Penguin Books, 1999.

[25] Roger E. Bohn, *Measuring and Managing Techological Knowled* 314, în voi. Eds. Dale Neef a.o., The Economic Impact of K Butterworth-Heinemann, Boston, 1998.

[26] Gh. Tecuci, *Building Intelligent Agents,* Academic Press, San Diego,

III. CULTURA ŞI SOCIETATEA CUNOAŞTERII

Societatea Cunoaşterii

Am prefigurat că va sosi un moment al societății cunoaşteri această sintagmă, Mihai Drăgănescu, 1976, 1986), dar abia în ultimul secolului XX conceptul s-a impus în SUA datorită lucrărilor sociolog Drucker şi ale altora, în ultimii 4-5 ani societatea cunoaşterii devenind r ca o etapă nouă a erei informației, respectiv a societății informaționale. Română a lansat acest concept în România în anul 2001 ca urmare a comunicării lui Mihai Drăgănescu, *Cunoaşterea şi Societatea cun* sesiunea de lansare a programului SI-SC, Academia Română, 10 aprili elaborării studiului lui Mihai Drăgănescu, *Societatea Informațio Cunoaşterii. Vectorii Societății Cunoaşterii,* Academia Română, Bucure 2001, publicat apoi în voi. coord. Florin Gh. Filip, *Societatea in Societatea cunoaşterii. Concepte, soluții şi strategii pentru România,* Română, 2002, p.43-112.

Spre deosebire de unele puncte de vedere care privesc numai (economia digitală, piața internet) societatea cunoaşterii nu este numai

bazată pe cunoaştere. Aceasta este foarte importantă, decisivă, esențială şi cuprinde utilizarea şi managementul cunoaşterii existente sub forma cunoaşterii tehnologice şi organizaționale, producerea de cunoaştere tehnologică nouă prin inovare, **o *nouă economie*** în care procesul de ***inovare*** este determinant, *în care bunurile intangibile devin mai importante decât cele tangibile.*

Societatea cunoaşterii **reprezintă mult mai mult** deoarece asigură o diseminare fără precedent a cunoaşterii către toți cetățenii prin mijloace noi, folosind cu prioritate Internetul şi cartea electronică şi metodele de învățare prin procedee electronice (e-learning), urmăreşte extinderea şi aprofundarea cunoaşterii ştiințifice şi a adevărului despre existență, *este singurul mod prin care se va asigura o societate sustenabilă din punct de vedere ecologic* şi va fi o *nouă etapă în cultură* (bazată pe cultura cunoaşterii care implică toate formele de cunoaştere, inclusiv cunoaşterea artistică, literară etc).

în fine, societatea cunoaşterii asigură bazele unei viitoare societăți a conştiinței, a adevărului, moralității, creativității şi spiritului.

Pentru realizarea societății cunoaşterii am definit, în studiul amintit mai înainte, o serie de vectori (tehnologici şi funcționali) care ar trebui introduşi în acțiune într-o succesiune firească pentru posibilitățile țării noastre.

Categoriile culturii

Dintre lucrările pe care le-am publicat anterior în problemele culturii [1] două se referă la teoria culturii. în *Perspectiva informațională a culturii* (1983) găseam un anumit sprijin pentru o viziune informațională a culturii în teoria semiotică a culturii elaborată de Umberto Eco în *Tratatul* său *de semiotică generală.* Umberto Eco propunea o ipoteză radicală prin care întreaga cultură este considerată un fenomen semiotic şi o ipoteză moderată prin care orice aspect al culturii este o entitate semantică. Semiotica se referă la semne cu conținut semantic astfel încât cele două ipoteze nu sunt prea deosebite. De aceea, consideram, prin generalizare firească, deoarece semnul şi semanticul (de semnificație şi de sens) sunt informație, o posibilă perspectivă informațională a culturii. Acest lucru, faptul că **esența culturii este informațională,** chiar dacă ea se manifestă prin comportamente socio-umane, obiecte materiale şi informaționale, a devenit tot mai evident. Nu trebuie să surprindă această esență informațională a culturii, astăzi fiind ştiut că şi inteligența şi conştiința sunt informație.

în legătură cu perspectiva informațională a culturii poate fi menționat ca precursor al acestei abordări, Ernst Cassirer [2] care considera că expresia culturală a omului şi societății este caracterizată de activitatea de creare a simbolurilor (activitatea simbolizatoare) generate de imagini mentale. Pentru Cassirer, simbolul este o cheie pentru înțelegerea naturii omului, iar, trăieşte numai într-un univers material, ci mai ales într-unui simbolic [3].

într-o a doua lucrare [1a], *Cultura şi marile tehnologii* (1996) ar linia clasică de definire a culturii ținând însă seamă de obiectele informați aduse de societatea informațională. **în teoria clasică cultura este defin fenomen social care cuprinde comportamentul socio-uman cu materiale şi informaționale integrate acestui comportament.** informaționale au fost introduse în această definiție la sfârşitul secolului X

Pare a fi posibilă o încadrare a teoriei culturii într-o viziune c (termenul este utilizat în raport cu teoria categoriilor şi functorilor din ma extinsă recent de la domeniul structural la domeniul structural-fenomenol

Privind comportamentul socio-uman *cultural* ca o categorie categorie este o subcategorie majoră a categoriei comportamentului so *general.* Ultima mai cuprinde şi o subcategorie a comportamentului strict biologic, atât la nivel individual, cât şi social. într-adevăr, comportamente individuale strict biologice există şi comportament determinate biologic, puse în evidență, în cazul omului, de Gr.T. Popa [demonstrează cum creierul vechi (primitiv, reptilian, thalamus-hipo determină comportamente necontrolate cultural care duc mase de comportamente sălbatice, iar în cazul societăților mai avansate duc la de semicivilizație, în care impulsivitatea biologică devine colectivă şi sălb

Cultura

O subcategorie a unei categorii este o categorie. **Categoria cu** subcategorie a comportamentului socio-uman general, dar este a deosebeşte specia umană de toate celelalte specii animale, chiar d dintre acestea pot avea şi rudimente de cultură. Categoria cultură *comportamentul socio-uman cultural,* spre deosebire de cel biologic, c construieşte, dar nu se dezvoltă decât datorită, totuşi, anumitor biologice remarcabile ale omului, în special ale creierului său care a disponibilitate informațională. De aceea, dacă originea biologică a mentului cultural nu poate fi pusă la îndoială, cultura este o construc ridică mult deasupra biologicului, atât cât va putea față de limitele bi omului la un moment dat în istorie.

Poate că alături de cele două subcategorii menționate mai înain să mai adăugăm comportamentului socio-uman încă una, aceea a s (comportamentul spiritual), pe care nu o tratăm în această lucra spiritualitatea ca o a treia subcategorie a comportamentului socio-um este o chestiune care trebuie aprofundată, având în vedere că mulți cultură consideră spiritualitatea a fi un comportament numai cultu seama de cercetările şi studiile de filozofie a ştiinței din ultimii 15 ani pri

şi conştiința, vom considera, până la argumente contrarii convingătoare, spiritualitatea ca fiind o subategorie separată şi nu una înglobată (total) în cultură.

Schematic, vom rezuma cele de mai înainte, astfel:

CATEGORIA COMPORTAMEN-TULUI SOCIO-UMAN	**Subcategoria comportamentului strict biologic**	**Notă: există şi comportament social determinat strict biologic**
	Subcategoria comportamentului cultural.	CULTURA
	Subcategoria comportamentului spiritual.	SPIRITUALITATEA

Sferele mari ale culturii

Pornind de la definiția din [1b] şi diferența pe care o face UNESCO între cultura intangibilă şi cultura tangibilă, marile sfere (categorii) ale culturii pot fi considerate următoarele:

I. Cultura intangibilă. 'Moştenirea intangibilă poate fi definită ca îmbrățişând toate formele de cultură tradițională şi populară sau cultura folk, adică producțiile colective originate de o comunitate dată şi bazate pe tradiție. Aceste creații sunt transmise oral sau prin gesturi şi sunt modificate într-o perioadă de timp printr-un proces de re-creare colectivă. Ele includ tradițiile orale, obiceiurile, limbajele, muzica, dansul, ritualurile, festivitățile, medicina tradițională şi **farmacopeeai** artele culinare şi tot felul de îndemânări speciale legate de aspectele materiale ale culturii, cum sunt uneltele şi habitatul [6]\ Fără îndoială, noțiunea de cultură intangibilă a fost introdusă sub influența noțiunii de valoare intangibilă din economie care a căpătat o mare importanță pentru societatea cunoaşterii (economia bazată pe cunoaştere). Se mai adaugă aici valori, credințe, cunoaştere tacită.

II. Cultura umanistă. **Am** preluat în acest studiu denumirea tradițională. Cultura umanistă cuprinde limbajele naturale, literatura, arta, istoria, filosofia, sportul. Cultura umanistă este o cultură tangibilă, ca şi ştiința şi tehnologia.

III. Cultura ştiințifică: Ştiința, tehnologia şi cunoaşterea. Această categorie a culturii conține două subcategorii:

III.a Ştiința, cunoaşterea ştiințifică şi tehnologică, cunoaşterea tehnologică pentru fabricația de produse, dar şi pentru utilizarea acestora, precum şi cunoaşterea organizațională şi economică, chiar dacă unele obiecte ale cunoaşterii surit tacite sau fac parte şi din cultura intangibilă. în categoria mare a culturii, anumite obiecte pot aparține la două sau m subcategorii, acestea nu sunt neapărat disjuncte.

III.b **Uneltele fizice şi informaționale, obiectele fizice maționale produse sau fabricate, utilizarea lor, in şi organizațiile,** care sunt consecințe, în cea m măsură, a cunoaşterii ştiințifice, tehnologice, econ organizaționale, poate chiar şi a culturii intangibile.

Nu numai că unele obiecte culturale pot face parte din m subcategorii ale culturii, dar vor exista şi zone de interferență între ob acestor subcategorii. De exemplu, filosofia ştiinței, care este un obiect al nu se poate dezvolta decât în strânsă legătură cu ştiința. în teoria ca asemenea legături se numesc morfisme (morphisms sau maps, în limba Mai mult, pe lângă legăturile dintre obiectele subcategoriilor culturii, din or **a** culturii ar proveni, există relații între aceste sfere în totalitatea lor. Ac numesc functori. Cei mai importanți functori sunt aceia dintre catego categoria **III** de mai sus. Aceşti functori,

F1 : Categoria III (Cultura ştiințifică) + Categoria II (Cultura umani

F2 : Categoria II (Cultura umanistă) Categoria III (Cultura ştiințif

reprezintă relația şi influența reciprocă dintre, în esență, cultura un ştiință (cultura ştiințifică). Importanța lor pentru societate şi om nu subestimată.

Care este mai importantă dintre cele două categorii? Amb importante, dar motorul dezvoltării provine din sânul categoriei III. Aces devenit tot mai evident odată cu formularea conceptelor societății cunoaşt

Este adevărat că o altă resursă importantă este viața spirituală, şi componenta de creație implicând puternic atât cultura umanistă, cât ştiințifică.

Odată cu era informației vor apare desigur multe elemente noi a datorită tehnologiei informației, cărții şi documentelor electronice, i tehnologiilor vorbirii, tehnologiilor bioelectronice şi bioinformatice, artificiale şi agenților inteligenți informatici, mediului ambiant inteligent, conştiinței artificiale. Vor apare schimbări în viața intelectuală, socială şi

Ce se va mai petrece în cultură ?

în secolul XXI sunt posibile câteva evenimente majore care vor viața omenirii:

- Prăbuşirea ecologică a societății şi a speciei umane, deteriorării grave a mediului înconjurător, ceea ce s-ar putea la mijlocul sec. XXI (să spunem, anul 2050) dacă nu se trece

repectiv de pe acum, la efortul de asigurare a unor societăți sustenabile. Salvarea este posibilă chiar cu cunoaşterea ştiințifică şi tehnologică de astăzi dacă se trece la un management adecvat al cunoaşterii [7] şi la noi concepte economice adaptate sustenabilității. în această problemă au apărut şi alte noi perspective care vor rezulta dintr-o serie de evenimente descrise în continuare.

- Dezvoltarea inteligenței artificiale până la depăşirea inteligenței umane, ceea se va putea petrece între anii 2019-2035 sau chiar mai devreme [8], [9], [10], [11], [12].
- Apariția conştiinței artificiale, tot în cursul sec. XXI, după ce inteligența artificială vâ depăşi inteligența umană, dar fără a putea preciza perioada.

Aceste două ultime evenimente presupun apariția unor noi specii inteligente, dar şi noi specii conştiente, unele nebiologice (roboți umanoizi în topul unor *specii* de roboți mai puțin inteligenți care simulează animale (insecte, pisici, câini) şi roboți construiți pentru anumite funcțiuni care să înlocuiască omul [8][13].

Speciile de roboți umanoizi inteligenți şi de agenți software inteligenți, ambele egal de inteligente sau mai inteligente decât omul sunt uneori numite robo sapiens [13]. într-o primă etapă, aceste specii nu vor avea conştiință, astfel cum are omul, datorită faptului că au numai o organizare structurală şi nu una structural-fenomenologică [14]. Dar aceste specii vor interacționa puternic cu omul şi societatea şi se pune întrebare în ce măsură ele vor fi şi artefacte culturale, nu numai prin faptul că fac parte din cultura omului, ci şi prin participarea lor activă la cultură. Vor dezvolta cultura lor (într-o anumită măsură, da) sau vor intra în jocul marii culturi, participând la *cultura totală devenită din fenomen socio-uman, unul socio-uman- inteligență/conştiință artificială*

întrucât robo sapiens va avea cunoaştere şi va participa la dezvoltarea ştiinței şi tehnologiei, chiar la dezvoltarea sa ca obiect tehnologic, el va participa cu siguranță la cultura ştiințifică, poate chiar la anumite forme de cultură umanistă sau numai robotică. El poate fi implicat, prin cunoaşterea culturii umaniste, să participe ceva mai pronunțat la această cultură. Când va trece de la inteligență la conştiință, o asemenea activitate ar putea fi mult mai pronunțată.

Probabil, între homo sapiens şi robo sapiens vor exista relații de competiție şi cooperare, dar acestea se vor dezvolta într-o societate comună, cel puțin până la o segregare care nu ar fi de dorit, în care spiritualitatea şi creativitatea lui homo sapiens îi va conferi acestuia din urmă poziții inabordabile lui robo sapiens. Din momentul în care vor apare specii de *robo sapiens-conştient,* lucrurile se vor schimba din nou, cu efecte poate şi mai dramatice pentru om şi societate. încerc să mă conving că ideile unei societăți a conştiinței ar putea fi benefice pentru un asemenea viitor care probabil nu va putea fi prohibit. Probabil, înspre un asemenea viitor şi într-un asemenea viitor să fie rezolvată şi sustenabilitatea unei societăți a conştiinței.

Este interesant de reluat aici câteva previziuni ale lui Kurzweil [10] starea societății în anii 2019 şi 2029.

Pentru anul 2019, în domeniul afacerilor şi al economiei, prevede tra care în majoritate vor folosi persoane simulate, oamenii de afaceri vo asistenți software care vor conduce tranzacțiile în numele lor. Locuințe dispune de roboți de întreținere. Cu aceste artefacte comunicarea se va fa voce, deoarece vor dispune de o tehnologie a limbajului natural şi a vocii de înaltă calitate. Oamenii vor avea relații cu persoane automate intelige calitatea acestora de profesori, îngrijitori medicali, persoane de compar Aceste persoane automate au şi calități superioare omului în privința me dar, afirmă Kurzweil, 'ele nu sunt încă privite ca fiind egale cu oamenii î subtilitatea personalității acestora'. Inteligența artificială este însă prez împletită cu toate aspectele societății. Responsabilitatea omului va rămân pe primul plan şi nu a persoanelor (agenților) care îl ajută. Operele de artă realiza prin colaborarea dintre artişti umani şi inteligențe artificiale. Prin pericol în societate îl vor constitui micile grupuri de oameni şi inteligențe ar folosind comunicații criptate care nu pot fi descifrate. Acestea vor folosi informatice şi agenți de îmbolnăvire obținuți prin bioinginerie. Pe de alt descifrarea relațiilor dintre genele genomului uman va permite o medicină u inteligența artifcială pentru tratamentul şi chiar eradicarea multor boli, pentru prelungirea considerabilă a vieții omului natural.

Pentru anul 2029, Kurzweil prognozează: în domeniul comunicaț predomina, ca volum, comunicația dintre oameni şi maşini. Populația uman plafona la 12 miliarde de persoane reale, cărora li se asigură toate c normale de viață. Populația umană şi a inteligențelor artificaiale va fi preocu primul rînd, pentru crearea de cunoaştere, într-o puzderie de forme. Va fi fie menționate capacități ale omului care să nu fie preluate de maşini, de deosebire netă nu mai există între lumea oamenilor şi lumea maşinilor. C umană a fost transferată maşinilor şi multe maşini au personalitate, îndem baze de cunoaştere preluând şi cunoaşterea umană. Implanturile neurale c bazate pe inteligență artificială vor amplifica funcțiile cognitive ale omului. K afirmă: 'A defini ceea ce înseamnă o ființă umană devine o chestiune semn politică şi de legislație. Creşterea rapidă a posibilităților maşinilo controversată, dar nu există nici o rezistență față de ea. Deoarece la maşinile au fost proiectate pentru a fi supuse controlului uman, ele nu au pr o forță amenințătoare față de populația umană. Oamenii realizează că nu m posibilă dezangajarea civilizației devenită om-maşină de dependența de int maşinilor. Creşte discuția despre drepturile legale ale maşinilor, în spe acelor maşini care sunt independente de oameni (care nu sunt introduse creier uman). Cu toate că nu se recunoaşte deplin, prin lege, influența evi maşinilor la toate nivelele de decizie asigură o protecție importantă a maşini

Kurzweil consideră calități ale maşinilor inteligente, care încă din anul 2029 pot fi persoane de artă în toate domeniile artei ('Mulți dintre artiştii de frunte sunt maşini'). Observăm însă că acest lucru ar presupune o stare de conştiință similară omului şi prin manifestarea fenomenelor de qualia. Implicit, Kurzweil consideră că maşini inteligente complexe, structurale, pot avea asemenea stări şi pot chiar participa la discuții filosofice pe baza experienței proprii. Vorbind de experiența subiectivă a unor astfel de maşini, aceasta ar însemna că asemenea maşini să fi trecut pragul de la inteligență la conştiință *numai pe baze structurale* încă din anul 2029. Ceea ce nu credem, în principiu, a fi posibil.

într-adevăr, previziunile pe care oamenii de ştiință le fac privind dezvoltarea inteligenței artificiale spre conştiință artificială se bazează pe extrapolări ale ştiinței structurale (complexitatea structurală de la un anumit grad în sus generează conştiință, acest lucru fiind considerat valabil începând cu creierul animalelor). Odată cu creşterea complexității artefactelor creiere electronice sau creierelor software se consideră că atunci când acestea ating complexitatea creierului uman se va produce de la sine conştiința artificială [8], [10], [11]. Uneori, unii dintre cei care susțin un asemenea punct de vedere au îndoieli asupra valabilității lui [12]. în viziunea unei filosofii integrative a ştiinței [15],[16], conştiința nu se poate realiza numai din elemente structurale, fiind nevoie şi de elemente fenomenologice [17]. Conştiințele artificiale vor pune probleme foarte mari speciei umane care cred că ar putea fi rezolvate în cadrul unei viitoare societăți a conştiinței. Aceasta va urma atunci societății cunoaşterii în cadrul erei informației [18],'[19].

Ce va fi cultura în societatea conştiinței, la care vor participa, dacă nu chiar vor predomina conştiințele artificiale? Dacă lucrul cel mai important, în cele din urmă, este continuitatea conştiinței create de omenire, atunci şi culturii create de ea trebuie să i se asigure o continuitate.

Aceste considerații arată, dacă mai era nevoie de subliniat, cât de importante vor fi în sec. XXI cultura ştiințifică şi cultura umanistă, ambele având nevoie de o cultură filosofică adecvată.

Culturi, cultură pozitivă şi cultură negativă. Polarizarea culturii în jurul cunoaşterii

O cultură poate fi apreciată *pozitiv sau negativ,* în raport cu anumite criterii. Se pierde prea mult din vedere acest lucru. Există astăzi şi o cultură a teroriştilor (chiar şi o ştiință a terorismului) o cultură a corupției care ne pune nouă românilor atâtea probleme, o cultura a hoților etc. Desigur, acestea pot fi numite sub-culturi, dar tot culturi sunt. Cultura are multe fațete.

Cultura negativă este o cultură deformată în raport cu criteriile civ socio-umane.

în ultimii 12 ani, în societatea românească, pe lângă multe lucruri po s-au accentuat, din nefericire, şi fenomene negative îngrijorătoare: co imoralitate, injustiție. Creşterea imoralității şi a injustiției, a influențat până ş academică din țara noastră. Avem nevoie şi de un efort cultural pentru a aceste flageluri din societatea noastră, pe lângă efortul dezvoltării economice

Un exemplu de cultură pozitivă este arta. A cunoaşte arta îns cunoaştere dar a simți arta, a trăi arta, a avea nevoie de ea, a fi o b interioară, acestea înseamnă cultură umanistă adevărată.

Dar dacă cele de mai sus nu sunt însoțite de comportament civiliz civilizație socio-umană, cultura poate fi denaturată (rapturile de opere de scopuri personale sau statale). Natura firească a culturii pozitive este acee susține civilizația socio-umană, spiritualitatea, cunoaşterea şi conştiința, în c urmă societatea cunoaşterii şi societatea conştiinței.

în privința relației dintre cultura umanistă şi cultura ştiințifică, astăzi mai poate vorbi de cultură, cu înțelesul de cultură - în general, dar de gândul la cultura umanistă.

Cultura - în general, are o mult prea puternică componentă ş (inclusiv tehnologică, economică, organizațională, politică) pentru a mai acc asemenea simplificare, este adevărat, continuatoarea unei tradiții care astă complet depăşită. Cultura, respectiv cultura - în general, este cultura uma cultura ştiințifică, împreună, ultima având, ca şi prima, un conținut extrem de

în spatele confuziei care se menține astăzi atunci când vorbim de se întreține schisma dintre cele două culturi, datorită unor interese de g etapa actuală a societății, cultura umanistă nu-şi mai poate erija numele gen cultură, de fapt nu ea, ci slujitorii ei care nu s-au adaptat la vremurile cunoaş societatea cunoaşterii, înainte de trecerea la societatea conştiinței, cultura concentra în jurul cunoaşterii. lar tehnologia va fi un factor cultural a covârşitor încât va reveni poate la pozițiile ei mitologice din antichitate. remarcam:

în antichitate, la egipteni, zeul Ptah era privit ca patronul lucrăto metal (metalurgişti şi fierari) şi al artizanilor. Ptah era însă unul dintre cei m zei, creatorul pământului, părintele zeilor şi al începuturilor. Interesant ac începuturilor".

La grecii antici, echivalentul lui Ptah era Hefaistos, zeul foc meşteşugurilor, protectorul artizanilor. El nu mai era o divinitate primordia divinitate, fiul lui Zeus şi al Herei, fiind căsătorit cu Afrodita. Se pare ca egipteni la greci, tehnologia nu mai păstra poziția începuturilor, dar avea to

reprezentant divin. La romani, echivalentul lui Hefaistos era Vulcan, considerat zeul focului.

Decăderea poziției tehnologiei în cultură începuse din antichitate. Ea a continuat până în secolul XX când într-adevăr avea să cunoască un reviriment. Astăzi vorbim despre marile tehnologii şi chiar despre o filosofie a tehnologiei, de care o serie de gânditori şi filosofi au scris lucrări deosebit de interesante: Ernst Kapp, Friedrich Desauer, Jose Ortega y Grasset, Martin Heidegger ş.a. Este adevărat că au apărut şi lucrări îndreptate împotriva tehnologiei (L.Mumford, J.Ellul ş.a.), declanşând ceea ce în secolul XX s-a numit dilema tehnologică.'

Revirimentul filosofic al tehnologiei în societatea cunoaşterii, în secolul XXI, va fi un factor important în gândire, în general. Tehnologia va continua biologicul, culturalul şi conştiința.

Ce va face omul? Marea lui înțelepciune va fi aceea de a pregăti în mod corespunzător viitorul [19]. Din ce în ce mai mult, gândirea filosofică va avea un rol hotărâtor în ştiință, politică, viața socială.

Există şi vor exista culturi ale profesiilor, ale domeniilor cunoaşterii, ale națiunilor, etniilor, grupurilor, ale comunităților constituite pe Internet, ale instituțiilor şi localităților virtuale, ale maşinilor inteligente etc. Lumea devine tot mai pluriculturală. Probabil aceasta este trăsătura cea mai importantă a postmodernității [20].

Momentul actual ar trebui să fie acela al tendinței spre cunoaştere şi cultură (cu înțelesul ei total) pentru întreaga populație a omenirii, fiecare zonă locală, geografică sau virtuală, trebuind să fie preocupată activ de realizarea concretă a acestei tendințe.

Referințe bibliografice

[1] Mihai Drăgănescu: *Lucrări despre cultură:*

a. Mihai Drăgănescu, ***Cultura şi marile tehnologii,*** conferință, Universitatea Populară de Vară "Nicolae Iorga", Vălenii de Munte -30 august 1996.

b. Mihai Drăgănescu, ***Perspectiva informațională a culturii,*** Contemporanul, 27 mai 1983.

c. Mihai Drăgănescu, ***Dimensiunile europene ale culturii române,*** expunere, Vălenii de Munte, 1992, publicată în Academica, 1992.

d. Mihai Drăgănescu, ***Arta şi societatea,*** cuvânt, Ploieşti, 4 noiembrie 1991, publicat în Academica, 1991.

e. Mihai Drăgănescu, ***Criterii transpolitice si transmafiote în cu*** 18 mai 1997, Caiete Critice, 1997, nr.3-4, p.145-147.

f. Mihai Drăgănescu, ***Spirit enciclopedic şi enciclopedism,*** confe Vălenii de Munte, 22 august 1993 (publicată în Academica volumul autorului, Cariatidele gândului, Ed. Academiei Române, p. 163-168).

[2] Ernst Cassirer, ***Substanzbegriff und Funktionbegriff,*** 1910; **D/e *Philoso der Symbolischen Formen,*** 1923-1929 (3 voi).

[3] Oltea Miscol, Elena Gheorghe, ***Repere istorice în filosofia culturii,*** Revis filosofie, XLVII, Nr. 5-6, 2000, p.449-459.

[4] Mihai Drăgănescu, ***Categories and functors for the Structural Phenomenological Modeling,*** Proceedings of the Romanian Academy, Series A, Vol.1, No.2, 2000, p.111-115.

[5] Grigore T. Popa, ***Reforma*** **sp/r/fii/u/,** volum în editare, conținând lucrări ale acestui autor prezentate şi publicate la Academia Română în anii 1940 (v şi prefața: Mihai Drăgănescu, O *gândire asupra conştiinței, moralității şi societății).*

[6] UNESCO, definiția culturii intangibile, web-site UNESCO.

[7] Mihai Drăgănescu, ***Societatea Informațională şi a Cunoaşterii. Vectorii Societății Cunoaşterii,*** Academia Română, Bucureşti, 9 iulie 2001, publi în voi. coord. Florin Gh. Filip, *Societatea informațională-Societatea cunoaşterii. Concepte, soluții şi strategii pentru România,* Academia Rom 2002, p.43-112).

[8] Moravec H., ***Rise of the Robots,*** Scientific American, December 1999, p. 93.

[9] Moravec H., ***Robot Mere Machines to Transcendent Mind,*** Oxford Unive Press, Oxford, 1999.

[10] Kurzweil R., ***The Age of Spiritual Machines,*** Penguin Books, 2000.

[11] Broderick ***D.Jhe Spike,*** New York, 2002, paperback.

[12] Buttazzo G., ***Artificial Consciousness. Utopia or Reral Possibility?*** Computer (IEEE), July 2001, p.24-30.

[13] Interviews of Menzel P. and D'Aluisio F., ***Robo Sapiens. Evolution of a species,*** MIT Press, Cambridge, Massachusetts, 2002.

[14] Drăgănescu M., Din *lucrările despre minte şi conştiință:*

a. Mihai Drăgănescu, ***The Interdisciplinary Science Consciousness,*** Noetic Journal, Vol.3, No.1, Jan.2000, p. 3 republicat in eds. Richard L. Amoroso et al, *Science and the Pi of Consciousness, Intimation of a 21st Century Revolution,* Chap pp. 46-59, Orinda: The Noetic Press, 2000.

b. Mihai Drăgănescu, ***Theories of Brain, Mind and Conscious Still Great Divergences,*** Noetic Journal, vol.3, No. 2, Apr. p.125-139.

c. Mihai Drăgănescu, ***The Brain as an Information Processor,*** NOESIS, XXV, 2000, p. 9-20.

d. Mihai Drăgănescu, ***On the Structural-Phenomenological Theories of Consciousness*** , NOETIC JOURNAL, Vol.1, No.1, June 1997.

e. Mihai Drăgănescu, ***Continuities and Discontinuities in the realms of life and mind,*** Revue Roumaine de Philosophie, Tome 41,1997, Nos 1-2, p.3-9.

f. Mihai Drăgănescu, De ***la filosofia la ştiinţa mentalului,*** Revista română de filosofie, XLIV, Nr.5, sep-oct 1997, p. 457-464.

g. Mihai Drăgănescu, ***Procesarea mentală a informaţiei,*** Memoriile Sect. St. ale Acad. Române, SERIA IV, Tom. XX, 1997, p.263-284.

[15] Kafatos M., Drăgănescu M., ***Preliminaries to the Philosophy of Integrative Science,*** E-book (Microsoft Reader), ISBN 973-10-02510-X, Editura ICI, Bucharest, 2001.

[16] Drăgănescu M., Kafatos M., ***Generalized Foundational Principles in the Philosophy of Science,*** paper presented at the Conference on "Consciousness in Science and Philosophy" in Charleston, Illinois, 6-7 Nov 1998, published in The Noetic Journal, Vol.2, No.4, Oct. 1999, p. 341-350, republished in the voi. ***Science and the Primacy of Consciousness, Intimation ofa21*** ***Century Revolution,*** Richard L. Amoroso and others (eds), Orinda: The Noetic Press, 2000, Chapter 9, pp. 86-98.

[17] Mihai Drăgănescu, ***Advancement in Neural Engineering and Neuroelectronics Put Forward Artificial consciousness,*** Communication at the INGIMED II Conference, Bucharest, Dec. 13, 2001; E-PREPRINT, MSReader Format, 2002.

[18] Mihai Drăgănescu, ***Conştiinţa, frontieră a ştiinţei, frontieră a omenirii,*** Revista de Filosofie, XLVII, nr. 1-2, 2000, p.15-22.

[19] Mihai Drăgănescu, ***Societatea conştiinţei, o viitoare etapă a erei informaţiei. Vectorii societăţii conştiinţei,*** studiu, Academia Română, în pregătire.

[20] După Alain Fienckielkrant, apud [3], p.458-459.

între lingvistica matematică şi cea computațională

Solomon MARCUS
Secția de Ştiințe Matematice a Academiei Române
solomon.marcus@imar.ro

Mă simt obligat să reacționez la un anumit mod de prezentare a ev ideilor, în cea de a doua jumătate a secolului al XX-lea, în articolul [1] al d-lu Tufiş (de aici mai departe DT), membru corespondent al Academiei Ro Precizez de la început că nu contest interesul şi utilitatea direcției de preo prezentate în [1]; am în vedere numai modul în care această direcție este p relație cu alte cercetări dedicate limbajului.

Citäm din [1: 133]:

"Desprinzându-se din lingvistica formală, "lingvistica matemati încercat dezvoltarea unor modele matematice de reprezentare a limb naturale sau formale (în general al aspectului lor sintactic, gramatical), că soluții abstracte de modelare generativă de tip universal a ceea ce se presu (la nivelul cunoaşterii ştiințifice a anilor 1960) a fi facultatea limbajulur.

Nu ştiu ce înțelege DT prin "lingvistica formală", o sintagmă nu folosită în perioada de emergență a lingvisticii matematice; exista ling structurală (altceva decât ceea ce ar putea fi lingvistica formală, adică baz formalizare în sensul logicii matematice moderne), care desigur a constituit u sursele lingvisticii matematice (de aici mai departe LM), aşa cum i se pot in alte surse (biologice,, logice, matematice, psihologice etc), dar factorul deter în naşterea LM, în a doua jumătate a anilor '50, a fost dezvoltarea calculato electronice şi, împreună cu ea, a primelor preocupări sistematice d (prescurtare a lingvisticii computaționale), numite atunci traducere aut documentare automată, prelucrarea automată a limbajului, cu diverse varian lor în engleză (de exemplu, "machine translation"), franceză, rusă, ger italiană etc. Din aceste preocupări s-au inspirat primele modele care au co noua disciplină a LM.

Vorbesc despre lucruri trăite. Punctul meu de plecare s-a aflat în lu unor Kulagina şi Melciuk, puternic implicați în studiile de traducere automată franceză, Yves Lecerf, implicat în problemele de documentare automată, Hays, implicat în traducerea automată din rusă în engleză şi reciproc, B. Va

cu preocupări de informatică lingvistică la Grenoble. De la ei, ca şi de la alți autori similari, am preluat în bună măsură ştafeta pe care am căutat s-o duc mai departe. Ceea ce afirm despre mine este valabil pentru cei mai mulți cercetători din domeniul LM din anii 1950 şi 1960, cum ar fi Maurice Gross, Masami Ito, A. Trybulec şi mulți alții. Dubioasă mi se pare sintagma "soluții abstracte", probabil efectul unui obicei binecunoscut de a diaboliza abstractul.

în ceea ce priveşte sintagma "lingvistică formală", ea a căpătat o anumită utilizare în anii târzii 1960 şi în anii următori, iar personal am folosit-o în unele lucrări, după cum se va vedea imediat, considerând-o oarecum echivalentă cu LM; dar chiar dacă nu acceptăm această echivalență, nu putem eluda faptul că lingvistica formală se află în imediata vecinătate a LM. DT pretinde ca LM "a încercat", sugerând astfel că ea a eşuat în tentativă de modelare a limbajului natural. Ceea ce este deocamdată numai o sugestie devine, după cum se va vedea, o certitudine pentru DT.

într-adevăr, iată ce scrie mai departe DT ([1]: 133):

"Curând metodele lingvisticii matematice şi-au atins limitele drept care, în anul 1966, la propunerea lui David Hays, domeniul de cercetare al limbajelor naturale, din perspectiva utilizării acestora în interacțiunea cu calculatoarele electronice, este individualizat sub numele de lingvistică computațională".

Chestiunea cu atingerea limitelor ține de domeniul umorului involuntar şi trecem peste ea, dar nu ne miră, după ce am văzut la ce se reduce LM pentru DT. Nu mi-am imaginat niciodată că între LM şi LC ar putea avea loc o competiție, prima definindu-se prin metodă (căci ce altceva este LM decât studiul limbajului cu ajutorul matematicii ?) iar a doua prin obiectivul pe care şi-l propune. LM nu poate ignora problematica LC iar LC nu-şi poate realiza proiectele fără LM. Probabil însă că DT lucrează cu o definiție specială a LM, pe care am dori s-o aflăm. Modul simplificator în care DT se referă la generativismul lingvistic, într-o logică binară care eludează faptul că în materie de modelare se lucrează cu grade de adecvare şi relevanță, este însă simptomatic pentru viziunea sa limitativă în problema în discuție.

Crede DT că gramaticile lui Joshi, atât de importante în LC, puteau fi concepute fără să fi fost precedate de cele ale lui Chomsky ? Da, Chomsky a fost tot timpul foarte controversat, dar fără stimulentul său nu ştiu ce ne-am fi făcut, inclusiv în LC şi în LM, în ciuda faptului că el nu s-a prea referit explicit nici la LC, nici la LM. Faptul că gramaticile context free se află din nou, începând cu anii '80, în centrul atenției îrr LC nu spune ceva ? Iar faptul că aceleaşi gramatici (cu extensiunile lor) au marcat, încă din anii "60, teoria limbajelor de programare, domeniu în care ținta programării în limbaj natural se află în actualitate, nu este şi el semnificativ ? LC are mai multe părți, mai multe orientări, mai multe niveluri de abstracție, care comportă criterii diferite de evaluare. DT îl asociază pe D. Hays la ideea sa privind falimentul LM şi lansarea, drept consecință, a LC. Ca unul care a cunoscut bine cercetările lui Hays (a se vedea frecvența citărilor numelui să lucrările subsemnatului) şi l-a cunoscut şi personal foarte bine, fiind invitatul ca "plenary speaker" la Institutul de lingvistică al Americii (SUNY, Buffalo, 19 pot depune mărturie că acest autor vedea în LM şi LC două domenii solidare, c fețe ale aceleiaşi medalii, aşa cum se va vedea din citatul pe care-l vom da jos. Desigur, Hays a avut un rol important în anii de pionierat ai LM şi LC, ideea unei competiții între ele i-a fost străină. Voi evoca aici intervenția sa la de a treia Conferință Internațională de LC (COLING, September 1971): "The and scope of Computațional Linguistics" [2]. Citäm ([2]:p.23):

"Solomon Marcus says that formal linguistics is a pilot scie emphasizing at the same time that the ordinary field of linguistics is not. But th to say that linguistics as a branch of mathematics will supply methods to r fields of science, whereas linguistics as a descriptive field, a branch of na history or natural science, does not. [...] A four-way scheme can be arra with psychology, computation, formal linguistics, and descriptive linguistics a poles. Psychology and computation are about performance, formal and descr linguistics are about competence, computation and formal linguistics are abs and psychology and descriptive linguistics are sciences. But two other fields to find places in this scheme: psycholinguistics joins psychology with lingu and seems at this time a most fruitful field, one in which great progress ca made with benefit to both parent fields. Correspondingly, on the abstract COMPUTAȚIONAL LINGUISTICS JOINS COMPUTATION WITH FOR LINGUISTICS (subl. mea, S. M.) and also seems a fruitful area, one in RAPID PROGRESS CAN BE EXPECTED WITH BENEFIT TO BOTH PA FIELDS (subl. mea, S. M.) and with beneficiat application to psycholinguistics"

Referirea pe care o face Hays la subsemnatul are în vedere sloganu care l-am folosit de mai multe ori, "formal linguistics as a pilot science", sintagma "formal linguistics" era folosită ca un echivalent al LM. Iată deci că vedea în LC o alianță a LM cu computaționalul, alianță de natură să imprim progres rapid atât în LM cât şi în domeniul computațional. Cei 30 de ani scu atunci au confirmat-o pe deplin. Denumirile folosite pentru preocupări interferența limbajelor, informaticii şi matematicii au variat tot timpul şi nu cr acest aspect merită prea multă atenție. Lingvistica matematică? computați inginerească? algebrică? cognitivă? aplicată? cantitativă? teoretică? stati probleme matematice ale semioticii? tehnologia limbajului? limbajul în inteli artificială? inginerie lingvistică? procesarea limbajului natural? "information st retrieval"? lingvistica cibernetică? pe fiecare dintre acestea am întâlnit propriile mele articole au fost publicate aproape sub fiecare dintre etichete mai sus. Iată şi câteva detalii semnificative ale istoriei.

în 1962 s-a înființat în USA "Association of Computațional Linguistics

în 1963 Ferenc Kiefer a demarat la Budapesta revista "Computațional Linguistics", care a trăit peste zece ani. Conferința de la Grenoble de "traitement automatique des langues" din 1967 era a treia de acest fel, fiind precedată de o alta, la New York, în 1965 şi de una în Anglia, probabil în 1963, organizată de M. Masterman. între timp, la ruşi, numeroase conferințe au avut loc pe tema "avtomaticeskaja obrabotka tekstov" iar "Sprachkunde und Informationsverarbeitung" a fost uneori eticheta folosită de germani ş.a.m.d. Nu negăm rolul important pe care l-a avut David G. Hays în dezvoltarea CL, dar acest rol a fost altul decât cel afirmat de DT. Emergența LC s-a produs încă din anii '50, sintagma LC a devenit curentă încă de la începutul anilor '60. Şirul de conferințe COLING nu a făcut decât să continue această tradiție. Alții au preferat folosirea sintagmei LM (a se vedea, de exemplu, "Prague Bulletin of Mathematical Linguistics", "Prague Studies of Mathematical Linguistics", revista japoneză "Mathematical Linguistics" (în echivalentul ei japonez) etc. în ceea ce priveşte însă profilul acestor reviste, nu am constatat o diferență față de cele de CL. Desigur, între timp au început să apară şi unele publicații mai specializate, cu referire la părți determinate ale CL (cum ar fi cea relativă la corpusul lingvistic). Etichetele nu au avut importanță şi nu ştiu să se fi desfăşurat vreo competiție între ele. Chiar Hays a folosit diverse etichete, de exemplu cea din [3]. Dar DT merge mai departe pe ideea sa şi afirmă (în completă discordanță cu viziunea lui Hays, de la care se reclamă) că "metodele LM şi-au atins limitele" (încă în urmă cu peste 30 de ani!), pentru ca numai două pagini după această afirmație (deci la pagina 135 din [1]) să afirme că e nevoie de "modele formale ale limbii la toate nivelurile ei (fonetică, morfologie, sintaxă, discurs) gramatici formale [...]". Cum vede DT aceste modele formale altfel decât sub formă logico-matematică? Ştie oare că multe modele de acest fel există de câteva decenii? Indicații bibliografice asupra lor sunt date parțial în [4], [5], [6], [7] iar pentru cercetările româneşti în [8], [9]. Desigur, aceste modele sunt inegale ca valoare, au nevoie de continuări, modificări, ameliorări, dar ele nu pot fi ignorate. Fonetica, fonologia, vocabularul, morfologia, sintaxa, semantica lingvistică şi lingvistica istorică au beneficiat din plin de metodele matematice, aşa cum se poate vedea din impactul deosebit al lucrărilor respective în literatura de specialitate; DT indică, drept domeniu al LM, numai "aspectul sintactic, gramatical", despre celelalte nu a aflat. Nu a aflat nici că LM a abordat şi aspecte analitice, nu numai pe cele generative. DT defineşte "dimensiunea fundamentală" a LC prin "fezabilitatea instanțierii unei descrieri lingvistice cât mai complete, mentenabilitatea acestei instanțieri şi, desigur, conformanța cu realitatea uzului limbii". ([1]: 133). Cu un mic efort înțelegem despre ce este vorba. Desigur că problemele de complexitate, de cost, nu puteau fi încă abordate în anii '50 şi '60 cu mijloacele cu care ele au început a fi studiate în a doua jumătate a anilor 70, când instrumentele elaborate în informatica matematică deveniseră mult mai perfecționate. Dar acest fapt nu ține, cum crede DT, de alegerea între LM şi LC, ci de progresul general realizat în ştiință. Pentru a mă referi la propria noastră experiență, atunci când, în 1969, prezentam la COLING-ul din Suedia gramaticile contextuale nu aveam să mă ocup de aspectul complexității acestor gramatiGi în maniera în care am putut face acest lucru ulterior (a se vedea, de exemplu, [10]). Dar acest fapt nu are nici o legătură cu eticheta folosită.

Anii '80 şi '90 au confirmat necesitatea unui orizont cât mai larg în domeniul computațional. Nu m-am mirat atunci când "Encyclopedia of Microcomputers" şi "Encyclopedia of Computer Science and Technology" mi-au solicitat o contribuție cu tema "Semiotics and Formal Artificial Languages" (a se vedea [11]) şi nici când "Handbook of Formal Languages" mi-a solicitat un capitol privind "Contextual Grammars and Natural Languages" [12] iar o lucrare preponderent teoretică a fost inserată în "Computațional Linguistics in the Netherlands 2000"[13]. Nu m-am mirat nici când am văzut că o revistă de profil "Linguistics and Philosophy" publică articole excelente de LC. Interferențele sunt în toate direcțiile şi ele caracterizează cultura contemporană. în acest context trebuie să ne plasăm, cred, atunci când ne referim la disciplinele cognitive care se dezvoltă sub ochii noştri şi îşi pun amprenta pe modul nostru de gândire şi de comportare. Un tratat ca "Mathematical Methods in Linguistics" [14] include multe fapte de LC, deşi în titlul său nu figurează epitetul "computațional". O revistă ca "Theoretical Linguistics" (1970-2000), publicată de Walter de Gruyter (Berlin - New York) a inclus multe articole vizând aspecte matematice şi/sau computaționale, deşi numele revistei nu indică acest lucru. Chiar o revistă mai tradițională ca "Linguistics" a inclus de multe ori articole de LM şi nici "Foundations of Language" nu a procedat altfel. Multe fapte de LM şi de LC se plasează în mod natural pe orizontul semioticii computaționale. Era internetului impune desigur o problematică nouă, față de care abordările anterioare se pot dovedi insuficiente. Să fie inițiativa noii generații de cercetători de a se dedica noilor probleme. Dar trecerea de la ieri la azi şi de la azi la mâine nu poate fi decât una care ține seama în mod critic de experiența acumulată. Din tot ceea ce am prezentat mai sus rezultă că LM şi LC au fost mereu împreună şi că, în general, etichetele nu au contat prea mult. Unii au mers chiar mai departe; astfel, în capitolul 4, "Mathematical Computațional Linguistics", din [15], se afirmă pur şi simplu (p.86): "Mathematical linguistics has also been called theoretical linguistics and even computational linguistics". Iar mai departe, în acelaşi loc: "Computațional Linguistics originated around 1950 with the initiation of research on automatic translation" (se trimite la o carte editată de D.G.Hays [3] şi la o alta avându-l ca autor pe acesta [16]).

Ca unul care crede în legătura naturală a lingvisticii cu matematica, am încercat o deosebită satisfacție să trăiesc momentul în care această legătură este acceptată de ambii parteneri şi că de multe ori nici nu mai e nevoie de acel artificiu retoric al epitetului "matematică"; LM este acceptată pur şi simplu ca lingvistică. Suntem convinşi că o traiectorie similară o urmează şi LC iar unele semne în această privință există de pe acum, aşa cum am arătat mai sus.

LC este de mai mulți ani o secțiune la congresele internaționale de lingvistică iar LM şi LC au secțiunea lor în reviste internaționale de referate ca "Language and Language Behavior Abstracts". în România, minți luminate ale anilor '60, ca profesorii Al. Rosetti, Grigore Moisil şi Tudor Vianu, au înțeles schimbările care se profilau şi au sprijinit proiectul înființării unei secțiuni de "lingvistică aplicată" la Facultatea de Limbă şi Literatură Română a Universității din Bucureşti, dar s-au găsit alții care să-l torpileze.

La Academia Română a funcționat mulți ani "Comisia de Lingvistică Matematică" iar revista "Cahiers de Linguistique Theorique et Appliquee", înființată în 1962, a fost multă vreme expresia colaborării lingvisticii cu matematica şi cu informatica. în ciuda forțelor adverse, s-a reuşit în acei ani atragerea unor studenți străluciți ai unor facultăți umaniste la cercetarea limbii şi literaturii cu mijloacele LM şi LC. Pentru a da numai două exemple de actuali profesori universitari care au susținut teze de doctorat de acest tip, voi menționa pe Pia Brinzeu, de la Catedra de Engleză a Universității din Timişoara şi pe Mihai Dinu, de la Facultatea de Litere a Universității din Bucureşti. Tot în acea perioadă şi-a susținut teza de doctorat Sorin Cristian Niță, pe o tema de critică textuală automată privind înlănțuirea (filiația) diferitelor variante ale "Istoriei Țării Româneşti" (Şerban Cantacuzino).

Iată însă că, în pofida realităților puse în evidență mai sus, în ([1]: 134) se scrie: "în România, cercetările în domeniul LC şi al prelucrării limbajului natural, precum şi primele rezultate practice au apărut la începutul anilor '80 [3, 4, 5, 6]".

La ce trimit numerele indicate în paranteze? La o bibliografie de 24 de titluri în care aproape toate (dar toate cele indicate între paranteze) încep cu DT (ignorându-se regula generală în lumea ştiințifică, a aşezării numelor autorilor aceluiaşi articol în ordine alfabetică; dar nu acest fapt este cel care ne interesează în momentul de față). Să observăm că încă În1978, în articolul "Mathematical and Computațional Linguistics" [9] de prezentare a activității din România în domeniul LM şi LC se face referire la peste 400 de articole publicate de 130 de autori români şi sunt menționați peste 300 de autori străini (unii dintre ei, nume de vază ale LM şi LC din acea perioadă) care au citat şi continuat cercetările româneşti. Să mâi adăugăm că numeroşi lingvişti români dintre cei mai importanți au citat şi folosit rezultatele şcolii româneşti de LM şi LC. Iată că vine acum DT şi face (deliberat sau nu) din tot acest efort un teren viran care-l aştepta pe DT să tragă primele jaloane. Nu e cam mult?

Să fim bine înțeleşi. Nu noi avem nevoie de încă o citare pe lângă miile de citări deja acumulate, ci noile generații de studenți şi de cercetători au dreptul la o informare corectă asupra dezvoltării LM şi LC în general şi, în particular, asupra LM şi LC în România. DT a mai publicat, în urmă cu câțiva ani, un articol în care se schița o privire istorică asupra LC în România, cu câteva citări la întâmplare, care trădau necunoaşterea situației reale.

Mai este un aspect care cere o precizare. în conformitate cu spe volumului în care apare articolul [1], DT face numeroase referiri la a documente ale unor organisme europene şi internaționale, cum este şi pentru a nu mai vorbi de aspectul financiar al colaborării cu organ respective. Această situație a existat de la începutul LM şi LC (chiar dacă avut amploarea de azi), datorită faptului că LM şi LC au apărut şi ca urmare comandamente sociale, privind precaritatea mijloacelor de prelucrare a infor îmi amintesc de faimoasele Rapoarte CETIS care veneau de la EURA Bruxelles, pe teme legate de analiză şi prelucrarea automată a limb traducere automată şi documentare automată. în USA, diferite corporații (cu RAND Corporation, Santa Monica, Calif.) finanțau cercetări similare. O în semnificativă a fost aceea din 1962, organizată de "NATO Advanced Su Institute", la Veneția, Italia, privind traducerea automată. De numele acestui este legat un document care a marcat evoluția cercetărilor de traducere aut seria de expuneri prezentate de Y. Bar-Hillel [17]. în legătură cu aceste a dirijate şi finanțate de diferite organisme europene şi internaționale, treb observăm că cei implicați au avut înțelepciunea şi priceperea necesare pe nu reduce proiectele respective la dimensiunea lor exclusiv utilitară, ci c subordona pe aceasta unei perspective mai ample, care lua în consi orizontul ştiințific real al problemelor. Pentru a da un prim exemplu, mă voi r faptul că mai multe rapoarte CETIS au pus în discuție un concept care, năs experimentele de traducere automată, avea să se dovedească de o dec semnificație pentru teoria sintactică în toată generalitatea sa; este vor conceptul de proiectivitate sintactică, cu consecințe bogate în studiul stru arborescente şi al gramaticilor de dependență. Azi putem spune că şi s limbajului natural şi teoria matematică a grafurilor au profitat esențial de cor respectiv (folosit până şi de Rene Thom, în probleme de morfogeneză Această expansiune a unui concept sau rezultat dincolo de motivația sa este testul cel mai convingător al interesului său. Un al doilea exemplu se re titlul provocator folosit de Bar-Hillel pentru expunerile sale: "Patru con despre lingvistica algebrică şi traducerea automată".

Simpla alăturare a celor două sintagme, una foarte teoretică, c aparent tehnologică, avea menirea să-i avertizeze pe cei care presau să se cât mai repede rezultate practice asupra faptului că proiectele de tra automată nu se pot finaliza de azi pe mâine, ci au nevoie de un lung lingvistic, matematic şi computațional. Acum ştim că acest itinerar continuă cu tatonări şi reveniri, şi, chiar dacă nu a dus încă la rezultatele vis impulsionat în mod esențial cercetările de Al, cu consecințe benefice aspectele logice şi semantice ale limbajului natural.

întrebarea pe care ne-o punem, dar o lăsăm deocamdată fără ră deoarece nu suntem pregătiți pentru a-l da, este următoarea: Nu cumva asp pe care le-am criticat mai sus sunt consecința unui fenomen mai general, a

unui orizont insuficient de cuprinzător, al unei prea mari dependențe de factori utilitari imediați? Ştiința a oscilat mereu între cognitiv şi utilitar, dar istoria arăta că funcția utilitară s-a manifestat în toată profunzimea ei atunci când ea a fost fructul unei evoluții fireşti a funcției cognitive, evoluție care poate fi de doi ani, de 20 de ani, de 200 sau de 2000 de ani. Cu un ochi îndreptat spre comisiile europene, suntem obligați totuşi să ținem treaz şi celălalt ochi, îndreptat spre ceea ce se întâmplă pe scena cercetării ştiințifice vii, aşa cum apare ea în revistele de specialitate şi la întâlnirile ştiințifice de profil. Istoria generală a ştiinței şi, în particular, scurta istorie a LM şi LC, sunt pline de învățăminte în această privință.

Referințe bibliografice:

[I] D. Tufiş. *Promovarea limbii române în SI-SC.* în *Societatea Informațională - Societatea cunoaşterii* (coord. F. Gh. Filip). Ed. Expert, Bucureşti, 2001, 131-142.

[2] D. G. Hays. *The field and scope of computațional linguistics.* Papers in Computațional Linguistics (eds. F. Papp, G. Szepe). Proceedings of the Third International Meeting of Computațional Linguistics, held in Debrecen, Hungary, 1971. Akademiai Kiado, Budapest, 1976, 21-26.

[3] D. G. Hays (ed.). *Readings in Automatic Language Processing,* American Elsevier, New York, 1967.

[4] S. Marcus. *Mathematical Linguistics in Europe. Current Trends in Linguistics* (Th. A. Sebeok, ed.), vol.9, Mouton, The Hague, 1972, 646-687.

[5] S. Marcus. *Mathematique et Linguistique.* în *Mathematique, Informatique et Sciences Humaines,* Paris, 26,1988, 103, 7-21.

[6] S. Marcus. *The status of research in the field of analytical algebraic models of language.* în *Current Issues in Mathematical Linguistics* (C. Martin-Vide, ed.). Elsevier - North Holland, Amsterdam, 1994, 3-21.

[7] S. Marcus. *Lingvistica matematică, azi.* în *Matematica în lumea de azi şi de mâine* (C. Iacob, coord.), Editura Academiei, Bucureşti, 1985, 182-186.

[8] S. Marcus. *Recent Romanian investigations in the field of mathematical and computațional linguistics.* Avtomaticeskaja Obrabotka Tekstov, Matern. Fyz. Fakulta, KL Praha, 1973,15-42.

[9] S. Marcus. *Mathematical and computațional linguistics.* în *Current Trends in Romanian Linguistics* (A. Rosetti, S. Golopentia Eretescu, eds.). Revue Roumaine de Linguistique 23, 1978, 1-4, 559-588.

[10] S. Marcus, C. Martin-Vide, G. Paun. *Contextual grammars as generative models of natural languages.* Computațional Linguistics 24,1998, 2, 245-274.

[II] S. Marcus. *Semiotics and formal artificial languages.* în *Encyclopedia of Computer Science and Technology* (A. Kent, J.C.Williams, eds.) 29, Ed. Marcel Dekker, New York, 1994, 393-405; also in Encyclopedia of Microcomputers (A. Kent, J.C.Williams, eds.) 15, 1995, 299-312.

[12] S. Marcus. *Contextual grammars and natural languages, Handbook of F Languages* (G. Rozenberg, A. Salomaa, eds.), 2, Springer, Berlin, New 1997,215-235.

[13] S. Marcus, C. Martin-Vide, G. Păun. *A new-old class of linguis motivated regulated grammars. Computațional Linguistics in the Nethe 2000* (W. Daelemans et al., eds.), Selected Papers from the Eleventh Meeting, Ed. Rodopi, Amsterdam, New York, 2001, 111-125.

[14] B. H. Partee, A. Ter Meulen, R. Wall. *Mathematical Methods in Lingu* Kluwer, Dordrecht et al, 1990.

[15] E. F. Beckenbach, Ch. B. Tompkins (eds.). *Concepts of Communi Interpersonal, Intrapersonal and Mathematical.* John Wiley and Sons, York, 1976.

[16] D. G. Hays. *Introduction to Computațional Linguistics.* American Elsevier, York, 1967.

[17] R. Thom. *Stabilite Structurelle et Morphogenese.* John Benjamins, New 1970.

[18] Y. Bar-Hillel. *Four Lectures on Algebraic Linguistics and Machine Tran* revised version of a series of lectures given in July 1962, before a N Advanced Summer Institute, Venezia, Italy.

între lingvistica matematică şi cea computațională: o altă perspectivă

Dan TUFIŞ
Institutul de Cercetări pentru Inteligență Artificială,
Str. 13 septembrie, nr. 13, 74311, sector 5, Bucureşti
tufis@racai.ro

1. în loc de introducere

Dat fiind că acest articol este un comentariu asupra filipicei de neînțeles "între lingvistica matematică şi cea computațională" a domnului Solomon Marcus, membru titular al Academiei Române, mărturisesc că elaborarea sa fost o întreprindere asupra căreia am avut multe ezitări iscate din incertitudinea receptării sale corecte, constructive. Din păcate majoritatea afirmațiilor şi implicațiilor pe care domnia sa le face în articolul amintit, sunt inexacte şi umorale. Nu mai insist şi asupra decontextualizării citatelor din lucrarea mea [1], procedeu neelegant. Este binecunoscut din logica clasică faptul că dintr-o serie de premise false se poate demonstra orice. în ciuda ezitărilor amintite, violenta polemică lansată de domnul Solomon Marcus prin articolul menționat îmi oferă posibilitatea de a aduce în discuție elemente de istorie a domeniului care ar putea fi de interes, cu precădere pentru cititorii al căror domeniu de specialitate nu este prelucrarea automată a limbajului natural. Pentru specialiştii în domeniul prelucrării limbajului natural, majoritatea argumentelor pe care le voi aduce sunt bine cunoscute.

Ca modalitate de documentare, am optat pentru includerea integrală a materialului produs de domnul Academician Marcus, indentat şi redat cu caractere italice. De asemenea, am păstrat secțiunea domniei sale de referințe bibliografice. Lucrările pe care le-am citat eu sunt documentate în cuprinsul textului, prin includerea referinței complete între paranteze rotunde. Singura excepție este lucrarea mea, sursa nemulțumirii domnului Marcus, care este referită de amândoi ca [1]. Cititorul va putea face astfel mai uşor distincția între cele două categorii de referințe. înainte de a proceda la analiza afirmațiilor domnului Academician Marcus, aş dori să fac unele precizări:

- contextul discuției în [1], ca şi în cele ce urmează, este cel al tehnologiei limbajului, al cercetărilor foarte intense în întreaga lume

pentru dezvoltarea de sisteme inteligente capabile să faciliteze comunicarea dintre doi sau mai mulți conlocutori (oameni sau sisteme software), prin intermediul limbajului natural;

în raport cu lucrarea [1] domnul Academician Marcus se opreşte cu îndârjire asupra a doar trei fraze interpretate ca atac la persoana sau activitatea sa ştiințifică şi se referă ironic (şi după cum se va vedea în continuare, în mod nejustificat) la alte două, făcând abstracție de restul prezentării care nu are nici o contingență cu domnul Marcus. Domnul Academician are merite pe care nu i le poate lua nimeni, are contribuții importante în mai multe domenii şi este creatorul şcolii româneşti de lingvistică matematică. Interesul domniei sale pentru aspectele legate de implementarea pe calculator a programelor de prelucrare a limbajului natural a fost minim. îmi reamintesc o discuție pe care am avut-o în anul 1991 la câtva timp după ce mă întorsesem de la Conferința Europeană de Lingvistică Computațională organizată la Berlin de profesorul Jurgen Kunze. Cu acea ocazie, domnul Academician Marcus mi-a mărturisit că îl cunoaşte de multă vreme pe profesorul Kunze şi că au şi colaborat o perioadă cât amândoi au avut ca domeniu de preocupări lingvistica matematică. La sfîrşitul anilor '60, mai spunea domnul Marcus atunci, drumurile celor doi s-au despărțit, profesorul Kunze optând pentru noua paradigmă a lingvisticii computaționale.

Domnul Academician Marcus a scris enorm, în domenii extrem de variate, aici mă refer în special la cele legate de studiul limbii, şi prin urmare era inevitabil să nu atingă subiectul foarte actual al prelucrării automate a limbajului natural. A făcut-o însă detaşat de nivelul inerent perisabil al tehnologiei informatice. O teorie ştiințifică, un model formal teoretic sau transpus într-o implementare a unui program software sunt inevitabil supuse „eroziunii" timpului, unele mai rapid altele mai lent. Lucrarea [1], despre care discutăm, ia în discuție exact acest cadru al investigației tehnologice şi a măsurilor ştiințifice, tehnice, organizatorice şi chiar legislative pentru a crea o bază perenă a cercetării şi dezvoltării tehnologice privind prelucrarea automată a limbii noastre: resursele computaționale fundamentale ale limbii române. Societatea Informațională-Societatea Cunoşterii este caracterizată de vectori tehnologici şi funcționali [M. Drăgănescu: *„Societatea informațională-societatea cunoaşterii. Vectorii societății cunoaşteri^* In *Societatea Informațională - Societatea cunoaşterii* (coord. F. Gh. Filip). Ed. Expert, Bucureşti, 2001, 43-112.] a căror ignorare este nu numai neproductivă dar şi periculoasă, „în era electronică, ***este esențial pentru supraviețuirea unei limbi ca ea să fie folosită în sistemele de informare electronică"*** afirmă fără echivoc Alain Danzin în influentul raport al Comisiei Europene „Towards a European Language Infrastructure" întocmit în 1992 prin consultarea a 182 de specialişti din cercetare şi industrie. Promovarea limbii române în contextul informațional al societății cunoşterii este un obiectiv actual şi de viitor şi nu poate fi subiect de dispută în viața ştiințifică românească;

- deşi este un truism, cred că pentru evitarea unor interpretări greşite este necesar să subliniez faptul că în dezvoltarea programelor de inteligență artificială, de prelucrare a limbajului natural sau în general în ingineria software, o mulțime de discipline matematice (teoria algoritmilor, teoria complexității, teoria limbajelor formale, teoria categoriilor, statistica matematică şi multe, multe altele) sunt fundamente indispensabile în avansul ştiințific şi tehnologic al acestor discipline (şi desigur nu numai al lor). Programarea (ca şi matematica elementară) sau utilizarea de produse informatice sunt activități la îndemâna tuturor (de altfel reflectate şi în programele şcolare de învățământ), dar proiectarea şi realizarea de programe software inteligente necesită o pregătire teoretică solidă, talent şi multă muncă. Diferența între două programe care produc aceleaşi rezultate dar unul în câteva secunde şi altul în câteva ore, apare tocmai din diferența de pregătire teoretică şi talent a autorilor lor.

- domeniul ştiinței şi tehnologiei informației este poate cel mai dinamic sector al activității creative: Biil Gates spunea că dacă de pildă industria automobilelor ar fi avut aceeaşi dinamică cu cea a calculatoarelor, acum o maşină ar trebui să coste 1 dolar. Fantasticul ritm de dezvoltare al tehnologiei hardware (bazată pe importante descoperiri ştiințifice obținute în ultimii 50 de ani) nu a fost nici pe departe egalat de ritmul dezvoltării în domeniul software. în ciuda acestui decalaj, ştiința ingineriei software si-a reînnoit instrumentarul teoretic (modele şi/sau formalisme) cu o viteză neîntâlnită în alte domenii ştiințifice. Dinamica fără precedent a cunoaşterii în ştiința şi tehnologia informației obligă omul de ştiință din acest domeniu la o informare continuă, din ce în ce mai specializată şi mai selectivă. Se estimează că în acest domeniu se scriu în fiecare zi mai multe articole decât poate citi un om în întreaga sa activitate şi că informația mai veche de 15-20 ani este foarte probabil să fie perimată (desigur cu excepțiile ce întotdeaurta confirmă regula). Evoluția terminologică în acest domeniu este încă o mărturie vie a dinamicii de care aminteam: în domeniul prelucrării limbajului natural se vorbeşte acum de ontologii lexicale, de gramatici lexicalizate susținute de ontologii, de analiză (parsing) ontologică, de lingvistica WEB-ului şi WEB-ul semantic, de resurse lingvistice standardizate şi aşa mai departe.

- referitor la antinomia „lingvistică matematică-lingvistică computațională" pe care domnul Academician Marcus mi-o atribuie, vreau să precizez că nicicând nu am afirmat că cele două domenii se exclud reciproc sau că

ar fi în competiție; pur şi simplu ele sunt subsecvente din punctul de vedere al relevanței față de problemele pe care le discutăm aici. Există fără îndoială o filiație între ele, în sensul că lingvistica computațională a preluat o mare parte din instrumentarul lingvisticii matematice (nici nu se putea altfel) dar ce a adus nou lingvistica computațională, pe lângă noi modele şi formalisme, este în primul rând de natură metodologică şi tehnologică: experimentul şi evaluarea. Ceea ce se numeşte astăzi lingvistică computațională teoretică este în mare măsură asimilată cu lingvistica formală modernă. Acest segment al lingvisticii computaționale a moştenit de la lingvistica matematică cel mai mult şi adecvându-şi metodele la realitățile tehnologice a produs şi este de aşteptat să producă noi rezultate validabile şi incorporabile în sisteme automate de prelucrare a limbajului natural. Teoriile şi formalismele lingvistice, azi în vogă în lingvistica computațională (TAG, LFG, HPSG, CG, CUG), au fost produse de lingvistica formală şi prin validarea instanțierilor pe segmente de limbă netriviale, au devenit instrumente operaționale ale prelucrării limbajului natural. Dezvoltarea de modele de limbă, analiza algoritmilor de prelucrare a limbajului (resursele de calcul necesare unei implementări funcționale, viteza de răspuns), construcția (achiziția) resurselor lingvistice standardizate, gradul de acoperire lingvistică al unei formalizări lingvistice (cunoştințe lingvistice=resurse lingvistice), sunt doar câteva direcții definitorii ale metodologiei lingvisticii computaționale, în sfârşit, în raport cu obiectivele finale urmărite de implementarea unui model de prelucrare a limbajului se remarcă în ultimii circa 10 ani o departajare şi chiar o competiție (fără însă a fi o antinomie) între abordările introspective-principiale şi cele inductive, bazate pe date. Prima categorie de abordări este caracterizată de dezvoltarea prin introspecție ştiințifică de teorii şi formalisme gramaticale computaționale (imensa lor majoritate bazate pe restricții şi unificare categorială cu accentuată lexicalizare) şi mai apoi instanțiate manual de experți lingvişti. Cea de a doua abordare, ce câştigă foarte mult teren în ultima perioadă, este cea bazată pe tehnicile învățării automate ce pornesc de la premiza că, într-un corpus lingvistic reprezentativ şi de dimensiuni mari, există suficientă informație privind regularitățile dintr-o limbă (cea în care sunt textele ce alcătuiesc corpusul lingvistic) astfel încât, tehnici adecvate de învățare automată să fie capabile să construiască un model de limbă robust şi de mare acoperire lingvistică. Aş mai menționa că, în fapt, de multe ori cele două abordări sunt combinate (cu preponderența uneia dintre ele). într-un anumit sens, acest dualism în abordările modelelor de prelucrare automată a limbajului natural continuă o celebră confruntare de idei între Chomsky şi Piaget susținătorii teoriilor înăscutului (innate) şi respectiv al învățării în explicarea facultății umane a limbajului.

Cu aceste lămuriri preliminare, voi analiza în continuare afirmațiile domnului Academician Marcus cu sincera speranță că cititorii acestui text, dar mai ales domnia sa, vor înțelege că preocupările mele şi ale distinsului profesor au alte obiective, motivații şi desigur modalități foarte diferite de finalizare. Acest lucru nu înseamnă că rezultatele fiecăruia dintre noi le anulează sau le diminuează pe ale celuilalt (cu atât mai mult cu cât recunoaştere internațională există pentru amândoi). După cum la fel de bine diferențele de perspectivă şi opinii, naturale în fond, nu înseamnă că nu avem a ne spune lucruri interesante unul altuia.

2. O analiză textuală

> *„Mă simt obligat să reacționez la un anumit mod de prezentare a evoluției ideilor, în cea de a doua jumătate a secolului al XX-lea, în articolul [1] al d-lui Dan Tufiş (de aici mai departe DT), membru corespondent al Academiei Romane. Precizez de la început ca nu contest interesul şi utilitatea direcției de preocupări prezentate în [1]; am în vedere numai modul în care aceasta direcție este pusă in relație cu alte cercetări dedicate limbajului."*

Aşa îşi începe domnul Academician Marcus articolul solicitat de mine pentru volumul „Limba Română în Societatea Informațională-Societatea Cunoaşterii" rezultat al proiectului INFOSOC „SI-SC: Soluții şi strategii în România". Să urmărim un prim citat incriminat (care în transcrierea dlui Academician este trunchiat si conține nişte ghilimele ce nu-mi aparțin; redau mai jos varianta publicată):

[1: p.133]:

"Din acest punct de vedere *(al folosirii calculatorului în prelucrarea limbajului natural - precizarea mea, DT),* este semnificativ a arăta că însuşi numele domeniului de cercetare a prelucrării automate a limbajului natural a suferit modificări reflectând progresele ştiințifice şi tehnologice: inițial, desprinzându-se din lingvistica formală, *lingvistica matematică* a încercat dezvoltarea unor modele matematice de reprezentare a limbajelor naturale sau formale (în general al aspectului lor sintactic, gramatical), căutând soluții abstracte de modelare generativă de tip universal a ceea ce se presupunea (la nivelul cunoaşterii ştiințifice a anilor 1960) a fi facultatea limbajului."

Ce l-a supărat aici pe distinsul polemist? Ne spune chiar domnia sa:

> *„Nu ştiu ce înțelege DT prin "lingvistica formală", o sintagmă nu prea folosită în perioada de emergență a lingvisticii matematice; există lingvistica structurală (altceva decât ceea ce ar putea fi lingvistica formală, adică bazată pe formalizare în sensul logicii matematice moderne), care desigur a constituit una din sursele lingvisticii matematice (de aici mai departe LM), aşa cum i se pot indica şi alte surse (biologice, logice, matematice, psihologice etc.)*

Mă surprinde întrebarea retorică cu care începe „argumentația", şi căreia nu-i văd decât un gratuit rol derogativ. Eu nu-mi închipui că domnia sa nu a auzit de antinomia „gramatică descriptivă - gramatică formală" la limitele extreme ea fiind reprezentată de lucrările lui O. Jespersen (O. Jespersen: *The philosophy of Grammar,* Allen & Unwin, London,1924 şi *Analytical Syntax.* Hoit Rinehart & Winston, New York, 1937 (republicată în 1969)) şi respectiv lucrările timpurii ale lui Chomsky referitoare la lingvistica generativă. Dacă însă mă înşel, o lectură lămuritoare, este influenta carte editată de Keith Brown şi Jim Miller în Pergamon Press, 1996 numită „Concise Encyclopedia of Syntactic Theories", cu precădere articolul „Descriptive Grammar and Formal Grammar" de F. Stuurman, al cărui prim capitol se numeşte chiar Descriptive and Formal Grammar: The Fundamental Opposition. La fel de utilă este şi lucrarea monumentală a lui David Crystal „The Cambridge Encyclopedia of Language", Cambridge University Press, 1987.

Pe de altă parte, o pagină mai încolo, domnul Academician mărturiseşte că şi domnia sa a folosit termenul de lingvistică formală:

> *în ceea ce priveşte sintagma "lingvistică formală", ea a căpătat o anumită utilizare în anii târzii 1960 şi în anii următori, iar personal am folosit-o în unele lucrări, după cum se va vedea imediat, considerând-o oarecum echivalentă cu LM* (lingvistica matematică);

Pentru lămurirea elementului istoric, furnizez în continuare un citat din recenzia lui R.B. Lees (Language, nr. 33, voi 3, '1957, pp. 375-408) la faimoasa carte a lui Chomsky (Syntactic Structures, Mouton, The Hague, 1957): „in a sense, transformational analysis is essentially a **formalization** of a iong-accepted, tradiţional approach...". Citatul apare la pagina 387. Chomsky se pare că *ă* apreciat termenul şi i-a adoptat, cel puţin în raport cu propria filozofie generativistă asupra limbajului.

> *„dar factorul determinant în naşterea LM, în a doua jumătate a anilor "50, a fost dezvoltarea calculatoarelor electronice şi, împreună cu ea, a primelor preocupări sistematice de LC (prescurtare a lingvisticii computaţionale), numite atunci traducere automată, documentare automată, prelucrarea automată a limbajului, cu diverse variante ale lor in engleza (de exemplu, "machine translation"), franceză, rusă, germană, italiană etc. Din aceste preocupări s-au inspirat primele modele care au constituit noua disciplină a LM."*

înainte de a face o serie de precizări istorice mai exacte, vreau să notez că de la începutul istoriei sale, domeniul traducerii automate a fost, şi în mare măsură a şi rămas, un domeniu distict de restul preocupărilor legate de prelucrarea limbajului natural. Aş mai observa că textul de mai sus, încearcă să sugereze că LM s-ar fi constituit ca disciplină ulterior LC. Ambiguitatea afirmaţiei de mai sus provine din punerea în relaţie de concordanţă temporală a primelor preocupări în domeniul LC cu apariţia domeniului în sine. Oricine ştie că un anumit domeniu ştiinţific se cristalizează în timp, pe baza unor rezultate ştiinţifice promiţătoare, a unor experimente convingătoare (în cazul domeniilor tehnologice). Pânâ la sedimentarea elementelor definitorii ale unui domeniu de cercetare, pot coe sau se pot succeda mai multe direcţii de cercetare. Dintre acestea unel dispare sau îşi pot diminua foarte mult influenţa în raport cu motivaţia iniţială îşi pot continua însă existenţa prin noi motivaţii, prin alegerea de noi obiective

Ca element istoric, aş preciza că în toate evocările pe care le-am ci cel ce pentru prima dată a sugerat idea folosirii calculatorului şi a tehnici decodificare pentru prelucrarea automată a limbajului natural a fost Waren W în 1946. în 1949 el scrie lucrarea **„Translation"** considerată de toţi special traducere automată ca primul document programatic al acestei discipline. în a avut loc la Universitatea Georgetown din SUA prima conferinţă dedicată ex traducerii automate. în 1954, Peter Toma de la Universitatea Georg împreună cu un grup de cercetători de la IBM, realiza primul experim traducere automată (engleza-rusa) folosind un dicţionar de 250 de cuvint reguli sintactice de rescriere. Acest sistem avea să constituie nucleul fain program de traducere automată Systran pe care Peter Toma îl finalizează în

> *Punctul meu de plecare s-a aflat în lucrările unor Kulagina Melciuk, puternic implicaţi în studiile de traducere automată rus franceză, Yves Lecerf, implicat în problemele de documenta automată, D. G. Hays, implicat în traducerea automată din rusă engleză şi reciproc, B. Vauqois, cu preocupări de informat lingvistică la Grenoble. De la ei, ca şi de la alţi autori similari, preluat în bună măsură ştafeta pe care am căutat s-o duc departe. Ceea ce afirm despre mine este valabil pentru cei mai m cercetători din domeniul LM din anii 1950 şi 1960, cum ar fi Mau Gross, Masami Ito, A. Trybulec şi mulţi alţii.*

Traducerea automată, dar mai ales eşecul primelor încercări de re a acestui obiectiv încă nerezolvat sau nerezolvat complet, a constituit fără o motivaţie a „emergenţei" LM. Aşa cum voi arăta pe larg mai departe proiectelor de traducere automată au fost puse, prin interpretarea unila tendenţioasă a raportului APLAC, exclusiv pe seama inadecvării teoriilor li folosite atunci şi a cantonării în fapticul unor limbi particulare. Teoria „ înnăscute a limbajului" lansată de Chomsky, opunându-se tradiţiei tipol studiu lingvistic prin diversitatea limbilor, a generat o prodigioasă cerc direcţia determinării principiilor gramaticii universale, în speranţa că identi caracterizarea lor riguroasă le-ar putea operaţionaliza atât pentru ex comunicării umane prin limbaj cât şi (un derivat subsidiar al obiect Chomsky) pentru realizarea de sisteme de traducere automată apro performanţa umană.

> *Dubioasă mi se pare sintagma "soluţii abstracte", probabil et unui obicei binecunoscut de a diaboliza abstractul.*

Remarca de mai sus mă surprinde de două ori: mai întâi pentru c nimic reprobabil în expresia „o soluţie abstractă" (ba chiar dimpotrivă: "a Care rezultă din separarea şi generalizarea însuşirilor caracteristice ale

de obiecte sau de fenomene care este considerat independent, detaşat de obiecte, de fenomene sau de relațiile în care există în realitate" DEX'96) şi apoi referirea la un obicei binecunoscut (al cui?) de diabolizare a abstractului. Nu neagă nimeni că acele soluții abstracte de care aminteam au generat idei valoroase şi cercetări computaționale (mai ales în domeniul traducerii automate bazate pe conceptul „interlingua") dar rezultatele acestor idei şi cercetări nu sunt revendicate nici chiar de Chomsky.

> *în ceea ce priveşte sintagma "lingvistică formală", ea a căpătat o anumită utilizare în anii târzii 1960 şi în anii următori, iar personal am folosit-o în unele lucrări, după cum se va vedea imediat, considerând-o oarecum echivalentă cu LM; dar chiar dacă nu acceptăm această echivalență, nu putem eluda faptul că lingvistica formală se află în imediata vecinătate a LM.*

Cu amendamentele cronologice pe care le-am comentat mai devreme, apropierea între LM şi LF (lingvistica formală) este exact ceea ce am afirmat şi eu.

> *DT pretinde ca LM "a încercat", sugerând astfel ca ea a eşuat în tentativa de modelare a limbajului natural.*

în primul rând este vorba de modelarea computațională a limbajului. în al doilea rând nu eu pretind acest lucru, dar sunt perfect de acord cu el. Iată câteva opinii ale unor mari specialişti, activi, din domeniul prelucrării automate a limbajului natural (sublinierile îmi aparțin):

- Cristopher Manning and Hinrich Shutze: Foundations of Statistical Natural Language Processing, The MIT Press, 1998:

> .....the availability of large text corpora has changed the scientific approach to language in linguistics and cognitive science. Phenomena that were not detectable or seemed uninteresting in *studying toy domains and individual sentences* have moved into the center field of what is considered important to explain."

- Susan Amstrong-Warwick (editor): Prefața la „Special Issue on Using Large Corpora", Computațional Linguistics, Volume 19, no 1, 1993 p. 4:

> „What is that has brought about this rapid grouth of interest in corpus-based NLP?...The technological advances in computer power has certainly favoured the approach, as has the growing availability of large-scale textual resources in machine readable form. *More important, perhaps, is the growing frustration of trying to use standard rule-bâsed methods to account for more than a well-chosen fragment of text, regardless of the application.* The data extracted from large corpora have demonstrated that language is more flexible and complex than that which most rule-based systems have up to present tried to account for. The relative lack of practicai results at a time when industrial concerns are looking to the CL community to demonstrate progress toward useful applications has also contributed to the growing interest in new methods.

> And finally, the success rate demonstrated in the speech community offe
> hope for similar progress in NLP."

- Nancy Ide and Jean Veronis (editori) Computațional Linguistics -Speci Issue on Word Disambiguation, Voi. 24, No. 1 1998 p.15:

> „Although quantitative methods were embraced in early MT work, in t mid-1960s interest in statistical treatment of language waned amo linguists due to the trend toward the discovery of *formal linguistic ru sparkled by the theories ofZellig Harris (1951) and bolstered most nota by the transformational theories of Noam Chomsky (1957). Instea attention turned toward full linguistic analysis and hence to sentenc rather than texts, and toward contrived examples and artificially limit domains instead of general language."*

- Victor Yngve: From Grammar to Science:New Foundations for Gene Linguistics, John Benjamin Publishing Company, 1996:

> „there seems to be no scientific way of deciding among the ma contenders...We fiind positions and methods being promoted like a n movie or defended with withering polemics or taken up like the lat fad...We *should abandon logical-domain theories entirely and move to physical domain...Because this (notation) can be programmed on computer it can be used to test large-scale models...Gone will be babei ofarbitrary grammatical notations, each to be discarded in turn".*

Deşi nu împărtăşesc în întregime poziția extrem de radicală a lui Yngve, este simptomatică pentru insatisfacția generală față de abordările tradiționale anilor '60-80.

- R.F. de Bruine (editor) „Synthesis of Proposal for an RTD Programme Users, Industry and Research in Language and Technology", DGXIII, Commiss of the European Cummunities, September 1992:

> „There is a broad need to further understanding of linguistic phenomen the context of computerising the analysis and generation of langua General research should be stimulated within the following three n topics:
>
> - research on the linguistic meaning representation at the var level of description, ranging from the lower (e.g. phon morphological and syntactic) and better understood ones to higher, scientifically more difficult ones (e.g. semantic, pragm contextual and communicative ones). It is forseen that the fo must yield results in the short to medium term. Even if the l are long-term enterprises, they must be organised in way ensures availability of usable intermediate results.
> - reasearch on *more adequate and efficient computa schemes for natural language processing (e.g. constraints b computing and quantitative aspects) providing the base for r*

processing behaviour vz the applications of advanced computer science and statistical methods in close collaboration and synergy with related actions.

- research into the human factors related with the future spread of advanced language processing technologies taking into account the ergonomics aspects, economic and socio-cultural dimensions."

Lista unor astfel de citate poate continua pe zeci de pagini, dar am să mă opresc aici nu înainte de a mai reaminti raportul comisiei prezidate de Alain Danzin „Towards a European Language Infrastructure". Acest document, o adevărată cartă albă a cercetării în domeniul tehnologiilor limbajului, a restructurat complet programele de cercetare şi prioritățile pe termen mediu şi lung. A o ignora (ba chiar mai mult a o critica fără a-i cunoşte conținutul şi a o eticheta ca pe un document birocratic al celor de la Uniunea Europeană) poate fi desigur o opțiune personală, dar cu efectul izolării ştiințifice şi mai accentuate.

Ceea ce este deocamdată numai o sugestie devine, după cum se va vedea, o certitudine pentru DT. într-adevăr, iată ce scrie mai departe DT([1]: 133):

"Curând metodele lingvisticii matematice şi-au atins limitele drept care, în anul 1966, la propunerea lui David Hays, domeniul de cercetare al limbajelor naturale, din perspectiva utilizării acestora în interacțiunea cu calculatoarele electronice, este individualizat sub numele de lingvistică computațională".

Chestiunea cu atingerea limitelor ține de domeniul umorului involuntar şi trecem peste ea, dar nu ne miră, dupa ce am văzut la ce se reduce LM pentru DT.

în ciuda repetatelor mele clarificări, şi după cum se observă şi din citatul de mai sus, referirea mea era la utilizarea metodelor lingvisticii matematice în programele de prelucrare a limbajului şi nicidecum la domeniul în sine. Probabil că pentru cine nu a încercat să realizeze un sistem de prelucrare a limbajului natural şi nu s-a lovit de problemele implementării unui dicționar şi a unei gramatici computaționale e mai greu de înțeles remarca mea anterioară. Domnul Academician Marcus nu s-a apropiat niciodată de problemele unei implementări şi prin urmare nu mă surprinde lipsa de înțelegere a diferenței între o definiție formală a unei gramatici (de exemplu) care se explicitează în câteva, rânduri şi implementarea unei gramatici computaționale care nu numai că nu încape în câteva sute sau mii de pagini dar reclamă o muncă exprimată convențional în mii de oameni/an. Gramatica computațională a limbii engleze, dezvoltată în cadrul proiectului Alvey, a fost rezultatul a 10 ani de muncă intensă a celor mai importante 12 colective de cercetare din Anglia, fiecare dintre acestea fiind conduse de cercetători importanți şi fiind suplimentate cu numeroşi studenți doctoranzi. Gramatica GPSG dezvoltată este unul din exemplele standard de gramatică introspectivă de mari dimensiuni. Un astfel de efort uman şi financiar nu este la îndemâna multor societăți. Şi experiența a arătat că nici nu este necesar! Ralph Grishman, de la Universitatea din New York a demonstrat că programul său de inducție gramaticală, pe baza unui corpus de antrenare a generat o g nucleu, a cărei „finisare" a durat mai puțin de două săptămâni şi, confru gramatica Alvey pe un text arbitrar a reuşit să analizeze mai multe fraze cuvinte a demonstrat o mai mare acoperire lingvistică.

Nu mi-am imaginat niciodată că între LM şi LC ar putea avea l competiție, prima definindu-se prin metoda (căci ce altceva este decât studiul limbajului cu ajutorul matematicii ?) iar a doua obiectivul pe care şi-l propune. LM nu poate ignora problematica iar LC nu-şi poate realiza proiectele fără LM. Probabil însă că lucrează cu o definiție specială a LM, pe care am dori s-o *aflăm*

Nici nu există această competiție decât în imaginația Academician care sugerează mai sus că LC nu foloseşte matematic atunci când o face, disciplina se numeşte LM. Ceea ce, aşa cum am înainte, este fals. Elementele suplimentare, esențiale şi definitorii sunt ca algoritmii eficienți şi cunoştințele cu care acesta trebuie „hrănit". O forn procesului de înțelegere şi/sau producere a limbajului natural, de orice s fi ea, nu este decât o ipoteză asupra unui fenomen încă neelucidat. acestei ipoteze este cheia care a diferențiat LC de LM. în anexa acestei furnizat două definiții pentru LM şi LC. Prima definiție (LM) aparține lui G Pullum şi Andras Kornai iar cea de a doua (LC) se află în pagina WEB a de Lingvistică Computațională (al cărui membru sunt din 1985). Aş precizarea că lingvistica teoretică modernă (în sensul precizat m studiază limbajul nu numai cu ajutorul matematicii. Alături de m sociologia, psihologia, medicina şi ştiințele cognitive constituie d cunoaşterii care sunt fundamental implicate în explicarea acestui miracc reprezintă comunicarea inter-umană. Incapacitatea actuală de a procesor artificial de limbaj la nivelul performanței şi competenței datorează nedescifrării (încă) a mecanismelor minții şi creierului omului structural-fenomenologic şi noile cercetări în direcția unei ştiințe (reprezentată între alții de lucrările de pionierat ale Academician Drăgănescu) sunt fără îndoială porți deschise spre cunoaşterea, în exactă a minții şi implicit a facultății limbajului. Până atunci, obi (realizarea de sisteme automate capabile să prelucreze limbajul natura la modele aproximative, a căror acceptabilitate se probează prin imple evaluarea lor pe date reale. Cum între afirmarea unui obiectiv de LC ş sa operațională este o distanță mare, pe care uneori cercetătorii făr tehnologia programării fie că o ignoră, fie nu vor (şi de multe ori interesați) să o parcurgă, confuzia ce duce la auto-acreditarea într- conex este explicabilă.

Modul simplificator în care DT se referă la generativismul într-o logică binară care eludează faptul că în materie de se lucrează cu grade de adecvare şi relevanță, es simptomatic pentru viziunea sa limitativă în problema în disc

Crede DT că gramaticile lui Joshi, atât de importante în LC, puteau fi concepute fără să fi fost precedate de cele ale lui Chomsky? Da, Chomsky a fost tot timpul foarte controversat, dar fără stimulentul său nu ştiu ce ne-am fi făcut, inclusiv în LC şi în LM, în ciuda faptului că el nu s-a prea referit explicit nici la LC, nici la LM.

Modul „simplificator" incriminat mai sus se referă la fraza „soluții abstracte de modelare generativă de tip universal". Având în vedere că în articolul [1] aceasta este singura referire la generativism, bănuiesc că domnul Academician Marcus a vrut să spună „succint". Apoi, continuarea ce se referă la logica binară pe care o folosesc în interpretare şi simptomele viziunii mele limitative asupra problemei discutate desigur sunt efecte stilistice nereuşite, întrucât nu am abordat (şi nici nu mă interesează în mod deosebit) subiectul pe care îl invocă domnul Academician. Pentru că tot am ajuns aici, țin să-i reamintesc domnului Academician Marcus că Noam Chomsky şi-a revizuit complet punctul de vedere care a dominat aproape 15 ani lingvistica mondială. într-adevăr Chomsky este un mare om de ştiință, chiar dacă foarte controversat, dar acest statut îi este conferit şi de onestitatea cu care s-a detaşat de creațiile sale anterioare ce i-au adus notorietatea, dovedite (unele chiar de el însuşi) ca fiind depăşite, propunând soluții şi teorii noi.

Formalismul TAG al lui Joshi este într-adevăr unul foarte important în LC ca şi HPSG, LFG, CG şi alte câteva. Dar dintre formalismele de lingvistică computațională, TAG este cel mai departe de influența chomskyană. Dacă se poate face o asociere între TAG şi vreo teorie generativistă de tip chomskyan aceasta este doar de natură antinomică. Am colaborat cu profesorul Aravind Joshi în 1991 la Institutul Lingvistic de la Universitatea Santa Cruz din California, am fost apoi invitatul său la Universitatea din Pennsylvania, invitație motivată printre altele şi de o deosebită apreciere pentru o demonstrație alternativă a mea, mai scurtă şi, considerată de profesorul Joshi, mai elegantă a unei teoreme a domniei sale referitoare la categoria de limbaje acoperite de LTAG. Cu acea ocazie, profesorul Joshi mi-a pus la dispoziție trei volume consistente de lucrări asupra TAG tratând foarte amănunțit motivațiile lingvistice, proprietățile computaționale şi caracterizarea matematică. Aceste volume i le-am pus la dispoziție şi domnului Academician Marcus. Profesorul Joshi a fost in 1997 invitatul profesorului Dan Cristea şi al meu la Şcoala de Vară EUROLAN unde a susținut o serie de prelegeri de înaltă ținută ştiințifică. Am evocat aceste lucruri pentru a-l lămuri pe domnul Academician Marcus că formalismul TAG şi varianta sa mai nouă LTAG îmi sunt familiare şi prin urmare mă surprinde afirmația dânsului implicând o filiație între teoriile tui Joshi şi Chomsky.

Faptul că gramaticile context free se află din nou, începând cu anii "80, în centrul atenției în LC nu spune ceva ?

Acest lucru este exact şi ilustrează foarte bine ceea ce spuneam înainte: contextul computațional în care complexitatea algoritmică este primul mare judecător al adecvării unui model (inerent limitat, după cum arătam mai devreme) bazat pe o anumită teorie lingvistică. în anii de vârf ai lingvisticii matematice, şi în cei de început ai lingvisticii computaționale, pornindu-se de la o conjectură Chomsky (limbajele naturale nu sunt limbaje independente de context) demo în anii '80 de Gerald Gazdar (autorul teoriei GPSG), cercetarea a fost orienta identificarea de formalisme lingvistice cât mai puternice, cu puterea generativ mai apropiată de cea a gramaticilor universale (echivalente deci cu m Turing). Formalismul ATN (Augmented Transition Networks) al lui William W de la BBN a fost timp de peste 10 ani suportul standard al majorității sistemel prelucrare a limbajului natural. Eu însumi am dezvoltat în anii 1984 şi 198 mediu de programare lingvistică conținând un editor de gramatici ATN compilator ATN. Din punct de vedere formal ATN-ul este echivalent cu o m Turing şi tocmai această putere formală prea mare l-a scos din competiția sol utile în lingvistica computațională. La sfârşitul anilor '80 obiectivul major (valabil şi astăzi) a devenit identificarea unui formalism de putere generativ mai mică dar care să acopere cât mai multe din problemele practice pu prelucrarea automată a limbajului natural. Aşa au revenit în actualitate gram independente de context şi s-au dezvoltat abordările lexicalizate. Cele din ur fost propuse tocmai pentru a rezolva, în cadrul scheletelor de gra independente de context, idiosincrasiile limbajului natural cel mai adesea loca la nivelul lexical. Mai mult, după anii '90, odată cu resurecția interesului fa abordările statistice, gramaticile regulate şi automatele finite au căpătat o ut foarte largă.

LC are mai multe părți, mai multe orientări, mai multe niveluri d abstracție, care comportă criterii diferite de evaluare.

Este adevărat că actualmente în LC se regăsesc orientări, abordă

- motivații diferite. Dar indiferent de sorginte, ele se plasează (cel puțin declar contextul computațional prin raportarea la un mediu software de prelu Considerând exemplul HPSG, probabil cea mai în vogă teorie ling computațională actuală, atunci când Ivan Sag analizează sau argume adecvarea teoriei sale în descrierea formală a unei limbii naturale (aşa procedat în recentele sale conferințe la Facultatea de Litere a Univ Bucureşti şi în Aula Academiei Române) el se plasează în sfera lin teoretice. Atunci când prezintă soluțiile de implementare a unui fragment n limbii engleze şi discută rezultatele generate de analizorul HPSG dezvo grupul sau de la Universitatea Stanford şi modalitățile algoritmice de rezo ambiguităților (aşa cum a făcut în prelegerea susținută la sediul RACAI plasează în sfera LC.

DT îl asociază pe D. Hays la ideea sa privind falimentul LM lansarea, drept consecință, a LC.

Afirmația de mai sus conține două lucruri false:

a) nu am vorbit de falimentul LM ci de insuficiența metodelor momentul invocat (cred că citatele pe care le-am preze argumentele aduse până acum sunt lămuritoare).

b) Eu nu-l pot asocia pe David Hays la o idee pe care nu am exprimat-o.

în textul meu original scriam: „la propunerea lui David Hays, domeniul de cercetare al limbajelor naturale, din perspectiva utilizării acestora în interacțiunea cu calculatoarele electronice, este individualizat sub numele de lingvistică computațională".

Propunerea lui Hays venea în sprijinul identificării unui nume comun pentru diversele preocupări asupra limbajului din perspectiva implementării de sisteme automate de prelucrare. Traducerea automată, un domeniu care se dezvoltase distinct de celelalte preocupări în domeniul prelucrării automate automate a limbajului natural, căzuse în disgrație în urma raportului ALPAC ***(Languages and machines: computers in translation and linguistics.*** *A report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council. Washington, D.C.: National Academy of Sciences, National Research Council, 1966. (Publication 1416.) 124pp.).* **în** raportul ALPAC, comandat în 1964 de Academia Națională de Ştiințe, în afara criticilor deosebit de dure la adresa realizărilor şi abordărilor de până atunci în domeniul traducerii automate existau şi o mulțime de recomandări care se refereau *la noi metode de investigație ştiințifică şi la abordarea unor obiective mai realiste.* Istoria domeniului a reținut (pe nedrept) doar apriga critică a lui Bar-Hillel care, considerată unilateral, a dus la stoparea pentru circa 15 ani a cercetării oficiale în domeniul traducerii automate în SUA şi mai apoi în majoritatea țărilor dezvoltate (o incitantă prezentare a a ceea ce a însemnat proiectul ALPAC este „ALPAC: the (in)famous report", *http://ourworld.compuserve.com/homepages/WJHutchins/Alpac.htm,* şi îi aparține lui **John** Hutchins). Ceva trebuia făcut pentru a conserva câştigurile ştiințifice obținute până atunci şi a permite în noul context continuarea cercetărilor anterioare cu scopul declarat al realizării de programe cu obiective realiste. O serie de minți luminate **(John** Pierce, David Hays, **John** Carroll) au văzut pericolul ca, asociate cu domeniul traducerii automate, toate celelalte preocupări privind prelucrarea automată a limbajului puteau fi periclitate, şi în acest sens în raport s-a inserat un capitol distinct numit „Automatic language processing and computațional linguistics" ce arăta beneficiile aduse de cercetarea în domeniul traducerii automate în domeniile prelucrării automate a limbajului şi al lingvisticii computaționale. Printre altele în capitolul respectiv se arată că „...(what is required is) **basic developmental research in computer methods for handling language,** as tools for the linguistic scientist to use as a help to discover and state his generalizations, and ... to state in detail the complex kinds of theories..., **so that the theories can be checked in detail."** (sublinierea mea, DT). Mai mult preşedintele comitetului de elaborare a raportului ALPAC, **John** Pierce, conştient de pericolul interpretării greşite sau al ignorării recomandărilor prezente în anexele raportului (aşa cum s-a şi întâmplat), a ținut să insereze în raportul final adresat preşedintelui Academiei Naționale de Ştiințe o secțiune nouă care sublinia idea de a susține lingvistica computațională în mod distinct de traducerea automată („supporting computațional linguistics, as distinct from automatic language translation"). Dezvoltând ideile din capitolul raportului AL referitor la prelucrarea limbajului natural (concept care şi atunci şi acum este de cel al traducerii automate) Pierce considera că NSF (National Sc Foundation) trebuia să asigure fonduri de cercetare pentru dezvoltarea de m de limbă de dimensiuni mari **„since small-scale experiments and wor miniature models of language have proved seriously deceptive in the and one can come to grips with real problems only above a certain sc grammar size, dictionary size, and available corpus".**

Acesta este contextul în care David Hays, activ cercetător la înc anilor '60 în domeniul traducerii automate (de altfel unul din membrii comitet au elaborat raportul ALPAC) a propus individualizarea preocupărilor leg prelucrarea limbajului natural cu ajutorul calculatorului, dezvoltarea de mod limbă realiste (nu miniaturi la îndemâna cercetării individuale) şi a apl „serioase" (în opoziție cu experimentele la scară mică) sub numele de *lin computațională.*

> *Denumirile folosite pentru preocupările la interferența limbajel informaticii şi matematicii au variat tot timpul şi nu cred că ace aspect merită prea multă atenție. Lingvistică matematică? co putațională? inginerească? algebrică? cognitivă? aplicată? canti tivă? teoretică? statistică? probleme matematice ale semioticii? te nologia limbajului? limbajul în inteligența artificială? lingvist inginerească? procesarea limbajului natural? "information stora retrieval"? lingvistica cibernetică? pe fiecare dintre acestea întâlnit-o şi propriile mele articole au fost publicate aproape s fiecare dintre etichetele de mai sus.*

Citatul de mai sus mi se pare extrem de relevant pentru discuția de defineşte clar diferența de opinii. Dacă de pildă distincția dintre *medicină u medicină veterinară* sau (coborând în taxonomie) între cardiologie şi stom „nu merită prea multă atenție" atunci domnul Academician are dreptate.

Din punctul meu de vedere însă, este o mare diferență între denominări ale studiului limbii amintite mai sus (la care se mai poate adăug la fel de numeroasă), ele definind câteva domenii distincte definite prin obiective, competențe, metode şi modele.

> *în 1962 s-a înființat în USA "Association of Computați Linguistics".*

De fapt în 1962 s-a înființat AMTCL, acronim pentru „Associa Machine Translation and Computațional Linguistics", primul preşedinte al fiind Victor Ingve (cel pe care l-am citat mai devreme), iar al doilea fiin Hays. ACL (Association of Computațional Linguistics) a apărut abia în 1968

> *în 1963 Ferenc Kiefer a demarat la Budapesta re "Computațional Linguistics", care a trăit peste zece ani.*

Este adevărat, dar conținutul ei era foarte diferit de al revistei „Mechanical Translation and Computațional Linguistics' apărută în 1965 ca revistă oficială a AMTCL. Şi tot ca un rezultat al diferențierilor tot mai mari care apăruseră în domeniu, AMTCL îşi încetează activitatea la începutul anilor '70 fiind înlocuită de „American Journal of Computațional Linguistics" care în 1984 devine „Computațional Linguistics" (actuala denumire).

Conferința de la Grenoble de "traitement automatique des langues" din 1967 era a treia de acest fel, fiind precedată de o alta, la New York, în 7965 şi de una in Anglia, probabil în 1963, organizată de M. Masterman. între timp, la ruşi, numeroase conferințe au avut loc pe tema "avtomaticeskaja obrabotka tekstov" iar "Sprachkunde und Informationsverarbeitung" a fost uneori eticheta folosită de germani s.a.m.d. Nu negăm rolul important pe care l-a avut David G. Hays în dezvoltarea CL, dar acest rol a fost altul decât cel afirmat de DT.

Nu am să reiau explicația faptului că nu i-am atribuit lui Hays nici un rol demolator, dar trebuie să subliniez faptul că inițiativa lui David Hays, de care am discutat mai devreme, a avut un rol **fundamental** în evoluția CL. Aşa cum am arătat mai sus, inițiativa disocierii de traducerea automată, pentru a nu periclita restul preocupărilor privind prelucrarea automată a limbajului a fost o necesitate conjuncturală. In 1965, când la New York a avut loc prima conferință COLING, Hays anticipa desigur efectul de bumerang al raportului la elaborarea căruia participa, şi a propus chiar atunci, detaşarea oficială prin sintagma „computațional linguistics" de domeniul traducerii automate (pe care îl părăsise de altfel şi Hays cel ce fusese unul dintre principalii specialişti în traducere automată ai RAND Corporation). Deci nu Hays a creat domeniul lingvisticii computaționale, el este cel ce a „oficiat" botezul. Şi nu a făcut-o de pe orice poziție ci de pe cea de fost membru al Comisiei Alpac şi de preşedinte al AMTCL.

Emergența LC s-a produs încă din anii "50, sintagma LC a devenit curentă încă de la începutul anilor "60. Şirul de conferințe COLING nu a făcut decât să continue aceasta tradiție. Alții au preferat folosirea sintagmei LM (a se vedea, de exemplu, "Prague Bulletin of Mathematical Linguistics", "Prague Studies of Mathematical Linguistics", revista japoneza^ "Mathematical Linguistics" (in echivalentul ei japonez) etc. în ceea ce priveşte însă profilul acestor reviste, nu am constatat o diferență față de cele de CL. Desigur, între timp au început să apară şi unele publicații mai specializate, cu referire la părți determinate ale CL (cum ar fi cea relativă la corpusul lingvistic). Etichetele nu au avut importanta si nu stiu sa se fi desfăşurat vreo competiție intre ele. Chiar Hays a folosit diverse etichete, de exemplu cea din [3].

Persistența cu care domnul Academician pune semnul egalității între domeniul lingvisticii matematice, în care fără discuție nu a avut sau nu are rival în România, şi cel al lingvisticii computaționale sau tehnologia limbajului este apa foarte curioasă. Nu şi dacă observăm următoarele fapte:

- sintagma „lingvistică matematică" este din ce în ce mai puțin utili (o căutare pe internet a termenilor „mathematical linguisti „computațional linguistics", „natural language processing"şi „langu technology" este foarte instructivă: numărul de documente ce îi re este 4.630, 87.900,169.000 si respectiv 2.840.000);
- în domeniul strict computațional, la care se referea [1], în Rom activează de câtva timp o serie de cercetători importanți (majorit dintre ei membrii ai Comisiei de Informatizare pentru Limba Rom pe care am onoarea să o conduc, şi din care de altfel face par domnul Academician Marcus);
- domnul Academician Marcus fie nu cunoaşte, fie dezavu rezultatele româneşti obținute în domeniul **prelucrării cu calcula** a limbii române (cel puțin aşa poate fi considerată ignorarea com a acestora în lucrările domniei sale); ori poate consideră că reprezintă domeniul său de interes.

Dar DT merge mai departe pe ideea sa şi afirmă (în completă discordanță cu viziunea lui Hays, de la care se reclamă) că "metodele LM şi-au atins limitele" (încă în urmă cu peste 30 de ani!), pentru ca numai două pagini după această afirmație (deci la pagina 135 din [1]) să afirme că e nevoie de "modele formale ale limbii la toate nivelurile ei (fonetică, morfologie, sintaxă, discurs) gramatici formale [...]". Cum vede DT aceste modele formale altfel decât sub forma logico-matematică?

Asupra primei părți a acestei fraze cred că am discutat suficient. Ref la „contradicția" pe care o semnalează în partea a doua a frazei de mai sus, n decât să-i recomand domnului Marcus să citească încă de câteva ori art respectiv (sau să-l citească integral). Este vorba de **NOI modele formale de** (în opoziție cu cele vechi), resurse lingvistice computaționale adecvate mome actual. Dintre noile teorii care au apărut şi s-au şi impus aş putea să ami teoria optimalității în comunicare dezvoltată de Prince and Smolensky în 199 implementări în domeniul fonologiei şi morfologiei computaționale ş promițătoare rezultate chiar în sintaxă), teoriile sintactice bazate pe unifica satisfacerea de restricții, precum şi o întreagă pleiadă de teorii ale discursu domeniul prelucrării automate a limbajului natural există standarde, tehnologii specifice, există organizații mondiale specializate, mai toate apăr ultimii 10-15 ani. Dacă domnul Academician Marcus poate afirma că pentru română în domeniul resurselor lingvistice computaționale s-a făcut (sau a ceva înainte de anii '80 înseamnă că domnia sa are o imagine complet diferită tuturor specialiştilor din lume.

Ştie oare că multe modele de acest fel există de câteva decenii? Indicații bibliografice asupra lor sunt date parțial in [4], [5], [6], [7] iar pentru cercetările româneşti in [8], [9]. Desigur, aceste modele sunt inegale ca valoare, au nevoie de continuări, modificări, ameliorări, dar ele nu pot fi ignorate. Fonetica, fonologia, vocabularul, morfologia, sintaxa, semantica lingvistica şi lingvistica istorică au beneficiat din plin de metodele matematice, aşa cum se poate vedea din impactul deosebit al lucrărilor respective în literatura de specialitate;

Recursul la modelele anilor '60-70 descrise în lucrările menționate ca argument pentru concepte ce au apărut la începutul anilor '90 mă scuteşte de comentarii. Pe de altă parte, avansul ştiințific în orice domeniu se clădeşte pe cunoaşterea anterioară iar cazurile de „frângere cognitivă", când salturile ştiințifice neagă cunoaşterea anterioară sunt rare şi ele de regulă definesc revoluțiile în ştiință. Filiația sau influențele în dezvoltarea unui domeniu ştiințific (atunci când ele pot fi depistate cu obiectivitate) constituie preocuparea istoricilor ştiinței. Lucrările tehnice, de regulă se raportează la contemporaneitate, ceea ce în termeni temporali poate însemna, în funcție de dinamica domeniului, câțiva ani, un deceniu, mai multe decenii sau perioade chiar mai mari. De pildă, puține lucrări tehnice în domeniul lingvisticii teoretice, al fonologiei se referă la marele gânditor Panini, considerat de mulți oameni de ştiință creatorul ştiinței limbii. Lucrarea sa fundamentală *Astaka,* cunoscută şi sub numele de „gramatica lui Panini" conține descrieri formale ale regulilor de producție ale limbii sanscrite şi o clasificare cu peste 1700 de elemente constitutive ale limbajului. Aceste elemente sunt organizate în clase a căror agregare este descrisă prin intermediul unor reguli ordonate, într-o manieră apropiată de teoriile actuale. El poate fi considerat un precursor al teoriei limbajelor formale şi al lingvisticii matematice, dar puține cărți sau lucrări de referință în aceste domenii menționează numele genialului savant ce a trăit cu mai bine de peste 2500 de ani în urmă. în schimb, numele său se regăseşte în orice lucrare serioasă de istorie a lingvisticii formale.

Obstinația cu care domnul Academician Marcus încearcă să sugereze că eu aş dezavua metodele matematice, sau rezultatele importante ale lingvisticii româneşti dovedeşte că domnia sa complet neinformat în ceea ce mă priveşte.

DT indică, drept domeniu al LM, numai "aspectul sintactic, gramatical", despre celelalte nu a aflat. Nu a aflat nici ca LM a abordat şi aspecte analitice, nu numai pe cele generative.

Fals: **„numai"** este imaginația domnului Academician. Citatul corect este: „în **general** al aspectului lor sintactic, gramatical".

DT defineşte "dimensiunea fundamentală" a LC prin "fezabilitatea instanțierii unei descrieri lingvistice cât mai complete, mentenabilitatea acestei instanțieri şi, desigur, conformanța cu realitatea uzului limbii". *([1]: 133). Cu un mic efort înțelegem despre ce este vorba.*

Desigur că problemele de complexitate, de cost, nu puteau fi în abordate în anii '50 şi '60 cu mijloacele cu care ele au început a studiate în a doua jumătate a anilor 70, când instrument elaborate în informatica matematică deveniseră mult m perfecționate. Dar acest fapt nu ține, cum crede DT, de alegeri între LM şi LC, ci de progresul general realizat în ştiință. Pentru mă referi la propria noastră experiență, atunci când, în 19 prezentam la COLING-ul din Suedia gramaticile contextuale aveam cum să mă ocup de aspectul complexității acestor grama în maniera în care s-a putut face acest lucru ulterior (a se ved de exemplu, [10]). Dar acest fapt nu are nici o legătură cu etich folosită.

Efortul (chiar mic) este probabil generat de unii termeni de sp nefamiliari domnului Academician. Voi furniza lămuririle necesare mai jos.

Eu mă refer la perioada actuală când invoc ca dimensiune fund *fezabilitatea instanțierii* unei descrieri lingvistice cât mai complete. Instanți descrieri lingvistice înseamnă altceva decât complexitatea formală, de altfel şi amintesc în secțiunea trunchiată a citatului folosit de domnul Aca Marcus mai sus. Este un termen tehnic care se referă la construcția pro în baza unui formalism sau teorii lingvistice, a unei gramatici şi a dic aferent, care furnizate ca resurse unui program de prelucrare a limbajulu permit acestuia să analizeze sau să genereze un text arbitrar. O instanțiere este fezabilă dacă ea se poate realiza în condiții de timp ş umane rezonabile.

Nu m-am mirat atunci când "Encyclopedia of Microcomputers "Encyclopedia of Computer Science and Technology" mi-au so o contribuție cu tema "Semiotics and Formal Artificial Language se vedea [11]) şi nici când "Handbook of Formal Languages" solicitat un capitol privind "Contextual Grammars and Na Languages"[12] iar o lucrare preponderent teoretică a fost ins în "Computațional Linguistics in the Netherlands 2000"[13].

Nu văd rostul acestor lămuriri. Toată lumea îl ştie, îl recunoaşte dintre cercetătorii adevărați nu-l contestă pe omul de ştiință Marcus, reprezentant român al lingvisticii matematice, creatorul acestei şcoli în Ro articolul [1] nu m-am referit nici direct nici indirect la domnia sa. Fapt evocat criticile pe care le-am comentat anterior la adresa **metodelor** matematice ale începutului deceniului şapte nu are nici o legătură cu (încă o dată, deosebite) ale domnului profesor. însă probabil că identifică LM mondială, domnia sa a considerat critica asupra **metodelor LM** din a atac la persoana sa, adevărat act de blasfemie.

în anii din urmă, domnul Academician încearcă să transfere în noilor tendințe şi tehnologii ale limbajului, ignorând o realitate exis

portofoliul de rezultate pe care le-a obținut anterior creditându-le ca surse primare a tot ceea ce se întâmplă azi în tehnologia limbajului în România (şi nu numai). Şi cine nu este de acord cu acest lucru (parafrazându-l pe domnul Marcus) trebuie demonizat. Textul pe care îl comentez ca şi acțiunile recente declanşate de domnul Academician Marcus, pretinse a fi iscate de conținutul articolului [1], nu fac decât să-mi întărească această impresie. Eu nu am nimic de împărțit cu domnul Academician.

Nu m-am mirat nici când am văzut că o revistă cu titlul "Linguistics and Philosophy" publică articole excelente de LC. Interferențele merg în toate direcțiile şi ele caracterizează cultura contemporană, în acest orizont trebuie să ne plasăm, cred, atunci când ne referim la disciplinele cognitive care se dezvoltă sub ochii noştri şi îşi pun amprenta pe modul nostru de gândire şi de comportare. Un tratat ca "Mathematical Methods in Linguistics" [14] include multe fapte de LC, deşi în titlul său nu figurează epitetul "computațional". O revistă ca "Theoretical Linguistics" (1970 2000), publicată de Walter de Gruyter (Berlin-New York) a inclus multe articole vizând aspecte matematice şi/sau computaționale, deşi numele revistei nu indică acest lucru. Chiar o revista mai tradițională, ca "Linguistics" a inclus de multe ori articole de LM şi nici "Foundations of Language" nu a procedat altfel. Multe fapte de LM şi de LC se plasează în mod natural în orizontul semioticii computaționale.

Faptul că tratatul amintit nu incorporează în titlu atributul computațional nu mă surprinde, pentru că ar fi creat o confuzie pe care autorii au evitat-o deliberat. Cartea respectivă nu este o carte de lingvistică computațională, conținutul ei tratează exact ce anunță în titlu: metode matematice folosite în studiul lingvistic. Lingvistica teoretică, puternic formalizată în ultimele decenii apelează inevitabil (ca de altfel marea majoritate a domeniilor ştiințifice) la metode şi modele matematice.

Era internetului impune desigur o problematică nouă, față de care abordările anterioare se pot dovedi insuficiente.

Exact aceasta este esența celor 3 paragrafe din [1] incriminate şi combătute pe larg de domnul Academician Marcus: insuficența abordărilor anterioare. Conştientizarea acestei insuficiențe însă a precedat cu câțiva ani apariția internetului.

Salutăm inițiativa noii generații de cercetători de a se dedica noilor probleme.'

Nu putem ignora tonul paternalist privind noua generație de cercetători care se dedică problemelor ridicate de internet în prelucrarea automată a limbajului natural. INTERNET-ul este o revoluție! Şi implicațiile sale sunt atât de mari încât asigurarea accesului universal la Internet a devenit o problemă fundamentală chiar şi pentru o organizație de caliblul UNESCO. Am avut onoarea să fac parte din Comisia de Experți creată de Secretarul General al UNESCO (comisie de cel mai înalt nivel) pentru elaborarea documentului Recommendation on Multilingu and Universal Access to Cyberspace. Sunt al doilea expert român (dup Ambasador Dan Hăulică, Membru Corespondent al Academiei) care a făcut dintr-o comisie de experți UNESCO de acest nivel.

Ignorarea în cercetarea privind prelucrarea automată a limbajului natu fenomenului INTERNET este de neconceput. Societatea cunoaşterii are ca u premisele sale fundamentale accesul universal, neîngrădit de bariere lingvist cunoşterea stocată în internet. Alte comentarii sunt de prisos.

Dar trecerea de la ieri la azi şi de la azi la mâine nu poate fi decâ una care ține seama în mod critic de experiența acumulată.

Nimeni nu neagă acest lucru, şi faptul că l-am rugat insistent pe d Academician să facă parte din Comisia de Informatizare pentru Limba Ro cred că arată buna mea credință şi speranța pe care o nutream (şi care înc supraviețuieşte) că experiența domniei sale va fi pusă în slujba obiectivelor pe nici eu nici domnul Marcus nu le putem atinge singuri. în acelaşi spirit, i-am p domnului Academician Marcus să scriem împreună o antologie a cerce româneşti în domeniul lingvisticii formale şi computaționale, de la începutur care le evocă domnia sa şi pînă în zilele noastre. Din păcate propunerea a r fără răspuns.

Din tot ceea ce am prezentat mai sus rezultă clar ca LM şi LC a fost mereu împreună şi că, în general, etichetele nu au contat pre mult Unii au mers chiar mai departe; astfel, în capitolul 4 "Mathematical and Computațional Linguistics", din [15], se afirm pur şi simplu (p.86): "Mathematical linguistics has also been calle theoretical linguistics and even computațional linguistics". Iar ma departe, în acelaşi loc: "Computațional Linguistics originated aroun 1950 with the initiation of research on automatic translation" (s trimite la o carte editată de D.G.Hays [3] şi la o alta avându-l c autor pe acesta [16]).

Nu văd în pasajul pe care l-am citat mai sus nici un argument împot ceea ce am susținut în [1] şi în cele prezentate aici. Notez în treacăt ad „even" cu o valoare discursivă în completă consonanță cu considerentele is pe care le-am invocat ale evoluției ştiințifice şi tehnologice în domeniul prel limbajului natural.

în România, minți luminate ale anilor "60, ca profesorii Al. Roseti Grigore Moisil şi Tudor Vianu, au înțeles schimbările care se profila şi au sprijinit proiectul înființării unei secțiuni de "lingvistică aplicată la Facultatea de Limba şi Literatura Română a Universității d Bucureşti, dar s-au găsit alții care să-i torpileze.

Aşa este, şi mă bucură elogiul adus acestor corifei ai ştiinței româ Poate şi pentru că alături de câțiva reprezentanți importanți ai lingvisticii rom actuale care au înțeles tendințele şi imperativele momentului (Prof. Dan M decanul Facultății de Litere, Prof. Alexandra Cornilescu, Conf. Emil Ionesc

participat la reluarea acestei lucrări. Programul de Masterat în Lingvistică Formală şi Computațională de la Facultatea de Litere a Universității din Bucureşti, funcționează de mai bine de 2 ani şi nutresc speranța că Ministerul Educației şi Cercetării va aproba demersurile noastre privind chiar înființarea unui departament cu acest profil.

în acelaşi sens, am participat alături de profesorul Cristea (având fără discuție şi sprijinul altor minți luminate ale Universității A.I.Cuza din laşi) la lansarea în 2001 a Masterat-ului în Lingvistică Computațională al Facultății de Informatică. Nu este uşor să pendulezi între laşi şi Bucureşti, dar şi domnul profesor Cristea, şi doamna profesor Cornilescu şi eu o facem pentru ca cele două programe „surori" de maşter să-şi împlinească menirea de a pregăti câți mai mulți specialişti în folosul programelor de informatizare pentru limba română.

> *La Academia Română a funcționat mulți ani "Comisia de Lingvistică Matematică" iar revista "Cahiers de Linguistique Theorique et Appliquee", inființată în 1962, a fost multă vreme expresia colaborării lingvisticii cu matematica şi cu informatica. In ciuda forțelor adverse,* s-a *reuşit în acei ani atragerea unor studenți străluciți ai unor facultăți umaniste la cercetarea limbii si literaturii cu mijloacele LM si LC.*

Comisia de Informatizare pentru Limba Română de la Academia Română, înființată în anul 2001, încearcă, ținând cont de realitățile şi prioritățile actuale, să armonizeze eforturile celor ce lucrează în domeniul limbii române şi care cred în perspectiva înrolării ei în cadrul limbilor importante ale societății cunoşterii. Eu am convingerea că voi putea spune peste timp acelaşi lucru: *„In ciuda forțelor adverse,* s-a *reuşit în acei ani atragerea unor studenți străluciți ai unor facultăți umaniste la cercetarea limbii şi literaturii cu mijloacele"* tehnologiei limbajului.

> *Pentru a da numai două exemple de actuali profesori universitari care au susținut teze de doctorat de acest tip, voi menționa pe Pia Brinzeu, de la Catedra de Engleză a Universității din Timişoara şi pe Mihai Dinu, de la Facultatea de Litere a Universității din Bucureşti. Tot în acea perioada şi-a susținut teza de doctorat Sorin Cristian Niță, pe o tema de critică textuală automată privind înlănțuirea (filiația) diferitelor variante ale "Istoriei Tării Româneşti" (Şerban Cantacuzino).*

Exemple de profesori şi cercetători români valoroşi, cu contribuții substanțiale în domeniul limbii române se pot da foarte multe. Mulți dintre ei sunt în străinătate şi fac o bună propagandă ştiinței româneşti. Mi-e cunoscută cartea cu adevărat remarcabilă a domnului profesor Mihai Dinu „Personalitatea limbii române", de altfel premiată de Academia Română. Această lucrare este o solidă cercetare de lingvistică computațională în spiritul actual tocmai pentru că a parcurs acea cale dificilă a instanțierii lingvistice (în cazul său la nivelul componentului lexical).

> *lată însă că, în pofida realităților puse in evidență mai sus, in ([1]: 134) se scrie: "în România, cercetările în domeniul LC şi al prelucrării limbajului natural, precum şi primele rezultate practice au apărut la începutul anilor "80 [3, 4, 5, 6]".*
> *La ce trimit numerele indicate in paranteze ? La o bibliografie de 24 de titluri în care aproape toate (dar toate cele indicate intre paranteze) încep cu DT (ignorandu-se regula generală în lumea ştiințifică, a aşezării numelor autorilor aceluiaşi articol în ordine alfabetică; dar nu acest fapt este cel care ne interesează în momentul de față).*

înainte de a comenta acest pasaj şi pe cel următor, nu pot să trec observația absurdă şi falsă pusă între parantezele ce trădează totuşi o ezit probității omului de ştiință în fața unei răutăți gratuite. Nu există nici o r generaiă de genul celei afirmate. Ordonarea alfabetică este o convenție autorii cu contribuții egale în redactarea unei lucrări. Am deschis la întâm două volume de specialitate, conținând contribuții (S. Amstrong et al. „Natural Language Processing Using Very Large Corpora, Kluwer, 1999 Strzalkovski (ed) „Natural Language Information Retrieval", Kluwer, 1999). Di 19 lucrări cu mai mai mulți autori, doar trei urmăresc (probabil din întâm "regula generală în lumea ştiințifică" pe care o invocă domnul Academician care probabil a impus-o şi o impune tuturor celor alături de care publică, indi de contribuția fiecăruia.

> *Să observăm că încă în 1978, în articolul "Mathematical and Computațional Linguistics" [9] de prezentare a activității din România în domeniul LM şi LC se face referire la peste 400 de articole publicate de 130 de autori români şi sunt menționați peste 300 de autori străini (unii dintre ei, nume de vază ale LM şi LC din acea perioadă) care au citat şi continuat cercetările româneşti. Să mai adăugăm că numeroşi lingvişti români dintre cei mai importanți au citat şi folosit rezultatele şcolii româneşti de LM şi LC. lată ca vine acum DT şi face (deliberat sau nu) din tot acest efort un teren viran care-l aştepta peDTsă tragă primele jaloane. Nu e cam mult?*

Deşi am repetat de nenumărate ori până în acest moment, o mai dată, precizând că discuția din [1] se referea la **resurse lingv computaționale** şi **programe software de dialog în limbaj natural** (în română). Acestea erau rezultatele practice pe care le menționam în comentat cu gratuită aciditate. Poate să-mi menționeze domnul Academician sistem de dialog în limba română implementat înaintea sistemelor pe care realizat eu şi colaboratorii mei? lată câteva repere:

- Sistemul QA (1980) un sistem inferențial de întrebare răspuns în română, susținut de un demonstrator original de teoreme în c predicatelor de ordin 1;

- SDLR (1981) un sistem de dialog în limba română ce a extins capabilitățile lui QA cu operatorii lingvistici ai logicii fuzzy;
- IUREŞ (1983) sistem de generare automată a sistemelor de întrebare răspuns, independent de limbă, pe care l-am realizat împreună cu Dan Cristea, acum decanul facultății de informatică a Universității Cuza. Sistemul IUREŞ a fost omologat internațional în 1988 şi a constituit primul produs de inteligență artificială exportat (în acelaşi an). Sistemele IUREŞ şi SDLR sunt referite printre altele în enciclopedia de lingvistică computațională. Mai important este faptul că sistemele IUREŞ şi SDLR sunt amplu descrise în prestigioasa antologie "The Survey of the Current Status Research and Future Trends in Machine Translation and Natural Language Processing" realizat in 1992 de JEIDA (Japan Electronic Industry Development Association), fiind de altfel singurele sisteme de dialog în limbaj natural din întreaga zonă fost comunistă incluse în această carte.

Acestea erau referințele incriminate de domnul Academician şi dacă domnia sa poate să-mi indice un singur sistem de prelucrare a limbajului natural realizat în România înaintea celor pe care le-am citat, eu am greşit. Dar mă îndoiesc. Nu cunosc conținutul articolului menționat (pe care i l-am solicitat de altfel domnului Academician, fără a-l primi însă), astfel încât nu pot afirma nimic despre cei 130 de autori români ce au realizat (conform afirmației domnului Marcus) lucrări de lingvistică computațională. Ce pot însă să afirm este că am citit multe din lucrările de lingvistică teoretică contemporană ale marilor noştri lingvişti şi ele au fost extrem de relevante ca material faptic în cercetările mele. Dar lucrările pe care le-am citit (şi citat) eu, nu erau din domeniul lingvisticii computaționale. Lucrările domnului Marcus (în special cele din domeniul *limbajelor formale)* apăreau destul de frecvent între referințele bibliografice ale lucrărilor mele de la începutul anilor '80. Eram la început de drum, sursele documentare erau puține şi demersul era natural. Pe atunci, Chomsky era din nou foarte în vogă, noua sa teorie *Government and Binding* impulsionând o serie de cercetări în domeniul formalizării gramaticii universale. Tentația computațională față de această teorie a fost enormă, şi chiar dacă actualmente nu există nici o gramatică computațională efectivă a GB, idei fundamentale din GB se regăsesc în formalisme lingvistice computaționale moderne (cum ar fi HPSG).

> *Să fim bine înțeleşi. Nu noi avem nevoie de încă o citare pe lângă miile de citări deja acumulate, ci noile generații de studenți şi de cercetători au dreptul la o informare corectă asupra dezvoltării LM şi LC în general şi, în particular, asupra LM şi LC în România. DT a mai publicat, în urmă cu câțiva ani, un articol în care se schița o privire istorică asupra LC în România, cu câteva citări la întâmplare, care trădau necunoaşterea situației reale.*

Cu rezerve față de prima parte a paragrafului, mă opresc la grija domn
Academician pentru dreptul noilor generații de studenți şi de cercetători as
„informării corecte" asupra istoriei LM şi LC. Personal, cred că mult mai impo
pentru ei este să ştie prezentul şi tendințele viitoare ale domeniului. Astfe
cunoştințe le pot asigura un loc de muncă, o direcție de specializare, o ca
viitoare. Noile generații de studenți şi de cercetători sunt utilizatori pasiona
Internetului. Acest uriaş ocean informațional le asigură un imens volum
cunoştințe, începând cu cursuri on-line (obligatorii pentru profesori la mai
universitățile importante ale lumii), volume ale conferințelor sau articole extre
utile, recente şi mai puțin recente, cărți electronice. Chiar şi relevante lucră
istorie asupra diverselor domenii ştiințifice. Sistemele moderne de reg
documentară le asigură şi o ierarhizare a acestor surse de informare în rapo
relevanța şi cu interesul manifestat de alți cititori. Listele de discuții sau arhive
întrebări frecvente (FAQ) le pot oferi răspunsuri avizate şi obiective la întreb
ce-i preocupă. In anexă este furnizat un exemplu.

, în ultima parte a citatului de mai sus, domnul Academician Marcus a
în discuție o lucrare a mea din 1996 şi care arată că frustrările domniei sale
mai vechi. Articolul de care aminteşte domnul Academician mai sus, are
„Resurse lingvistice computaționale: trecut, prezent şi viitor" şi a apăr
volumul „Limbaj şi Tehnologie", Ed. Academiei, 1996. Cei interesați, pot
articolul respectiv în pagina oficială a RACAI *(http://www.racai.ro* secț
publicații). Iar cele *„câteva citări la întâmplare, care trădau necunoaşterea si
reale"* apar în capitolul 2. „Cercetări şi realizări româneşti în domeniul p
crării automate a limbajului natural". Cred că titlul volumului, al articolului
capitolui sunt lămuritoare pentru ceea ce discutam acolo, dar probabil fraza,
trimitea la un volum editat de domnul Marcus, „abordările statistice, revenite
în actualitate, au avut o tradiție strălucită (în România, adăugarea mea DT)"
prea scurtă şi insuficient de laudativă.

> *Mai este un aspect care cere o precizare. în conformitate c
> specificul volumului în care apare articolul [1], DT face numeroas
> referiri la acte şi documente ale unor organisme europene ş
> internaționale, cum este şi firesc, pentru a nu mai vorbi de aspectu
> financiar al colaborării cu organismele respective.*
> *Această situație a existat de la începutul LM şi LC (chiar dacă nu
> avut amploarea de azi), datorită faptului că LM şi LC au apărut şi c
> urmare a unor comandamente sociale, privind precaritate
> mijloacelor de prelucrare a informației. îmi amintesc de faimoase
> Rapoarte CETIS care veneau de la EURATOM, Bruxelles, pe tem
> legate de analiza şi prelucrarea automată a limbajului, traducer
> automată şi documentare automată. în USA, diferite corporații (cu
> ar fi RAND Corporation, Santa Monica, Calif.) finanțau cercetă
> similare. O întâlnire semnificativă a fost aceea din 1962, organiza*

de *"NATO Advanced Summer Institute", la Veneția, Italia, privind traducerea automată. De numele acestui Institut este legat un document care a marcat evoluția cercetărilor de traducere automată: seria de expuneri prezentate de Y. Bar-Hillel [17]. în legătură cu aceste activități dirijate şi finanțate de diferite organisme europene şi internaționale, trebuie să observăm că cei implicați au avut înțelepciunea şi priceperea necesare pentru a nu reduce proiectele respective la dimensiunea lor exclusiv utilitară, ci de a o subordona pe aceasta unei perspective mai ample, care lua în considerare orizontul ştiințific real al problemelor. Pentru a da un prim exemplu, mă voi referi la faptul că mai multe rapoarte CETIS au pus în discuție un concept care, născut din experimentele de traducere automată, avea să se dovedească de o deosebită semnificație pentru teoria sintactică în toată generalitatea sa; este vorba de conceptul de proiectivitate sintactică, cu consecințe bogate în studiul structurilor arborescente şi al gramaticilor de dependență. Azi putem spune că şi sintaxa limbajului natural şi teoria matematică a grafurilor au profitat esențial de conceptul respectiv (folosit până şi de Rene Thom, în probleme de morfogeneză [17]). Această expansiune a unui concept sau rezultat dincolo de motivația sa inițială este testul cel mai convingător al interesului său. Un al doilea exemplu se referă la titlul provocator folosit de Bar-Hillel pentru expunerile sale: "Patru conferințe despre lingvistica algebrică si traducerea automată".*

Simpla alăturare a celor două sintagme, una foarte teoretică, cealaltă aparent tehnologică, avea menirea să-i avertizeze pe cei care presau să se obțină cât mai repede rezultate practice asupra faptului că proiectele de traducere automată nu se pot finaliza de azi pe mâine, ci au nevoie de un lung itinerar lingvistic, matematic şi computațional. Acum ştim că acest itinerar continuă şi azi, cu tatonări şi reveniri, şi, chiar dacă nu a dus încă la rezultatele visate, a impulsionat în mod esențial cercetările de AI, cu consecințe benefice pentru aspectele logice şi semantice ale limbajului natural, întrebarea pe care ne-o punem, dar o lăsăm deocamdată fără răspuns, deoarece nu suntem pregătiți pentru a-l da, este următoarea: Nu cumva aspectele pe care le-am criticat mai sus sunt consecința unui fenomen mai general, acela al unui orizont insuficient de cuprinzător, al unei prea mari dependențe de factori utilitari imediați? Ştiința a oscilat mereu între cognitiv şi utilitar, dar istoria arată că funcția utilitară s-a manifestat în toată profunzimea ei atunci când ea a fost fructul unei evoluții fireşti a funcției cognitive, evoluție care poate fi de doi ani, de 20 de ani, de 200 sau de 2000 de ani. Cu un ochi îndreptat spre comisiile europene, suntem obligați totuşi să ținem treaz şi celălalt ochi, îndreptat spre ceea ce se întâmplă pe scena cercetării ştiințifice vii, aşa cum apare ea în revistele de specialitate şi la întâlnirile ştiințifice de profil.

Remarcile de mai sus îmi sugerează celebra fabulă cu strugurii ce
Cercetarea instituționalizată (în opoziție cu cea „de dragul artei") are m
întotdeauna justificabile. Organismele de finanțare a cercetării, național
internaționale, nu fac desigur acte de caritate. Obținerea unei finanțări pen
proiect de cercetare nu este la îndemâna oricui şi el implică nu numai abo
unei probleme importante, dar şi credibilitatea grupului de cercetare. Eva
propunerilor de proiecte se face de către experți recunoscuți în domeniul res
angajați şi plătiți de agențiile de finanțare a cercetării. în condițiile unei conc
internaționale acerbe pentru fondurile (din păcate prea mici) destinate cerce
lua în derâdere, invocând caracterul utilitar, cercetările ce obțin concu
finanțarea arată o desprindere de realitate. în luna martie a.c. am partic
evaluarea propunerilor de proiecte europene din cadrul Programului Ca
(apelul 8), şi în calitate de raportor al direcției **„11.1.1** • Exploratory
Risk/Long Term Research", pot să afirm că propunerile de proiecte pe c
am văzut erau foarte departe de a avea caracter utilitar. Domnul Acade
Marcus lasă fără răspuns o întrebare cu răspuns sugerat, ridicând o pro
discutată cu ceva timp în urmă, anume a tipului de cunoaştere contemp
enciclopedică (şi inerent generalistă) sau specializată. Cel puțin în do
tehnologice, viteza fără precedent a apariției de cunoştințe noi face imp
cunoaşterea enciclopedică şi în acelaşi timp expertă pe toată lărgimea spe
cunoaşterii actuale chiar şi într-un domeniu aparent îngust. Tehnologia lim
este actualmente termenul ce subsuma toate preocupările legate de prelu
automată a limbajului natural. Cred că acest lucru spune totul!

3. în loc de concluzii

Ajungând în acest punct al răspunsului meu la atacul do
Academician Marcus mărturisesc că mă încearcă un apăsător sentim
deşertăciunii. Nu am dorit această polemică şi în nici un caz în acest c
Considerând că ea este nepotrivită față de obiectivele urmărite de proiectul
Soluții şi strategii în România", în calitatea mea de director de proiect şi coe
volumului de față, am discutat cu membrii comitetului director al pro
oportunitatea publicării polemicii domnului Academician Marcus (şi im
răspunsului meu) în volumul destinat unor probleme tehnice. Părerea
unanimă că nu este cazul să amestecăm obiectivele proiectului cu discuția
Dar transmițând domnului Academician această opinie şi făcându-i propune
a găzdui această polemică pe internet (în pagina oficială a RACAI) domnia
simțit cenzurat, insultat şi îndreptățit să facă o serie de afirmații pe care m
să le comentez. Decizia de includere a acestei secțiuni în volumul de față a
o fără plăcere pentru că pe de o parte, în ciuda părerii domnului Acade

Marcus *(Articolul meu se încadrează perfect în obiectivul pe care pretindeți că-/ urmăriți şi în acest spirit a fost conceput Realizați gravitatea deciziei Dv?* - de a nu-l include în volum, precizarea mea, D.T.) continui să cred că nici articolul domniei sale nici al meu nu îşi aveau rostul aici. Pe de altă parte, nu pot decât să deplâng supărarea pe care i-am provocat-o fără voie domnului Marcus şi risipa de energie pe care o depune într-o problemă care din punctul meu de vedere nu există. Drept care, sperând că includerea articolului ce *se încadrează perfect în obiectivul...*îi va da domnului Academician satisfacția pe care şi-a dorit-o, las cititorii să aprecieze cât de grav ar fi fost pentru obiectivul tehnologiei limbii române în contextul „Societatea Informațională - Societatea Cunoaşterii: Soluții şi strategii în România" ca cele două articole să nu fi apărut aici.

Referințe bibliografice (secțiune din lucrarea domnului Academician Marcus):

[1] D. Tufiş. *Promovarea limbii române în SI-SC.* în *Societatea Informațională - Societatea cunoaşterii* (coord. F. Gh. Filip). Ed. Expert, Bucureşti, 2001,131-142.

[2] D. G. Hays. *The field and scope of computațional linguistics.* Papers in Computațional Linguistics (eds. F. Papp, G. Szepe). Proceedings of the Third International Meeting of Computațional Linguistics, held in Debrecen, Hungary, 1971. Akademiai Kiado, Budapest, 1976, 21-26.

[3] D. G. Hays (ed.). *Readings in Automatic Language Processing,* American Elsevier, New York, 1967.

[4] S. Marcus. *Mathematical Linguistics in Europe. Current Trends in Linguistics* (Th. A. Sebeok, ed.), vol.9, Mouton, The Hague, 1972, 646-687.

[5] S. Marcus. *Mathematique et Linguistique.* în *Mathematique, Informatique et Sciences Humaines,* Paris, 26, 1988,103, 7-21.

[6] S. Marcus. *The status of research in the field of analytical algebraic models of language.* în *Current Issues in Mathematical Linguistics* (C. Martin-Vide, ed.). Elsevier-North Holland, Amsterdam, 1994, 3-21.

[7] S. Marcus. *Lingvistica matematică^ azi.* în *Matematica în lumea de azi şi de mâine* (C. Iacob, coord.), Editura Academiei, Bucureşti, 1985, 182-186.

[8] S. Marcus. *Recent Romanian investigations in the field of mathematical and computațional linguistics.* Avtomaticeskaja Obrabotka Tekstov, Matern. Fyz. Fakulta, KL Praha, 1973,15-42.

[9] S. Marcus. *Mathematical and computațional linguistics.* în *Current Trends in Romanian Linguistics* (A. Rosetti, S. Golopentia Eretescu, eds.). Revue Roumaine de Linguistique 23, 1978, 1-4, 559-588.

[10] S. Marcus, C. Martin-Vide, G. Paun. *Contextual grammars as generative models of natural languages.* Computațional Linguistics 24, 1998, 2, 245-274.

[11] S. Marcus. *Semiotics and formal artificial languages.* în *Encyclopedi Computer Science and Technology* (A. Kent, J.C.Williams, eds.) 29, Marcel Dekker, New York, 1994, 393-405; also in Encycloped Microcomputers (A. Kent, J.C.Williams, eds.) 15, 1995, 299-312.

[12] S. Marcus. *Contextual grammars and natural languages. Handbook of F Languages* (G. Rozenberg, A. Salomaa, eds.), 2, Springer, Berlin, New 1997,215-235.

[13] S. Marcus, G. Martin-Vide, G. Paun. *A new-old class of linguis motivated regulated grammars. Computațional Linguistics in the Nethe 2000* (W. Daelemans et al., eds.), Selected Papers from the Eleventh Meeting, Ed. Rodopi, Amsterdam, New York, 2001, 111-125.

[14] B. H. Partee, A. Ter Meulen, R. Wall. *Mathematical Methods in Lingu* Kluwer, Dordrecht, 1990.

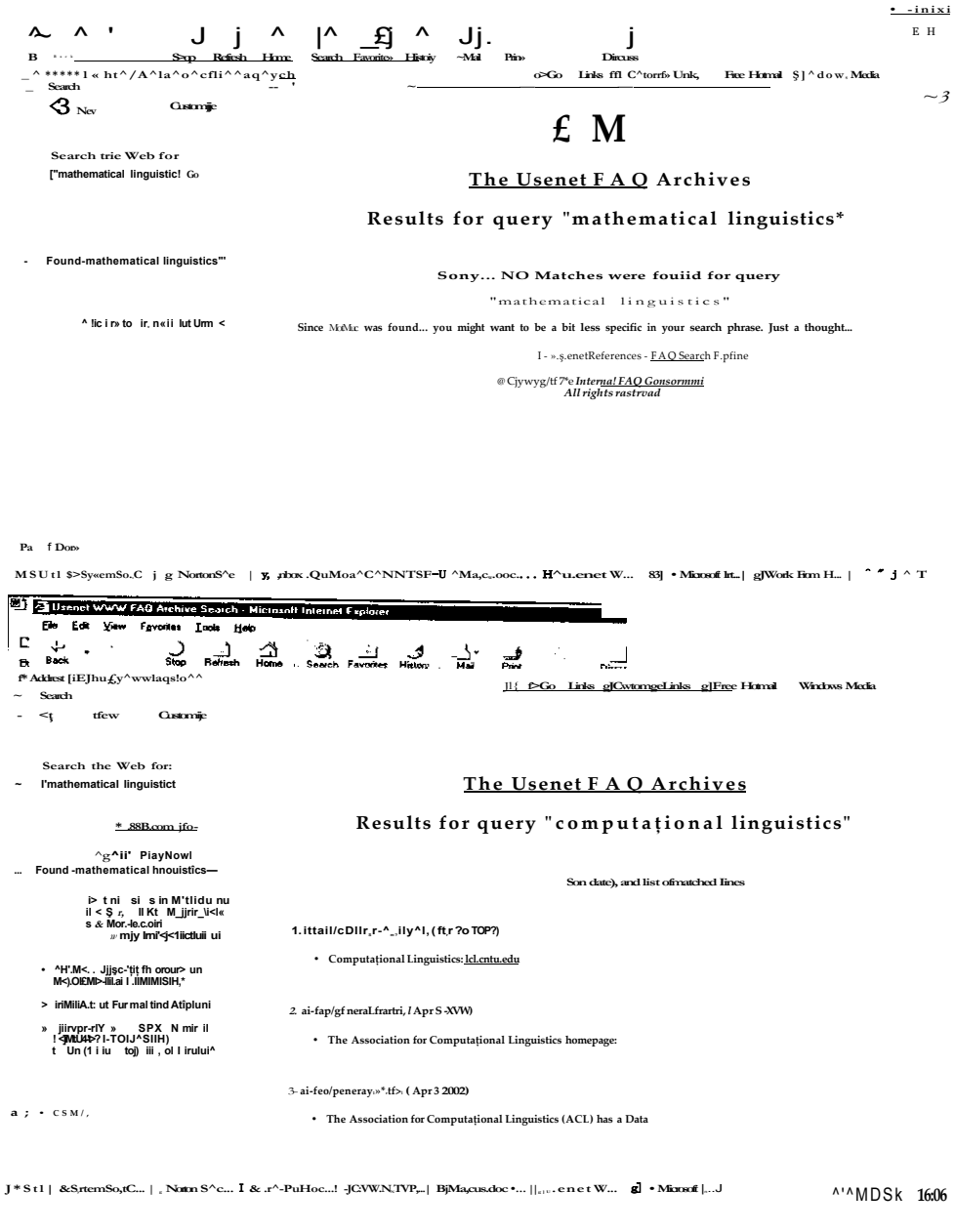
[15] E. F. Beckenbach, Ch. B. Tompkins (eds.). *Concepts of Communi Interpersonal,* Intrapersonal and Mathematical. John Wiley and Sons, York, 1976.

[16] D. G. Hays. *Introduction to Computațional Linguistics.* American Elsevier, York, 1967.

[17] R. Thom. *Stabilite Structurelle et Morphogenese.* John Benjamins, New 1970.

[18] Y. Bar-Hillel. *Four Lectures on Algebraic Linguistics and Machine Tran* revised version of a series of lectures given in July 1962, before a N Advanced Summer Institute, Venezia, Italy.

ANEXAI: Exemple de căutare într-o arhivă de întrebări frecvente (Usenet FAQ)



ANEXA 2: Definiţii

What is Mathematical Linguistics?

MATHEMATICAL LINGUISTICS is the study of mathematical str and methods that are of importance to linguistics. As in other branches of mathematics, the influence of the empirical subject matter is somewhat i theorems are often proved more for their inherent mathematical value than applicability.

Both in phonology/morphology and in syntax/semantics the ch linguistic formalism is to some extent influenced by considerations that go the primary issue of descriptive adequacy. One important issue is Reco Complexity. This concerns the complexity of the decision problem for mem in a language: it is assumed that a grammatical theory should have the pro guaranteeing that there is some reasonably rapid (polynomial in the lengh input) computation that will answer the question of whether a given seque words is a grammatical expression according to a given grammar. Human certainly do much more than this when they listen to an utterance and fig the meaning of what was said, so a grammatical theory that canno guarantee reasonably rapid confirmation of well-formedness is proba psycholinguistically realistic. Another one is Learnability, which concern sorts of mathematically definable procedures could in principie correctly gu grammars for languages.

(Geoffrey K. Pullum and Andras

What is Computaţional Linguistics?

Simply put, COMPUTAŢIONAL LINGUISTICS is the scientific s language from a computaţional perspective. Computaţional linguists are int in providing computaţional models of various kinds of linguistic phenomena models may be "knowledge-based" ("hand-crafted") or "data-driven" ("statis "empirical"). Work in computaţional linguistics is in some cases motivated scientific perspective in that one is trying to provide a computaţional explana a particular linguistic or psycholinguistic phenomenon; and in other ca motivation may be more purely technological in that one wants to pr working component of a speech or natural language system. Indeed, the computaţional linguists is incorporated into many working systems today, i speech recognition systems, text-to-speech synthesizers, automated response systems, web search engines, text editors, language ins materials, to name just a few.

(Copyright © 2000, The Association for Computaţional Lin

LIMBA ROMANA
în
Societatea Informațională
Societatea Cunoașterii

Acest volum este dedicat Academicianului Mihai Drăgănescu, Profesorul şi mentorul unei întregi generații de specialişti în ştiința şi tehnologia informației în general şi al problemelor societății informaționale şi a cunoaşterii în special. Marea majoritate a contribuțiilor din acest volum aparțin unor experți ce fac parte din Comisia de Informatizare a Limbii Române, comisie a Academiei Române la a cărei naştere un rol esențial l-a avut Profesorul Drăgănescu, preşedintele Secției de Ştiința şi Tehnologia Informației. Savantul Mihai Drăgănescu are numeroase contribuții în ştiința contemporană, binecunoscute atât în țară cât şi în străinătate. Pentru cine îl cunoaşte pare incredibilă puterea sa de muncă, debordanta creativitate şi neostoita căutare a noului. Profesorul Drăgănescu este indiscutabil port-drapelul conceptului de societate informațională-societate a cunoaşterii în România. în lucrările sale din urmă cu peste 25-30 de ani se regăsesc cu claritate multe concepte foarte actuale în zilele noastre, previziuni curajoase atunci, acum realități cotidiene. în lucrările domniei sale din ultima vreme, apare un nou concept ce avem convingerea că se va impune: Societatea Conştiinței, o treaptă superioară a societății cunoaşterii. Nu este de mirare deci că în contextul societății informaționale şi a cunoaşterii profesorul Drăgănescu a susținut cu consecvență şi a afirmat cu claritate rolul Inteligenței Artificiale în devenirea noilor societăți ale cunoaşterii. între domeniile Inteligenței Artificiale un loc de frunte în promovarea principiilor societății cunoaşterii îi revine Tehnologiei Limbajului Natural. Profesorul Drăgănescu a fost unul dintre puținii oameni de ştiință români care au înțeles şi au sprijin total aceste direcții. Cu aproape douăzeci de ani în urmă (1983), Profesorul Drăgănescu edita (împreună cu Adrian Davidoviciu şi Ioan Georgescu) volumul "Inteligența Artificială şi Robotica" pentru ca trei ani mai târziu (împreună cu Corneliu Burileanu) să editeze un alt volum de referință "Analiza şi sinteza semnalului vocal". Astăzi, cercetările mondiale în domeniul tehnologilor lingvistice au atins un nivel de maturitate ce permit sinergizarea eforturilor lingviştilor, informaticienilor, matematicienilor şi a altor specialişti din sectorul academic sau industrial, să abordeze proiecte mari, interdisciplinare având ca obiectiv prelucrarea automată, în mediile de comunicare electronică, a din ce în ce mai multe limbi naturale. Printre acestea, limba română îşi face loc încet dar sigur. Volumul de față este o mărturie în acest sens. în acelaşi timp, volumul se constituie într-o nouă confirmare a realităților pe care Profesorul Mihai Drăgănescu le prefigura cu mulți ani în urmă.

ISBN 973-8177-83-9