Dan Tufis, Florin Gh. Filip (coordonatori)

Limba Româna
în
Societatea Informationala - Societatea Cunoasterii

EDITURA Expert

© EDITURA Expert

Bucuresti, România
Editor: **Valeriu IOAN-FRANC**
Coperta si prezentarea grafica: **Nicolae LOGIN**, **Luminita LOGIN**

ISBN 973- 8177- Aparut 2002

Dan Tufis, Florin Gh. Filip (coordonatori)

Limba Română
în
Societatea Informațională -
Societatea Cunoaşterii

EDITURA Expert

DEDICATIE

Acest volum este dedicat Academicianului Mihai Draganescu, Profesorul si mentorul unei întregi generatii de specialisti în stiinta si tehnologia informatiei în general si al problemelor societatii informationale si a cunoasterii în special. Marea majoritate a contributiilor din acest volum apartin unor experti ce fac parte din Comisia de Informatizare a Limbii Române, comisie a Academiei Române la a carei nastere un rol esential l-a avut Profesorul Draganescu, presedintele Sectiei de Stiinta si Tehnologia Informatiei. Savantul Mihai Draganescu are numeroase contributii în stiinta contemporana, binecunoscute atât în tara cât si în strainatate. Pentru cine îl cunoaste pare incredibila puterea sa de munca, debordanta creativitate si neostoita cautare a noului. Profesorul Draganescu este indiscutabil port-drapelul conceptului de societate informationala-societate a cunoasterii în România. În lucrarile sale din urma cu peste 25-30 de ani se regasesc cu claritate multe concepte foarte actuale în zilele noastre, previziuni curajoase atunci, acum realitati cotidiene. În lucrarile domniei sale din ultima vreme, apare un nou concept ce avem convingerea ca se va impune: Societatea Constiintei, o treapta superioara a societatii cunoasterii. Nu este de mirare deci ca în contextul societatii informationale si a cunoasterii profesorul Draganescu a sustinut cu consecventa si a afirmat cu claritate rolul Inteligentei Artificiale în devenirea noilor societati ale cunoasterii. Între domeniile Inteligentei Artificiale un loc de frunte în promovarea principiilor societatii cunoasterii îi revine Tehnologiei Limbajului Natural. Profesorul Draganescu a fost unul dintre putinii oameni de stiinta români care au înteles si au sprijin total aceste directii. Cu aproape douazeci de ani în urma (1983), Profesorul Draganescu edita (împreuna cu Adrian Davidoviciu si Ioan Georgescu) volumul „Inteligenta Artificiala si Robotica" pentru ca trei ani mai târziu (împreuna cu Corneliu Burileanu) sa editeze un alt volum de referinta „Analiza si sinteza semnalului vocal". Astazi, cercetarile mondiale în domeniul tehnologilor lingvistice au atins un nivel de maturitate ce permit sinergizarea eforturilor lingvistilor, informaticienilor, matematicienilor si a altor specialisti din sectorul academic sau industrial, sa abordeze proiecte mari, interdisciplinare având ca obiectiv prelucrarea automata, în mediile de comunicare electronica, a din ce în ce mai multe limbi naturale. Printre acestea, limba româna îsi face loc încet dar sigur. Volumul de fata este o marturie în acest sens. În acelasi timp, volumul se constituie într-o noua confirmare a realitatilor pe care Profesorul Mihai Draganescu le prefigura cu multi ani în urma.

Dr. Dan Tufis, m.c.A.R, Acad. Florin Gh. Filip

CUPRINS

SECTIUNEA III:
TEHNOLOGII ALE LIMBAJULUI VORBIT

SECTIUNEA IV:
DEZBATERI SI DISCUTII

INTRODUCERE

Programul de cercetare aplicativa „Strategii si solutii pentru Societatea Informationala – Societatea Cunoasterii în România (SI-SC), din subprogramul A-strategic, al Programului National INFOSOC a avut ca principale obiective stabilirea unui program de veghe conceptuala pentru mentinerea pe linia tendintelor mondiale ale avansului SI-SC, sensibilizarea factorilor de decizie si a publicului larg, crearea unui cadru de reflectie prospectiva pe temele prioritare ale SI-SC: economice, sociale, culturale, tehnologice, ambientale, precum si operationalizarea unor solutii de interes prioritar pe plan national. În cadrul acestui proiect a fost elaborat volumul „*Societatea Informationala – Societatea Cunoasterii. Concepte, solutii si strategii pentru România"* (publicat la Ed. Expert in anul 2000), realizat sub coordonarea Academicianului Florin Gheorghe Filip.Acest volum avea ca scop construirea unei viziuni si continea o serie de studii si cercetari care au aprofundat rezultatele programului prioritar al Academiei Române privind *Societatea Informationala – Societatea Cunoasterii* si au identificat o serie de orientari strategice cerute de sustinerea unei dezvoltari de tip "salt" a SI-SC în România. Prin prisma obiectivelor proiectului, au fost analizate principalele aspecte conceptuale ale SI-SC, probleme legate de infrastructurile informatice si de comunicatii ale SI-SC, formarea profesionala si pregatirea generala a populatiei în si pentru SI-SC, rolul stiintei, cercetarii si inovarii, aspecte sociale si juridice, institutiile statului si relatia lor cu cetateanul, dezvoltarea economiei si afacerilor, dimensiunea culturala a SI-SC, actorii sociali ai crearii si difuzarii tehnologiei informatiei si comunicatiilor în contextul SI-SC. Studiile tematice, ancheta Delphi pentru consultarea opiniei expertilor privind tendintele globale si optiunile posibile de raportare la ele, scenariile de evolutie elaborate au sustinut functia prospectiva a proiectului.

Functia operativa a acestui proiect, respectiv identificarea de solutii tehnice privind rezolvarea principalelor prioritati identificate în faza analizei prospective urma sa se manifeste în perioada imediat urmatoare, printr-o dintr-o serie de cercetari/dezvoltari tehnologice ce vor trata pe larg problematica specifica a fiecaruia dintre directiile amintite anterior. Aceasta serie este deschisa prin prezentul volum ce înglobeaza contributii ale unor specialisti români reprezentativi în domeniul prelucrarii automate a limbajului natural si a resurselor lingvistice necesare utilizarii limbii române în mediile de comunicare electronica.

În [1] este definit conceptul de "Societate Informationala – Societate a Cunoasterii" (SI-SC) precum si principalii sai vectori tehnologici si functionali. În acest context „internetul dezvoltat" (ca vector tehnologic) si "managementul utilizarii morale a cunoasterii la nivel global" (ca vector functional) sunt prezentati ca factori motrici esentiali ai Societatii Cunoasterii, si în perspectiva, a Societatii Constiintei. „Din momentul în care intervine Internetul cu marile avantaje pe care acesta le aduce (e-mail, comert electronic si tranzactii electronice, piata Internet, distributia de 'continut') prin cuprinderea în sfera informatiei electronice a unui numar cât mai mare de cetateni se trece la societatea

informationala. Cunoasterea este informatie cu înteles si informatie care actioneaza. De aceea societatea cunoasterii nu este posibila decât grefata pe societatea informationala si nu poate fi separata de aceasta. În acelasi timp, ea este mai mult decât societatea informationala prin rolul major care revine informatiei–cunoastere în societate." [1]

În 1984, William Gibson, un dizident cognitiv - dupa cum se auto-caracterizeaza, publica volumul SF „*Neuromancer*" (Ace Book, July 1984, ISBN: 0-441-56959-5), carte care pe lânga o multime de premii literare i-a adus notorietatea si 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 facut cariera, actualmente fiind o notiune care din punct de vedere tehnic subsuma conceptul „Internet"(scris cu majuscula): "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 "i") 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/tg2k / _cyberspace.html, 05.08.2002).

Dimensiunea tehnica (evocata mai sus) a notiunii de „*ciberspatiu"* este complementata de dimensiunea socio-culturala si din aceasta perspectiva problemele „satului global" previzionat de Societatea Informationala – Societatea Cunoasterii. Idea atenuarii schismei dintre specialistii din domeniul tehnic si cei din zona stiintelor umaniste in contextul SI-SC este sustinuta puternic si de M. Derouzos [5] ,cel care a propus conceptul de "piata informationala", pe care il considera mai realist decat cel de "ciberspatiu". De altfel , dimensiunile socio culturale ale SI-SC au fost evocate in capitolele 2,3,4 si 6 ale volumului „*Societatea Informationala – Societatea Cunoasterii. Concepte, solutii si strategii pentru România"*

Printre componentele socio-culturale ale SI-SC, utilizarea limbii materne si a accesului universal la ciberspatiu [2, 3, 4] constitue o prioritate .

În contextul actual, al comunicarii mediate de tehnologia informatiei si de telecomunicatii, limba devine obiect al investigatiei tehnice. Tehnologia limbajului impune metodologii specifice de cercetare/dezvoltare, dezvoltarea sau adaptarea resurselor lingvistice fundamentale cum ar fi dictionarele, tezaurele, corpusurile si gramaticile computerizate, în conformitate cu standardele sau recomandarile existente. În functie de resursele lingvistice disponibile, de volumul si calitatea lor, de compatibilitatea codificarii lor în raport cu recomandarile si standardele internationale etc., se poate vorbi de *nivelul de tehnologizare* al unei limbi naturale. Nivelul de tehnologizare al unei limbi naturale este în corespondenta directa cu statutul de limba de *circulatie electronica.* Aceasta sintagma, o parafraza la expresia *limba de circulatie internationala*, încearca sa elimine antinomia, pe

cât de cunoscuta pe atât de goala în continut spiritual si cultural, „limbi mari/limbi mici". Conceptul de „limba de circulatie electronica", pe lânga semnificatia lui directa, are profunde implicatii culturale, sociale si nu în ultimul rând economice implicând dreptul fiecarui cetatean de a avea acces în propria limba la cunostintele, informatiile si serviciile ciberspatiului.

Promovarea limbii române în SI-SC presupune informatizarea limbii române ca factor infrastructural fundamental (vector functional) si precum si stimularea utilizarii curente (prin vectori tehnologici) a limbii române în utilizarea tehnologiilor si a serviciilor informatice. Acestui obiectiv presupune un eforturi umane si materiale substantiale si de dimensionarea lor se leaga orizontul de timp al realizarii sale.

Volumul de fata reuneste lucrari ce trateaza aspecte specifice prelucrarii limbajului natural, în marea lor majoritate cu aplecare directa asupra limbii române. Inerent, volumul de fata nu poate acoperi întreaga arie problematica a domeniului dupa cum nici reprezentarea specialistilor români în domeniul tehnologiei limbajului nu este completa, dar cititorul va gasi un larg evantai de directii de cercetare, în care specialistii români au obtinut rezultate importante.

Volumul este structurat în patru parti (aspecte teoretice si probleme de terminologie, prelucrarea limbajului scris, prelucrarea limbajului vorbit, dezbateri si discutii) care pot fi citite în mod independent, în functie de interesul specific al cititorului.

Prima parte „Lingvistica teoretica si formala; lexicografie" cuprinde 7 lucrari din domeniul lexicografiei, sintaxei si terminologiei.

În lucrarea „Resurse lingvistice elaborate la Institutul de Lingvistica «*Iorgu Iordan*»" Ioana Vintila Radulescu face o trecere în revista a celor mai importante resurse lingvistice realizate în cei peste 50 de ani de activitate la Institutul de Lingvistica «*Iorgu Iordan*».

Angela Bidu-Vrânceanu prezinta în lucrarea „Contributia lingvisticii la studiul terminologiilor stiintifice" concluziile a trei contracte de cercetare stiintifica având ca obiect studiul terminologic al limbajului folosit în diverse domenii (matematica, filozofie, mineralogie, arte plastice).

Articolul „Gramaticile nontransformationale" al lui Emil Ionescu face o prezentare generala a gramaticilor bazate pe unificare si constrângeri precum si a principalelor realizari, în contextul acestei paradigme, în cercetarea lingvistica din România.

Niculai Curteanu propune în lucrarea „Catre o teorie X-bar functionala" o reconsiderare a teoriei clasice X-bar prin perspectiva modelului propriu SCD (Segmentare - Coeziune-Dependenta).

Ana-Maria Barbu prezinta în lucrarea sa „Teoria HPSG: studiu de caz: acordul încrucisat" principalele caracteristice ale teoriei HPSG si discuta în acest context un caz de dependenta încrucisata specific limbii române, respectiv clauzele relative în care pronumele relativ este precedat de articolul genitival.

O serie de probleme legate de terminologia computationala sunt prezentate în ultimele doua lucrari ale primei sectiuni. În articolul „ Dupa 10 ani de experienta terminologica: noul model de date terminologice al TERMROM" Dan Matei prezinta modelul dezvoltat în conformitate cu noile tendinte si standarde în domeniu si adoptat de Asociatia Româna de Terminologie – TERMROM.

Lucrarea lui Sorin Getaru „Probleme de reprezentare a datelor terminografice într-o baza de date relationala" aduce în discutie aspecte specifice reprezentarilor standardizate necesare realizarii dezideratului de interschimb si interoperabilitate intre diverse tezaure terminologice si discuta elementele distinctive ale standardului ISO-12200 MARTIF (**Ma**chine-**R**eadable **T**erminology **I**nterchange **F**ormat).

Sectiunea a doua a volumului („Tehnologii ale limbajului scris") este deschisa de lucrarea lui Dan Tufis si Dan Cristea „RO-BALKANET – ontologie lexicalizata în context multilingv pentru limba româna" care descrie stadiul dezvoltarii unui dictionar, pentru limba româna, structurat ca o retea semantica, de tip EuroWordNet, rezultat al unui program european ce-si propune extensia EuroWordnet (în prezent implementat pentru 10 limbi europene) cu înca 5 limbi.

Articolul lui Dan Gâlea, Niculai Curteanu si Constantin Lintes „Algoritmi de segmentare a textului în unitati de tip clauzal" trateaza o problema delicata a prelucrarii limbajului natural, respectiv cea a identificarii, în raport cu un anumit criteriu functional, a structurilor „clauzale" si prezinta contrastiv doi algoritmi diferiti (unul dintre ei apartinând autorilor), atât prin prisma modelarii lingvistice cât si al performantei computationale.

Rada Mihalcea si Vivi Nastase prezinta în articolul lor o metoda de inserare automata a caracterelor diacritice în texte scrise (cu studiu de caz pentru limba româna) fara diacritice si comenteaza rezultatele proprii în comparatie cu cele ale altor metode dezvoltate pentru rezolvarea aceleiasi probleme.

Adriana Vlad si Adrian Mitrea prezinta în lucrarea lor „Contributii privind structura statistica de cuvinte în limba româna scrisa" rezultate recente în caracterizarea statistica a limbii române scrise prin aproximarea ei ca un lant Markov ergotic multiplu cu ordin de multiplicitate mai mare decât 30, rezultate obtinute prin analiza riguroasa a unui corpus foarte mare de texte.

Articolul „Dezambiguizarea semantica automata în corpusuri paralele" al lui Dan Tufis prezinta o alternativa la spinoasa problema a dezambiguizarii cuvintelor polisemantice, bazâdu-se pe extragerea cunostintele implicite existente într-un corpus multilingv (creat de traducatori profesionisti) si apelând la tehnici si euristici ale linvisticii corpusului.

Dan Cristea prezinta în articolul „Referentialitate si cursivitate în structura discursului" elementele definitorii ale teoriei sale asupra structurii discursive a textelor (teoria nervurilor) si îsi exemplifica argumentatia prin analiza dihotomiilor structura-referentialitate si structura-coerenta.

În lucrarea „DLIR - un sistem de cautare documentara multilingv" Amalia Todirascu prezinta o abordare bazata pe logici terminologice, ontologii si tehnici de

prelucrare a corpusurilor în implementarea unui sistem de regasire documentara bilingv (romtna si franceza).

Partea a doua a volumului se încheie cu articolul lui Stefan Trausan-Matu „Mediu hermenofor pentru asistarea învatarii unor concepte într-o limba straina" care dupa o prezentare a notiunilor cu care opereaza în lucrare, descrie un modul de prelucrare a metaforelor utilizate în limbaje specializate (studiu de caz: limbajul financiar) incorporat într-un sistem de instruire inteligenta în învatarea conceptelor într-o limba straina, sistem distribuit dezvoltat în cadrul unui proiect european.

Sectiunea a treia a volumului este dedicata problemelor de prelucrare a vorbirii. Corneliu Burileanu si Luigi Bojan se opresc asupra tehnicilor de recunoastere a vorbitorului ca etapa distincta si strict necesara pentru recunoasterea automata a vorbirii si prezinta o parte a rezultatelor obtinute de catre autori.

Lucrarea lui Dragos Burileanu „Prelucrarea initiala a textului de intrare în cadrul unui sistem de sinteza a vorbirii pornind de la text în limba româna" abordeaza problemele sintezei limbajului vorbit pornind de la un text în format electronic si detaliaza etapa de preprocesare a textului ca etapa primara în procesul transformarii sale în semnal vocal inteligibil si coerent.

Tot în domeniul sintezei vorbirii se plaseaza si lucrarea lui Horia Nicolai Teodorescu „Utilizarea tehnicilor nuantate (fuzzy) ?i de dinamic? neliniar? pentru sinteza adaptiv? a vorbirii" ce subliniaz? rolul esen?ial al prozodiei ?i al model?rii sale algoritmice ?n realizarea unor sinteze vocale de calitate, purt?toare de informa?ie emo?ional?.

Un proiect de anvergura, este prezentat de Dumitru Todoroi, Diana Micusa, Zinaida Todoroi, Ion Linga, Ion Covalenco, Nicolae Objeleanu, Stefan Spataru, Stela Lungu, Virginia Turcanu, Elana Cozlov, Nadejda Ambrozii, Victor Slobodeanu, Igor Coseru si Catalina Suruceanu în lucrarea „Dictionarele multimedia ale limbii române. Secvente de implementari si experimentari ".

Sectiunea a treia a volumului se încheie cu lucrarea elaborata de Silviu Bejinariu, Vasile Apopei si Mariana Roman „Mediu pentru editarea transcrierilor fonetice în Limba Româna. Realizarea Atlasului Lingvistic Român pe Regiuni" ce prezinta un instrument ce permite realizarea facila a transcrierilor fonetice într-un limbaj standardizat (IPA), ofera extensii specifice de adnotare fonetica (realizate pâna acum manual) si prefigureaza realizarea variantei computerizate a atlaselor lingvistice românesti.

Ultima sectiune a volumului (Dezbateri si discutii) contine trei contributii. Prima dintre ele, elaborata de Academician Mihai Draganescu, „Asupra a doi vectori functionali ai Societatii Cunoasterii: Managementul Cunoasterii si Învatarea Electronica. Cultura si Societatea Cunoasterii" reprezinta liantul dintre volumul precedent (*Societatea Informationala – Societatea Cunoasterii. Concepte, solutii si strategii pentru România*, coordonator Academician Fl. Gh. Filip) si volumul de fata, rafinând clasificarea din lucrarea anterioara si adâncind o serie de probleme ridicate în [1].

Ultimele doua contributii reprezinta doua puncte de vedere asupra problematicii prelucrarii limbajului natural, prima pozitie „Între lingvistica matematica si cea

computationala" fiind sustinuta de Academician Solomon Marcus, iar cea de a doua „Între lingvistica matematica si cea computationala: o alta perspectiva" fiind prezentata de Dan Tufis.

Multumiri

Coordonatorii acestui volum, multumesc tuturor celor care au participat la realizarea proiectului „Strategii si solutii pentru societatea informationala-societatea cunoasterii în România" derulat cadrul programului national INFOSOC. Multumiri speciale se cuvin directorului programului INFOSOC, Profesor Doina Banciu, care a sustinut si a manifestat un interes deosebit fata de desfasurarea acestui proiect.

Referinte bibliografice

[1] M. Draganescu "Societatea informationala si a cunoasterii. Vectorii societatii cunoasterii" în *F. G. Filip (coord.) Societatea Informationala – Societatea Cunoasterii. Concepte, solutii si strategii pentru România*. Academia Româna, Editura Expert, ISBN 973-8177-42-1, 2001, pp. 43-112

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

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

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

[5].M. Dertouzos. "What It will Be ". Harper Edge. New York,1977(trad. In lb. Romana "Ce va fi", ed. Tehnica Bucuresti, 2000).

SECTIUNEA I

LINGVISTICA TEORETICA SI FORMALA; LEXICOGRAFIE

Resurse lingvistice pentru limba româna elaborate la Institutul de Lingvistica „Iorgu Iordan”

Ioana VINTILA-RADULESCU
Institutul de Lingvistica „Iorgu Iordan – Al. Rosetti”
Bucuresti, Calea 13 Septembrie 13
e-mail: ioanar@fx.ro

1. Consideratii generale

Întelegând prin *resursa* în general o „rezerva sau sursa de mijloace (materiale sau spirituale) susceptibile de a fi valorificate într-o împrejurare data”[1], întelegem prin *resurse lingvistice* pentru limba româna izvoarele fundamentale de informatii cu privire la aceasta, stocate convenabil (chiar daca înca preponderent în maniera traditionala) si care, în calitate de componente ale *culturii* în sensul cel mai larg, sunt susceptibile de a fi valorificate pentru studierea limbii române, precum si în diverse scopuri conexe, inclusiv aplicative, în cadrul *societatii informatice* actuale.

Cât priveste Institutul de Lingvistica „Iorgu Iordan”[2], acesta nu mai exista formal ca atare, deoarece la începutul anului 2002, printr-o hotarâre de guvern adoptata la propunerea conducerii Academiei Române, s-a produs re-unirea sa si a Institutului de Fonetica si Dialectologie „Al. Rosetti”. (Spunem reunire întrucât cercetarile de fonetica si de dialectologie formasera initial obiectul unui sector, respectiv al unei sectii a Institutului de Lingvistica din Bucuresti al Academiei Române (înfiintat în 1949), devenita din 1961 centru si apoi institut independent.) Întrucât în 1998 fusese oficializata, tot prin hotarâre de guvern, propunerea celor doua institute, aprobata de Prezidiul Academiei, de a-si adauga fiecare în titulatura numele fostului sau director, institutul în cadrul caruia cele doua nuclee care au fuzionat acum îsi continua de fapt activitatea poarta numele dublu de Institutul de Lingvistica „Iorgu Iordan – Al. Rosetti”.

Fara îndoiala, cele mai numeroase si mai importante resurse lingvistice pentru limba româna s-au realizat la acum fostul Institut de Lingvistica „Iorgu Iordan”, înglobând, pâna în 1961 direct si apoi numai indirect, si contributia colegilor foneticieni si

[1] *[] *** (1975).* Dictionarul limbii române (DLR). Serie noua. Tomul IX, Litera R, *Bucuresti,* s.v.
[2] *[] Pentru o imagine de ansamblu asupra activitatii acestui institut si a istoriei sale v. Mioara Avram, Marius Sala, Ioana Vintila-Radulescu (coordonatori) (1999).* Institutul de Lingvistica „Iorgu Iordan“. 50 de ani de existenta (1949-1999), *Bucuresti.*

dialectologi[3], precum si, în unele cazuri, în colaborare cu alte institute de specialitate din tara ale Academiei – Institutul de Lingvistica si Istorie Literara „Sextil Puscariu" din Cluj si Institutul de Filologie Româna „Alexandru Philippide" din Iasi – si cu cadre didactice de la facultatile de profil mai ales ale Universitatii din Bucuresti. Aceasta activitate este continuata si în noul cadru organizatoric de sectoarele fostului institut, pe care în cele ce urmeaza îl vom numi, pe scurt, *Institutul*.

2. Resurse lexicografice

Dintre resursele lingvistice traditionale dezvoltate pâna în prezent de Institut, cele mai importante din punctul de vedere care intereseaza aici sunt cele **lexicografice – dictionarele** (mono- si bilingve) –, activitatea lexicografica din Institut, începuta înca de la înfiintarea sa, desfasurându-se din 1959 în cadrul unui sector specializat cu acest profil, condus pâna în 1985 de Mircea Seche, iar de atunci încoace de Ion Danaila[4].

2.1. Dictionare monolingve

2.1.1. Dintre dictionarele românesti monolingve se distinge, prin anumite trasaturi ale sale, dictionarul „explicativ general academic" intitulat pur si simplu ***Dictionarul limbii române*** – dar mai cunoscut ca „Dictionarul Academiei" –, a carui realizare se apropie de sfârsit si care va cuprinde o mare parte a „tezaurului" lexical al limbii române – fara a putea si nici a intentiona sa includa însa ansamblul cuvintelor românesti folosite în toate epocile, în toate regiunile si în toate domeniile[5]. În ciuda marilor sale calitati, care sunt bine cunoscute si asupra carora nu credem deci ca mai este nevoie sa insistam aici, acest dictionar prezinta un dezavantaj major din punctul de vedere al utilizarii sale ca resursa de baza (pe lânga faptul ca nu se prezinta si sub forma unei variante electronice, care nici nu putea fi imaginata pâna nu de mult) si anume caracterul sau fatalmente neunitar, datorat faptului ca a fost elaborat pe parcursul a aproape un secol[6], de unde marile deosebiri dintre cele doua parti ale sale: cea publicata între 1907 si 1949 sub conducerea marelui lingvist Sextil Puscariu si cea care a început sa apara din 1965 si a carei publicare se apropie, în fine, de sfârsit. „Seria veche" a dictionarului academic, desemnat de aceea prin sigla DA,

---

[3] *Acestia au produs mai ales „resurse" de un tip specializat, concretizate în principal în atlase lingvistice si în arhiva fonogramica a limbii române, de care nu ne vom ocupa în mod direct aici, dar care, ca si contributiile similare ale altor institute, au avut si un aport indirect la resursele fundamentale despre care vorbim, printre izvoarele carora s-au numarat.*

[4] *[] Pentru detalii cu privire la lucrarile acestuia v. Ion Danaila (1999).* Sectorul de lexicologie si lexicografie, *în Mioara Avram, Marius Sala, Ioana Vintila-Radulescu, op. cit., p. 98-113.*

[5] *[] Ideea, relativ utopica si controversata, a înregistrarii si chiar a descrierii semantice a întregului inventar lexical al limbii române (ILEX) din toate timpurile, incluzând atât numele comune, cât si cele proprii (v. Ion Danaila (1993).* Pentru un inventar general al limbii române*, în „Limba româna" XLII, nr. 2, p. 61-68), nici nu a început a fi pusa în practica.*

[6] *[] V., printre altele, Marius Sala (1999).* Institutul de Lingvistica „Iorgu Iordan" la 50 de ani, *în Mioara Avram, Marius Sala, Ioana Vintila-Radulescu, op. cit., p. 35-37.*

cuprinde literele *A-C* (inclusiv putinele neologisme scrise acum cu *k-,* iar în DA cu *ch-*) si *F–J* complet, iar literele *D* si *L* partial (pâna la cuvântul *de*, respectiv *lojnita*), totalizând 3.142 de pagini de tipar, format mare, dintre ele lipsind în întregime, dupa cum se observa, litera *E*. Aceasta prima jumatate a dictionarului se distinge prin lista de cuvinte, bogata mai ales sub aspectul fondului traditional, prin tratarea amanuntita a semantismului, bazata pe numeroase citate, prin dimensiunile si valoarea comentariului etimologice, precum si prin traducerea sensurilor în limba franceza[7]. Desigur, nu aveau cum figura în aceste prime volume numeroasele neologisme încetatenite în româneste dupa elaborarea ei, ilustrarea sensurilor prin utilizarea lor de catre autori mai noi si în general toate aspectele care sunt rodul evolutiei ulterioare a limbii române, al cercetarilor dialectale, etimologice, filologice etc. mai recente si al dezvoltarii lingvisticii si metodelor ei, în general. Din 1965 dictionarul si-a reînceput aparitia, în format asemanator, ca *Serie noua* (de data aceasta sub o sigla diferita, mentionata în titlu, DLR), cu litera *M,* sub conducerea, la început, a lui Iorgu Iordan, Alexandru Graur si Ion Coteanu, iar actualmente a lui Gh. Mihaila si Marius Sala. Noua serie pastreaza, în mare, principiile lui Sextil Puscariu, dar beneficiaza de toate avantajele elaborarii sale mai aproape de zilele noastre: ea include modificari si amplificari reflectând evolutia limbii române, a lexicografiei românesti si a studiului limbii române în ansamblu, precum si a lingvisticii în general, dar nu mai cuprinde, în schimb, traducerea sensurilor (în anii ’60 nefiind considerat oportun acest lucru, desi era util mai ales pentru cunoasterea limbii române de catre straini, fara a fi, este drept, uzual într-un dictionar monolingv explicativ), iar sectiunea etimologica a fost redusa, dictionarul pastrându-si însa caracterul istoric (sensurile sunt date în ordinea atestarii lor în texte si în alte surse)[8]. Institutul bucurestean a redactat literele *M, N, P, S* si Z[9] si este pe cale de a încheia reluarea si terminarea literei *D* absenta din prima parte (trei volume); numai primele patru litere elaborate la Bucuresti însumeaza 51.847 de cuvinte si variante, totalizând 5.839 p. Institutului din Cluj i-au revenit literele *O, R, T, T* (totalizând 2.044 de pagini de tipar), *U* (aflata sub tipar) si, din prima parte, reluarea si terminarea unei parti din litera *L*, iar celui din Iasi – literele *S, V* (*S* si prima parte din cele trei ale literei *V* – singura dintre acestea aparuta pâna acum –totalizând 599 de pagini de tipar), *W, X, Y,* precum si, din prima parte, elaborarea literei lipsa *E* si reluarea si terminarea unei parti din litera *L*; pentru etimologii au fost consultati specialisti din mai multe centre universitare. Majoritatea literelor au

[7] *[] Pentru o descriere amanuntita a DA v. Mircea Seche (1969).* Activitatea lexicografica a lui Sextil Puscariu, *în* Schita de istorie a lexicografiei române*, vol. II,* De la 1880 pâna astazi, *Bucuresti.*

[8] *[] V. si Mircea Seche (1969).* Seria noua a Dictionarului academic general*, în* Schita de istorie a lexicografiei române*, vol. II,* De la 1880 pâna astazi, *Bucuresti, p. 72-79.*

[9] [] Iorgu Iordan, Al. Graur, I. Coteanu (red. resp.) et al. (1965-2000). *Dictionarul limbii române (DLR*). *Serie noua,* Bucuresti: T. VI, *Litera M*, 1965–1968 (aparut initial în fascicule); VII, Partea 1, *Litera N,* 1971; Partea a 2-a, *Litera O, 1969*; VIII, *Litera P.* Partea 1, *P-PAZUI*, 1972; Partea a 2-a, PE–PÎNAR, 1974; Partea a 3-a, *PÎNA-POGRIBANIE*, 1977; Partea a 4-a, *POGRIJENIE–PRESIMTIRE*, 1980; Partea a 5-a, *PRESIN–PUZZOLANA*, 1984; IX, *Litera R*, 1975; X. *Litera S.* Partea 1, *S–SCLABUC*, 1986; Partea a 2-a, *SCLADA–SEMÎNTARIE*, 1987; Partea a 3-a, *SEMN–SÎVEICA*, 1990; Partea a 4-a, *SLAB–SPONGHIOS*, 1992; Partea a 5-a, *SPONGIAR–SWING*, 1994; XI Partea 1, *Litera S,* 1978; Partea a 2-a, *Litera T, T-TOCALITA,* 1982*; Partea a 3-a, TOCANA-TWIST,* 1983*;* XII, Partea 1, *Litera T,* 1994*; XIII, Partea 1, Litera V, V-VENI,* 1997; XIV, *Litera Z*, 2000.

aparut, unele pe sarite (*M* între 1965 si 1968, *N* în 1971, *O* în 1969, *P* între 1972 si 1984, *R* în 1975, *S* între 1986 si 1994, *S* în 1978, *T* în 1983, *T* în 1994, prima parte din *V* (pâna la *a veni*) în 1997 si *Z* în 2000) – în total 20 de volume –, cu exceptia literelor *D, E, K, L, U,* a putinelor cuvinte începând cu litera *Q* si a ultimelor parti ale literei *V* (începând cu *venin*), la care se adauga literele *W, X* si *Y.* Deosebirea cea mai importanta consta în tipurile de cuvinte reprezentate în cele doua serii: la majoritatea primelor litere ale alfabetului (cu exceptia celor care s-au redactat abia acum), neologismele sunt slab reprezentate, nu numai din cauza faptului ca foarte multe nici nu se încetatenisera înca în limba româna la vremea elaborarii volumelor respective, dar si din cauza reticentei lui Puscariu cu privire la acest sector al vocabularului; într-o situatie asemanatoare se afla termenii regionali, deoarece cercetarile dialectale se aflau în acea vreme abia la început. Prima parte prezinta în schimb avantajul de a putea servi ca baza pentru o prelucrare bilingva, întrucât includea si traducerea sensurilor în limba franceza, la care a trebuit sa se renunte în perioada comunista. Reluarea si completarea acestui dictionar, absolut necesara, nu ni se mai pare astazi recomandabil si nici posibil de realizat prin mijloace traditionale (fise etc.), ci exclusiv pe baze informatizate. Ea ar trebui sa valorifice, printre altele, si banca de texte si cea de inovatii a limbii române, despre care va fi vorba mai departe. Ar fi necesar ca partea publicata înainte de 1949 sa fie reluata si adusa la zi, cu atât mai mult cu cât putine persoane si chiar biblioteci poseda dictionarul în întregime (chiar în cazul seriei noi, tirajele diferitelor litere au fost diferite si în continua scadere), iar îmbatrânirea hârtiei în cazul seriei vechi o face fragila si greu de consultat. Având în vedere ca pentru noua serie a dictionarului s-au adunat, manual, peste sase milioane de fise cu extrase si atestari (dintre acestea, în DLR au fost incluse cca 3.200.000 de citate[10], reprezentând aproximativ 88% din totalul textului), este de sperat ca la reluarea, într-un viitor mai mult sau mai putin apropiat, se va putea uza de avantajele elaborarii computerizate, valorificându-se bancile de date în curs de elaborare în institut, despre care va fi vorba mai departe.

Având în vedere diferentele semnalate (dintre care unele se regasesc si între primele si ultimele litere din seria noua), este foarte binevenita ideea actualilor responsabili ai DLR de a se publica, pentru operativitate, un *Supliment* – care se poate realiza relativ mai lesne – „care sa înregistreze neologismele adoptate de limba literara de la începutul secolului" 20 „pâna în prezent, precum si o serie de cuvinte regionale incluse în atlasele lingvistice si în culegeri de pe teren sau termeni vechi extrasi din documente ale secolelor al XVI-lea – al XVIII-lea, editate în ultimele decenii"[11].

2.1.2. Din motivele expuse mai sus, la care se adauga si faptul ca DA/DLR este accesibil mai ales specialistilor si mai putin publicului larg, institutul bucurestean pregateste între timp, la sugestia conducerii Academiei Române, o sinteza a marelui dictionar academic, fara citate si izvoare si cu un sistem foarte economic de prezentare a

[10] *[] În legatura cu reflectarea noilor norme ortografice ale limbii române în volumele DLR elaborate dupa 1993, semnalam faptul ca forma* sânt*, reflectând un fonetism real, vechi si popular, este pastrata în citatele în care nu era folosit* sunt.

[11] *[] Marius Sala, G. Mihaila (2000).* Cuvânt înainte*, în* Dictionarul limbii române (DLR). Serie noua. Tomul XIV. Litera Z*, Bucuresti, p. VI.*

informatiilor lexicografice. Acest ***Mic dictionar academic*** (MDA)[12] (care va avea totusi patru volume), inclus, alaturi de DLR, printre lucrarile fundamentale ale Academiei Române, va avea cca 175 000 de intrari (cc125 000 de cuvinte si cca 50.000 de variante); primul volum (A-C) a fost publicat în anul 2001 de editura Univers Enciclopedic. Proiectul *Micului dictionar academic*, numit astfel în opozitie cu „marele" dictionar academic, si-a propus sa reduca decalajul dintre cele doua serii ale acestuia, îmbogatind primele litere pe baza unor surse lexicografice mai noi. La rândul sau, acest nou dictionar prezinta însa dezavantajul de a fi fost obligat, prin dimensiuni, sa renunte la citatele ilustrative, ceea ce limiteaza posibilitatea utilizarii lui ca sursa de informatii morfologice, gramaticale si stilistice; numarul neobisnuit de mare de abrevieri netransparente, utilizate din acelasi motiv de economie, constituie un argument suplimentar în favoarea realizarii unei versiuni electronice a MDA care sa permita regasirea automata a informatiilor.

2.1.3. Spre deosebire de DA/DLR, o reflectare în general unitara a vocabularului limbii române ofera ***Dictionarul explicativ al limbii române***[13], despre a carui sigla, DEX, se afirma, pe drept cuvânt, ca a devenit un apelativ; denumirea, care ar fi trebui protejata prin înregistrare, a fost preluata abuziv de *Noul dictionar explicativ al limbii române* publicat pe CD-Rom de firmele Litera în sigla NODEX, sugerând ca ar fi „un nou DEX". Prima editie, un volum de 1.049 de pagini, cuprinzând 56.569 de cuvinte si variante, a fost urmata de un ***Supliment la*** **Dictionarul explicativ al limbii române** *(DEX-S)*[14]. Editia a doua a DEX[15] totalizeaza 1.204 pagini; aceasta editie, care se publica în continuare în tiraje succesive, totalizase numai în primii patru ani de la aparitie 65.000 de exemplare vândute, dupa un calcul sumar rezultând ca la 42 de locuitori ai României revenea un *DEX*. Actualmente, se poate într-adevar afirma ca, prin *DEX*, *best-sellerul* lingvisticii românesti, Institutul a intrat în marea majoritate a caselor din România. Se preconizeaza ca DEX sa fie realizat, în fine, într-un viitor relativ apropiat, si în format electronic. El a fost deja supus, de catre Centrul de Cercetari Avansate în Învatarea Automata, Prelucrarea Limbajului Natural si Modelarea Conceptuala al Academiei Române, codificarii conform TEI[16]. Se estimeaza ca editia a III-a a DEX, conceputa sub conducerea lui Ion Danaila, va avea în plus fata de precedenta cca 30 000 de cuvinte. Sub conducerea lui Ion Coteanu si Ion Danaila, la sectorul de specialitate al Institutului a fost conceput si un ***Nou dictionar explicativ al limbii române*** (NEX), cu caracteristici diferite de cele ale *DEX*: inventar de

---

[12] *[] V. I. Danaila (1994).* De ce este nevoie de un MDA?*, în „Limba româna" XLIII, p. 397-406 si Marius Sala (2001).* Prefata*, în* Micul dictionar academic (MDA)*, vol. I,* A-C*, Bucuresti.*

[13] *[] I. Coteanu, Luiza Seche, M. Seche (conducatorii lucrarii) et al. (1975, 1996).* Dictionarul explicativ al limbii române *(DEX), Bucuresti.*

[14] *[] Ion Coteanu, Ion Danaila, Nicoleta Tiugan (conducatorii lucrarii) et al. (1988).* Supliment la Dictionarul explicativ al limbii române (DEX-S)*. Bucuresti.*

[15] *[] Ion Coteanu, Lucretia Mares (sub conducerea) et al. (1996),* Dictionarul explicativ al limbii române *(DEX), editia a II-a, Bucuresti.*

[16] *[] Dan Tufis (2000).* Cercetare si colaborare internationala în ingineria lingvistica la RACAI, *în „Terminologia în România si în Republica Moldova", Cluj-Napoca, p. 34-36 si* Recherche et collaboration internationale en industries de la langue r l'Académie Roumaine, *în „Terminometro Hors-série n^{o} 4. La terminologie en Roumanie et en République de Moldova", p. 38-40.*

cca 100.000 de cuvinte si variante (deci aproape de doua ori mai multe decât prima editie a DEX), definitii mai concise, prin eliminarea sinonimelor si – din pacate!–, neincluderea etimologiei cuvintelor; revizuit de cei doi responsabili, el asteapta introducerea în calculator, în vederea efectuarii corelatiilor semantice definitionale si sinonimice.

2.1.4. DEX a scos practic din circulatie dictionarele explicative mai vechi, limitate la limba româna literara, DLRLC si DM[17]. Prima sigla reprezinta ***Dictionarul limbii române literare contemporane***[18], elaborat de institutele din Bucuresti si Cluj pornind de la „baza manuscrisa" a DA si aparut între 1955 si 1957 în patru volume. El se mai foloseste si astazi – desi din el lipsesc cuvintele, sensurile si citatele neconforme cu ideologia vremii – pentru citatele cu care, spre deosebire de dictionarele de dimensiuni comparabile mai noi, sunt ilustrate sensurile cuvintelor (chiar daca, pentru unele neologisme, citatele provin, asa cum era obligatoriu în epoca, din traducerile „operelor clasicilor" marxism-leninismului!). Dintre acestea, primul mai merita însa atentie în virtutea faptului ca, spre deosebire de DEX si de MDA, include citate ilustrative, care din pacate au fost eliminate din dictionarele urmatoare.

2.1.5. O versiune prescurtata a acestui dictionar, cu un inventar putin marit si cu adaugarea etimologiei cuvintelor, dar cu eliminarea citatelor, a fost publicat de Institutul din Bucuresti în 1958 sub titlul ***Dictionarul limbii române moderne***[19] (abreviat DM).

2.1.6. Un dictionar de un tip special, cu o utilitate mult mai larga decât aceea care i se recunoaste de obicei, elaborat de data aceasta de colectivul de gramatica al Institutului (condus pâna de curând de Mioara Avram[20]), este ***Dictionarul ortografic, ortoepic si morfologic al limbii române*** (DOOM)[21]. Este singurul dictionar al limbii române (mai bogat decât DEX_1) care contine ample informatii cu privire la formele flexionare ale cuvintelor variabile incluse, putând servi astfel (chiar daca aceste informatii nu sunt exhaustive) ca sursa pentru studii si aplicatii de morfologie. Institutul are în prezent în lucru, sub conducerea subsemnatei, o a doua editie, partial revazuta si adaugita, a DOOM (care va cuprinde si cuvinte neînregistrate în nici un dictionar românesc pâna în prezent). Aceasta va aparea în anul 2003, inclusiv pe CD-Rom, si va trebui sa serveasca drept baza unui nou corector ortografic si morfologic, care sa tina seama de modificarea unor recomandari oficiale în raport cu cele înca în vigoare.

---

[17] *[] Pentru detalii cu privire la aceste doua dictionare v. Mircea Seche (1969).* Dictionarele explicative ale limbii române literare*, în* Schita de istorie a lexicografiei române*, vol. II,* De la 1880 pâna astazi*, Bucuresti, p. 135-147.*

[18] *[] D. Macrea, E. Petrovici (sub directia) et al. (1955-1957).* Dictionarul limbii române literare contemporane *(DLRLC), Editura Academiei, Bucuresti, vol. I,* A-C*; II,* D-L*, 1956; III,* M-R*, 1957; IV,* S-Z*, 1957.*

[19] *[] D. Macrea (sub directia) (1958).* Dictionarul limbii române moderne*, Bucuresti.*

[20] *[] Pentru activitatea acestuia v. Mioara Avram (1999).* Colectivul de gramatica*, în Mioara Avram, Marius Sala, Ioana Vintila-Radulescu, op. cit., p. 113-125.*

[21] *Mioara Avram (red. resp.) et al. (1982).* Dictionarul ortografic, ortoepic si morfologic al limbii române *(DOOM), Bucuresti, 1982.*

2.1.7. În fine, un dictionar mai putin obisnuit, ***Dictionarul invers***[22], în care cuvintele sunt ordonate alfabetic pornind dinspre sfârsitul lor, este deosebit de util specialistilor pentru studierea terminatiilor, a desinentelor si a sufixelor, dar si poetilor, fiind utilizabil si ca dictionar de rime. Aceasta lucrare – care, spune „legenda", a valorificat experienta din copilarie a uneia dintre autoare, care folosise în joaca o *pasareasca* de acest fel – ar merita si ea o noua elaborare, pe baza unui inventar mai bogat si actualizat de cuvinte si a unui program care sa permita „rasturnarea" lor automata.

2.1.8. Institutul a publicat, înca din 1968, un dictionar al lexicului unui autor, primul ales neputând fi altul decât Eminescu – *Dictionarul limbii poetice a lui Eminescu*[23], care însa, la acea vreme, nu se putea baza, evident, pe stabilirea concordantelor asa cum se realizeaza ea în zilele noastre.

2.1.9. Institutul a elaborat de asemenea o serie de dictionare ale limbii române pe epoci sau pe probleme, cum sunt *Dictionarul limbii române literare vechi*[24] si *Dictionarul împrumuturilor latino-romanice în limba româna veche*[25], publicate de sectorul de limba literara, filologie si poetica[26], condus de Ion Ghetie, iar în prezent de Alexandru Mares – si *Dictionarul elementelor românesti din documentele slavo-române*[27], elaborat la sectorul de slavistica[28] – dictionare destinate în primul rând specialistilor.

2.1. 10. Un cercetator din institut, Constant Maneca, a publicat, împreuna cu Florin Marcu, un extrem de util, cu toate criticile care i s-au adus, *Dictionar de neologisme*[29], reluat si dezvoltat, dupa moartea celui dintâi, de Florin Marcu, în numeroase variante, de diverse dimensiuni, la diferite edituri, inclusiv pe CD-Rom.

2.1.11. Se afla în lucru si ***Dictionarul etimologic al limbii române*** (DELR) – coordonator: Marius Sala –, alta lucrare fundamentala a Academiei Române, la care colaboreaza cercetatori din toate sectoarele Institutului, cercetatori din Cluj si Timisoara si cadre didactice de la universitatile din Bucuresti, Cluj si Timisoara.

---

[22] *[] *** (1957).* Dictionar invers, *Bucuresti. V. si Mircea Seche (1969).* Schita de istorie a lexicografiei române*, vol. II,* De la 1880 pâna astazi*, Bucuresti, p. 254-255.*

[23] *[] Tudor Vianu (sub redactia) et al. (1968).* Dictionarul limbii poetice a lui Eminescu, *Bucuresti.*

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

[25] *[] Gh. Chivu, Emanuela Buza, Alexandra Roman Moraru (1992).* Dictionarul împrumuturilor latino-romanice în limba româna veche (1421-1760)*, Bucuresti.*

[26] *[] V. Ion Ghetie (1999).* Colectivul de limba literara si filologie, *în Mioara Avram, Marius Sala, Ioana Vintila-Radulescu, op. cit., p. 132-143.*

[27] *[] Gheorghe Bolocan (redactor responsabil) et al. (1981).* Dictionarul elementelor românesti din documentele slavo-române. 1374-1600*, Bucuresti.*

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

[29] F. Marcu, C. Maneca (1961-1978). *Dictionar de neologisme*, Bucuresti, 1961; ed. II revazuta si adaugita, 1966; [3]1978. V. si Mircea Seche (1969). *Schita de istorie a lexicografiei române*, vol. II, *De la 1880 pâna astazi*, Bucuresti, p. 154-159.

2.1.12. Pe lânga resursele privitoare la numele comune, Institutul a elaborat si importante lucrari consacrate numelor proprii[30].

Astfel, în domeniul toponimiei, dupa clasica lucrare a lui Iorgu Iordan[31], s-a realizat în Institut ***Dictionarul 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 aparut în perioada 1993-2001 primele trei volume, precum si al doilea dictionar din serie, consacrat ***Munteniei*** si aflat în curs de definitivare.

În domeniul onomasticii, de asemenea urmând altei lucrari clasica a lui Iorgu Iordan[33], Institutul colaboreaza si la proiectul international PatRom, care realizeaza un dictionar istoric de antroponimie romanica, în care este reprezentata si limba româna, si din care pâna acum a fost publicat un prim volum de prezentare[34]

2.2. Dictionare bilingve si multilingve

2.2.1. Pe lânga dictionarele monolingve ale limbii române, Institutul a realizat si unele din cele mai importante dictionare bilingve[35] (englez-român[36], german-român[37], rus-român[38], ceh-român[39] si sârb-român[40] – perechea sa, dictionarul român-sârb, fiind în curs de redactare; un dictionar francez-român a ramas nepublicat) si frazeologice (spaniol-român, sub tipar, si român-spaniol, în curs de elaborare), carora li se adauga dictionare bilingve[41] – care au început a fi transpuse si pe CD-Rom – si dictionare frazeologice

---

[30] *[] Pentru activitatea în acest domeniu v. Gheorghe Bolocan, Ecaterina Mihaila (1999).* Colectivul de onomastica *si Domnita Tomescu (1999).* Grupul de lucru PatRom, *în Mioara Avram, Marius Sala, Ioana Vintila-Radulescu, op. cit., p. 125-132.*

[31] [] Iorgu Iordan (1952-1963). *Nume de locuri românesti în Republica Populara Româna* Bucuresti,1952; *Toponimia româneasca,* Bucuresti, 1963.

[32] [] Gh. Bolocan (sub redactia) et al. (1993-2001). *Dictionarul toponimic al României. Oltenia* (DTRO), vol. I-III, Craiova, Editura Universitaria.

[33] *[] Iorgu Iordan (1983).* Dictionar al numelor de familie românesti, *Bucuresti,* ***Editura Academiei. DE VERIF.***

[34] **** (1997).* Dictionnaire historique d'anthroponymie romane (PatRom). Présentation d'un projet, *Tübingen.*

[35] *[] V. si Ilinca Constantinescu. (1999).* Fostul sector de germanistica, *în Mioara Avram, Marius Sala, Ioana Vintila-Radulescu (coordonatori) (1999), op. cit., p. 174-179.*

[36] [] L. Levitchi (red. resp.) et al. (1974). *Dictionar englez–român*, Bucuresti. Suplimentul la acest dictionar, care nu a mai aparut, coordonat de Ilinca Constantinescu, va fi inclus într-o noua editie, mult marita, a dictionarului, aflata sub tipar si care va reprezenta cel mai bogat dictionar englez-român.

[37] *[] M. Isbasescu, Maria Iliescu (coord. si revizie) et al. (1966, 1988).* Dictionar german–român*, Bucuresti, 1966; editia a II-a revazuta si îmbogatita, Bucuresti, 1988.*

[38] *[] Gheorghe Bolocan (redactor responsabil) (1964).* Dictionar rus–român*, Bucuresti.*

[39] *[] S. Stati (red. resp.) et al. (1967).* Dictionar ceh–român*, Bucuresti.*

[40] *[] M. Tomici (1998-2000).* Dictionar sârb-român*, 3 vol., Timisoara.*

[41] [] Gh. Bolocan (1972). *Dictionar bulgar–român*, Bucuresti – Sofia; Gh. Bolocan et al. (1980). *Dictionar român–rus*, Bucuresti – Moscova; Al. Calciu, C. Duhaneanu, D. Munteanu (1979). *Dictionar român–spaniol*, Bucuresti; Ana Canarache (coord.) (1967, 1978). *Dictionar român-francez*, Bucuresti, ^{2}1978; M. Isbasescu (red. resp.) (1963), *Dictionar român-german*, Bucuresti; Valeria

românesti[42] si bilingve[43] elaborate de unii membri ai Institutului; D*ictionarul elen-român*, lucrare colectiva, se apropie si el de sfârsit.

2.2.2. Institutul a colaborat si la mai multe dictionare multilingve[44], dintre care ase distinge în mod deosebit un lexicon multilingv de un tip special – o adevarata premiera internationala – este ***Dictionarul elementelor latinesti savante din limbile romanice***, elaborat la sectorul de romanistica (condus initial de marele romanist Iorgu Iordan, apoi de Marius Sala si în prezent de subsemnata)[45], în colaborare cu cadre didactice de la Facultatea de Limbi si Literaturi Straine a Universitatii din Bucuresti si în coordonarea prof. dr. Sanda Reinheimer Rîpeanu, decanul Facultatii: Negasindu-si un editor „clasic" din cauza costurilor prea ridicate, acest dictionar va fi publicat direct pe Internet, sub auspiciile Universitatii din Bucuresti.

3. Banci de date

3.1. Institutul a avut în proiect înca din anii 1978-80 realizarea primei banci computerizate de date lingvistice din România (***Banca de date fono-morfo-semantice a limbii române – BANDASEM***)[46], cel dintâi modul fiind cel de semantica, proiectat pentru un *Dictionar confruntativ de sinonime, de analogii si de asociatii al limbii române (DCSAAs)*. Redactarea acestuia, care a ajuns la litera *S,* s-a facut însa cu mijloace traditionale, desi prin colaborarea cu Centrul de Calcul al Universitatii din Bucuresti se elaborase un modul de program în sistemul Socrate pentru recunoasterea si selectarea, ca proba, a analogiilor si a asociatiilor cuvântului *blitz*. Elaborarea DCSAAs a fost întrerupta pentru un timp în favoarea lucrarilor prioritare al Academiei, iar reluarea lui se va putea face, speram, cu mijloacele informatice disponibile actualmente[47].

3.2. O mininbanca initiata în cadrul sectorului de gramatica al Institutului, a carei alimentare a fost din pacate întrerupta în favoarea concentrarii fortelor pentru realizarea

---

Neagu (2001). *Dictionar român-spaniol* (cu transpunere pe CD-Rom), Bucuresti.

[42] *[] V. Breban et al. (1969).* Dictionar de expresii si locutiuni românesti*, Bucuresti.*

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

[44] **** (1981).* Dictionnaire de la presse écrite et audiovisuelle. Espagnol–français–italien–portugais–roumain*, Paris; *** (2001).* Usage Dictionary of Anglicisms in Selected European Languages (UDASEL) *Oxford s.a.*

[45] *[] Cu privire la activitatea acestuia v. Marius Sala (1999).* Sectorul de limbi romanice si clasice, *în Mioara Avram, Marius Sala, Ioana Vintila-Radulescu (coordonatori) (1999), op. cit., p. 147-164.*

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

[47] *[] Partea de fonetica/grafematica si de morfologie a* BANDASEM *a fost cedata institutului omolog din Cluj, pentru care v. Felicia Serban et al. (2000).* Baza de date a limbii române, *în „Terminologia în România si în Republica Moldova", Cluj-Napoca, p. 37-38 si* La base de données de la langue roumaine, *în „Terminometro Hors-série n^{o} 4. La terminologie en Roumanie et en République de Moldova", p. 40 -42.* ***VERSIUNEA INTEGRALA ÎN TUFIS ...***

editiei a doua a „Gramaticii Academiei", este ***Banca de inovatii a limbii române***, bazata pe monitorizarea presei scrise si audiovizuale actuale.

3.3. Având în vedere ca în DOOM informatia este atomizata, în folosul cititorului neprofesionist, în cadrul fiecarui cuvânt-titlu în parte, dar este greu de sistematizat de catre specialist, Institutul are în proiect, începând din 2003, realizarea unui baze de date care sa permita nu numai elaborarea unui ***Nou dictionar ortografic, ortoepic si morfologic al limbii române*** si a unor dictionare specializate de un tip asemanator, precum si aducerea lor permanenta la zi, ci si gruparea cuvintelor în clase în functie de caracteristicile lor fonetice, grafice si morfologice[48].

3.4. Institutul are în proiect si elaborarea sau definitivarea unor resurse terminologice[49] (dictionare terminologice bi- si multilingve, valorificând cele elaborate în cadrul proiectului PRACTEAST din cadru programului COPERNICUS al Comisiei Europene[50] si un dictionar al termenilor oficiali); de altfel, mai multi membri ai Institutului au colaborat la realizarea ***Bancii de date terminologice*** (BDT) multilingve a Asociatiei Române TermRom[51], care, cu sprijinul Directiei de terminologie si inginerie lingvistica a Uniunii Latine, este accesibila pe site-ul TermRom gazduit de CIMEC (http://www.cimec.ro/tr) si, de curând, si pe CD-Rom. Reprezentarea României (prin subsemnata) în Reteaua Panlatina de terminologie (Realiter)[52] si în Reteaua Francofona de

---

[48] *[] Clasificarea cuvintelor românesti conform modului lor de flexiune, realizata de Alf Lombard, Constantin Gâdei (1981).* Dictionnaire morphologique de la langue roumaine, *Lund – Bucuresti, bazata pe inventarul DEX$_I$, prezinta unele inexactitati din cauza insuficientei cunoasteri de catre autori a limbii române actuale; ea constituie una din bazele realizarii, în Republica Moldova, a unui pachet de programe destinat elaborarilor de nivel morfologie, pentru care v. Elena Boian et. al. (2000).* Instrumentar pentru aplicatii lingvistice, *în „Terminologia în România si în Republica Moldova", Cluj-Napoca, p. 38-40 si* Instruments pour applications lingustiques, *în „Terminometro Hors-série n^o 4. La terminologie en Roumanie et en République de Moldova", p. 42-44; o grupare pe tipuri a unui numar limitat de cuvinte ale limbii române a fost realizata de Flora Suteu, Elisabeta Sosa (1999) în* Îndreptar ortografic si morfologic*, Bucuresti.*

[49] *[] V. Ioana Vintila-Radulescu (1999).* Institutul de Lingvistica „Iorgu Iordan" din Bucuresti*, în „Terminologia în România si în Republica Moldova", Cluj-Napoca, p. 13-15, si* L'Institut de Linguistique Iorgu Iordan de Bucarest*, în „Terminometro Hors-série n^o 4. La terminologie en Roumanie et en République de Moldova", p. 22-13.*

[50] *[] Nicoleta Petuhov. (2000).* Colaborarea românească la proiectul PRACTEST, *în „Terminologia în România si în Republica Moldova", Cluj-Napoca, p. 58-59 si* La collaboration roumaine au projet Practeast, *în „Terminometro Hors-série n^o 4, La terminologie en Roumanie et en République de Moldova", p. 64-66.*

[51] *[] Dan Matei (2000).* Banca de date terminologice a TermRom, *în „Terminologia în România si în Republica Moldova", Cluj-Napoca, p. 29-30 si* La banque de données terminologiques de TermRom*, în „Terminometro Hors-série n^o 4, La terminologie en Roumanie et en République de Moldova", p. 32-33.*

[52] *[] Dan Matei (2000).* Prezenta românească în reteaua panlatina de terminologie (Realiter), *în „Terminologia în România si în Republica Moldova", Cluj-Napoca, p. 56-58 si* La présence roumaine dans le Réseau panlatin de terminologie Realiter*, în „Terminometro Hors-série n^o 4. La terminologie en Roumanie et en République de Moldova", p. 63-64.*

Amenajare Lingvistica (Rifal)[53] vor constitui desigur un sprijin în dezvoltarea resurselor terminologice pentru limba româna în conformitate cu normele si recomandarile internationale.

În afara numelor comune, si numele proprii au constituit obiectul preocuparilor institutului si ale unor membri ai sai.

4. Corpusuri

O alta categorie importanta de resurse lingvistice o constituie corpusurile, la Institut fiind în curs de realizare o ***Banca de texte românesti,*** care cuprinde texte din secolele al XVI-lea – al XVIII-lea, introduse integral în calculator, si în care se prevede introducerea câtorva sute de texte din toate epocile. Initiata de directorul institutului, acad. Marius Sala. *Banca,* a fost deja valorificata în elaborarea unor teze de doctorat, printre altele la aceea a Janei Balacciu-Matei. Pentru exploatarea ei deplina în vederea identificarii primelor atestari ale cuvintelor limbii române din fondul vechi, necesare MDA si *Dictionarului etimologic al limbii române* (DER) (în curs de elaborare sub conducerea acad. Marius Sala), a îmbogatirii dictionarelor limbii române în general si a dezvoltarii studiilor privind istoria limbii române literare si a limbii noastre în ansamblu este necesara achizitionarea unor programe de ultima ora, precum si specializarea unor persoane pentru utilizarea lor eficienta. Speram de asemenea ca într-un viitor nu prea îndepartat se va realiza si dorita jonctiune cu Banca de texte din faza moderna si contemporana a limbii române, proiectata a se realiza la Centrul de Studii Românesti de pe lânga Universitatea din Anvers, inaugurat în primavara anului 2000 sub conducerea cunoscutei romaniste si românisteLiliane Tasmowski.

5. Resurse bibliografice

Amintim pe scurt si principalele resurse bibliografice privitoare la limba româna elaborate de Institut sau de membri ai acestuia. *Bibliografia limbii române*, initiata de Al. Rosetti si definitivata de Aurel Nicolescu, a ramas nepublicata. [54]*Bibliografia româneasca de lingvistica (BRL)* referitoare la lucrarile de lingvistica aparute în tara începând din 1944 apare anual în revista „Limba româna“; în 1999, ea totalizase deja 64.340 de titluri, în peste 3.300 de pagini de tipar; se preconizeaza introducerea în calculator a tuturor numerelor din BRL în vederea publicarii unui volum cu itemurile ordonate pe autori si pe domenii

[53] *[] Ioana Vintila-Radulescu (2000).* Colaborarea în cadrul ACCT/Agentiei Interguvernamentale a Francofoniei si al Rifal, *în „Terminologia în România si în Republica Moldova”, Cluj-Napoca, p. 51-52 si* La coopération dans le cadre de l'ACCT (Agence itergouvernementale de la Francophonie)*, în „Terminometro Hors-série n° 4. La terminologie en Roumanie et en République de Moldova”, p. 57-58.*

[54] *[] I. Coteanu, I. Danaila (1970).* Introducere în lingvistica si filologia româneasca. Probleme. Bibliografie, *Bucuresti; T. Vianu (red. resp.) et al. (1972).* Bibliografia analitica a limbii române literare. 1780–1866, *Bucuresti; Gh. Chivu, Mariana Costinescu (1974).* Bibliografia filologica româneasca. Secolul al XVI-lea, *Bucuresti.*

(descrise si separate mai amanuntit decât în forma aparuta, cronologic, cu indice de domenii, materii, cuvinte, autori etc.).

Pentru domeniul terminologiei s-au realizat bibliografii ale dictionarelor terminologice, respectiv ale studiilor de terminologie[55] si ale standardelor românesti de/cu terminologie[56], precum si un repertoriu bio-bibliografic al terminologilor din România[57], care va fi inclus în repertoriul international al terminologilor din domeniul neolatin pregatit de Uniunea Latina, fiind în curând accesibil pe Internet.

6. Concluzii

Nu ne vom referi aici la alte tipuri de lucrari (gramatici[58], tratate[59], enciclopedii[60] etc.) elaborate de Institut sau de cercetatori ai acestuia ori la alte tipuri de resurse care ar merita sa fie elaborate de noul institut, pentru a înlocui lucrari mai vechi si a valorifica posibilitatile oferite culturii de societatea informationala, de exemplu un nou dictionar de frecventa al limbii române s.a.

Desi dictionarele pe CD-Rom si cele pe Internet sunt solicitatete de tot mai multi utilizatori din tara si din strainatate, care cer tot mai des informatii cu privire la eventuale dictionare românesti on-line, pâna în prezent a existat la noi o anumita reticenta a editurilor proprietare ale drepturilor asupra editiilor pe suportul traditional de hârtie fata de acest nou mod de difuzare. Nu trebuie însa sa existe temerea ca folosirea si a noilor suporturi ar diminua vânzarea cartilor, în conditiile în care, în ciuda tuturor eforturilor, un procent înca infim din populatia Românei are acces la PC-uri. De altfel, practica altor tari a aratat ca, în mod neasteptat, difuzarea si în format electronic chiar a sporit desfacerea cartilor, carora le-

---

[55] *[] Anca Fezi et al. (2000).* Bibliografia lucrarilor de terminologie (1990-1999). România, *în „Terminologia în România si în Republica Moldova", Cluj-Napoca, p. 103-113 si pe discheta anexata revistei „Terminometro Hors-série n° 4. La terminologie en Roumanie et en République de Moldova", 2000.*

[56] *[] Aurora Petan, Edy Savescu (2000).* Standarde românesti de/cu terminologie (1990-1999). România, *în „Terminologia în România si în Republica Moldova", Cluj-Napoca, 2000, p. 117-126 si pe discheta anexata revistei „Terminometro Hors-série n° 4. La terminologie en Roumanie et en République de Moldova", 2000.*

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

[58] *[] *** (1954, 1963).* Gramatica limbii române, *Bucuresti, ed. I, 1954; ed. a II-a, revazuta si adaugita, 1963; Mioara Avram (1986, 1997, 2001).* Gramatica pentru toti, *Bucuresti, 1986; [2]1997; [3]2001.*

[59] *[]Al. Rosetti (redactor responsabil) et al. (1965, 1969).* Istoria limbii române. *Bucuresti, vol. I.* Limba latina, *vol. al II-lea; Al. Graur, Mioara Avram (1970-1989).* Formarea cuvintelor în limba româna, *Bucuresti: I. Fulvia Ciobanu, Finuta Hasan (1970).* Compunerea*; II. Mioara Avram et al. (1978).* Prefixele*, 1978; III. Laura Vasiliu (1989).* Sufixele*, 1.* Derivarea verbala *etc.*

[60] *[] Marius Sala, Ioana Vintila-Radulescu (1981).* Limbile lumii. Mica enciclopedie, *Bucuresti; (1984).* Les langues du monde. Petite encyclopédie, *Bucuresti – Paris; Marius Sala (coord.) et al. (1989).* Enciclopedia limbilor romanice, *Bucuresti; (2001),* Enciclopedia limbii române*, Bucuresti.*

a facut în felul acesta reclama si care prezinta, la rândul lor, alte avantaje în utilizare în raport cu CD-Romurile, cele doua tipuri specializându-se si în functie de necesitati. Astfel, având în vedere culegerea lor computerizata, atât DEX, cât si MDA si $DOOM_2$ ar putea fi primele dictionare ale Institutului difuzate în viitor si pe CD-Rom.

Credem ca si diverse lucrari valoroase ale Institutului, care, exclusiv din motive financiare, nu-si gasesc editori de ani de zile, nici în tara, nici în strainatate (ca *Bibliografia limbii române, Dictionarul spaniolei americane* s.a.), ar putea fi valorificare prin aducerea lor la cunostinta celor interesati pe aceasta cale, tot mai utilizata în societatea informationala actuala. O conditie pentru viitor este realizarea din capul locului a lucrarilor institutului pe calculator, care a devenit posibila prin tot mai buna dotare tehnica a Institutului, realizata prin eforturile directorului sau, precum si prin însusirea, de catre un numar tot mai mare de cercetatori din Institut, în special din generatiile tânara si mijlocie, a cunostintelor de operare pe calculator, inclusiv, în unele cazuri, a lucrului cu baze de date.

Prin realizarea proiectelor de editare pe CD-Rom si pe Internet vom recupera relativa întârziere în acest domeniu fata de difuzarea în România, de catre Grupului Editorial Litera din Republica Moldova si firma Litera International, cu sediul în Bucuresti, a unor CR-Romurile cuprinzând, în diverse combinatii, mai multe titluri[61]. Speram ca CD-Romurile consacrate unor dictionare ale Institutului vor fi, desi tot protejate, mai usor de instalat decât cele de la Litera si ca vor oferi mai multe facilitati în utilizare decât acestea, care nu sunt foarte practice, mai ales pentru cercetatori, în ciuda structurii lor modulare si a interfetei lor comune, despre care în reclama se spune ca permit activarea simultana a tuturor dictionarelor.

Pentru progresul cercetarilor si dezvoltarea si prelucrarea resurselor la nivelul exigentelor pe plan mondial, credem ca în viitor se impune o mai buna colaborare, în interes reciproc, între lingvisti si informaticienii preocupati de probleme asemanatoare.

[61] *[]* Corectorul electronic ORTO 2001 ROM SP, Dictionarul ortografic al limbii române, Gramatica uzuala a limbii române, Noul dictionar explicativ al limbii române, Marele dictionar de neologisme *de Florin Marcu*, Dictionarul de dublete etimologice ale limbii române *de Marcu Gabinschi si un* Dictionar de termeni de afaceri englez-român.

Contributia lingvisticii la studiul terminologiilor ştiinţifice

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

1. Se admite „laicizarea" stiintelor [1] sau importanta lor socio-culturala, economica si pedagogica tot mai mare în societatile moderne. Aceasta înseamna ca *limbajele specializate* si *terminologiile* lor nu mai reprezinta coduri total inaccesibile vorbitorilor obisnuiti, nespecializati sau de alta specialitate. În directia deschiderii, chiar si partiale a codurilor stiintifice, *dictionarele generale* [2], care includ un numar destul de mare de termeni stiintifici joaca un rol deosebit pentru a asigura accesul la sensul specializat oricarui vorbitor insuficient informat, pentru a-l ajuta sa rezolve ambiguitatile de diferite tipuri si chiar sa utilizeze adecvat o terminologie. Permanenta raportare la dictionarele generale ca forme institutionalizate de reglare a uzului nu numai al cuvintelor din limba comuna, ci si a termenilor specializati constituie premisa de la care pornim pentru a sustine importanta lingvisticii în descrierea terminologiilor stiintifice, în receptarea si utilizarea lor adecvata chiar si de catre nespecialisti.

Pe aceste pozitii s-a situat activitatea în cadrul a trei contracte de cercetare stiintifica pe anii 1997, 1999 si 2000, finantate de CNCSIS (Consiliul National de Cercetare Stiintifica). Au fost studiate limbajul **filozofic**, terminologiile **matematica, mineralogica** si din **artele plastice** si, dintr-o perspectiva mai limitata **medicina, lingvistica** si **stiintele politice**. Rezultatele cercetarilor au fost publicate în doua volume: *Lexic comun, lexic specializat* [3], care contine studii cu caracter monografic si *Lexic stiintific interdisciplinar* [4], reprezentând o sinteza a lexicografiei generale si specializate pentru termenii din fiecare dintre domeniile studiate care apar mai mult decât într-o terminologie stiintifica.

În toate cercetarile întreprinse s-a urmarit adoptarea *unei grile metodologice comune* atât pentru clase de cuvinte din limba comuna (*abstractele*), cât si pentru termenii specializati din orice domeniu. S-a obtinut atât caracterizarea fiecarei terminologii studiate în parte, cât si desprinderea unor trasaturi generale ale terminologiilor stiintifice, relevante din punct de vedere lingvistic. S-au avut în vedere aspecte *paradigmatice* privind diferitele modalitati de definire a sensului, relatiile semantice (*monosemie/polisemie, hiponimie, sinonimie*) din perspectiva necesitatii ca termenii stiintifici sa fie monoreferentiali, univoci din punct de vedere semantic si sa nu aiba sinonime. Analiza **sintagmatica** a gradului de non-determinare contextuala ca o conditie de exprimare a sensului specializat a indivi-

dualizat terminologiile stiintifice studiate, de la o *libertate contextuala* mai mare (terminologia **matematica, mineralogica**) sau relativa (terminologia **filozofica**) pâna la o *stricta determinare contextuala* (terminologia **politica** si din **artele plastice**). Acolo unde independenta contextuala e mai mare, determinarile contextuale exprima în mod similar în diferite terminologii (**matematica**, **filozofica**, **lingvistica**) subcategorii stiintifice care dezambiguizeaza lexicul stiintific intedisciplinar. Caracterizarea termenilor stiintifici prin marci diastratice în dictionarele generale si enciclopedice ca tipuri de informatii sintagmatice reprezinta un aspect foarte important pentru uzajul adecvat de catre specialisti, aspect deficitar, inegal rezolvat.

De pe pozitia receptorului nespecializat care decodeaza sensul total sau partial, un rol important îl are *definitia lexicografica* care, spre deosebire de cea *terminologica* trebuie sa fie mai mult sau mai putin *naturala* si prin aceasta accesibila. Existenta celor doua tipuri de definitii ale termenilor specializati este în general admisa si compararea lor este favorizata de prezentarea sintetica, sinoptica propusa de noi [4]. Chiar si în cazul definitiilor strict terminologice, Em. Vasiliu [5] a sustinut si demonstrat prin diferite exemple relevanta diferita a unor componente de sens pentru vorbitorul specialist sau non-specialist. Pornind de la aceste constatari de principiu, ar fi justificat ca termenii stiintifici sa aiba *definitii alternative*, stiintifice si pre-stiintifice [6], conditionate atât de o interpretare semantica, cât si de una pragmatica. Din aceasta perspectiva, definitiile termenilor stiintifici în dictionarele generale ar trebui sa difere de cele din dictionarele specializate pentru a facilita deschiderea codurilor stiintifice si pentru a dezambiguiza lexicul stiintific interdisciplinar (din principiu, de interes mai larg) sau tangentele cu limba comuna. Din pacate, cu mici exceptii (**matematica**) selectia termenilor stiintifici si definirea lor nu difera aproape deloc în dictionarele generale si în cele specializate.

2. Din perspectiva lingvistica, terminologiile investigate prezinta o serie de particularitati:

Matematica se caracterizeaza prin cel mai mare grad de abstractizare si de ermetism la nivelul sensurilor si definitiilor lor. Compararea definitiilor specializate cu cele din dictionarele generale arata ca acestea din urma definesc diferit si mai accesibil termenii, fara a afecta precizia lor semantica. Sensurile univoce, fara sinonime nu sunt conditionate contextual; sintagmele mai mult sau mai putin fixe diferentiaza subcategorii conceptuale (de ex. *sistem de ecuatii, ~ de curbe, ~ de numeratie, ~ de referinta*) si nu afecteaza independenta semantica a acestora. Aceasta terminologie dispune de cea mai buna marcare diastratica în DEX, chiar daca exista numeroase situatii în care apartenenta la matematica rezulta numai din definitie (maniera de caracterizare practicata sistematic si nu întotdeauna convenabil de DEX în cazul altor terminologii). **Matematica** are cel mai bogat lexic stiintific interdisciplinar, cei mai numerosi termeni comuni fiind cu **fizica, filozofia, logica**, dar si cu **lingvistica, biologia, arhitectura** s.a.; termenii interdisciplinari îsi pastreaza aproape neschimbat sensul, indiferent de domeniul în care se utilizeaza. Daca în unele cazuri (relatia cu **fizica, logica, filozofia**) punctul de plecare pentru lexicul interdisciplinar nu se poate stabili cu certitudine, în destule alte situatii, **matematica** este sursa „împrumutului" facut de alte stiinte (**arte plastice, arhitectura, lingvistica** s.a.)

Mineralogia reprezinta si ea un grad mare de ermetism sau închidere a codului, majoritatea termenilor fiind univoci semantic, monoreferentiali si implicit, independenti contextual. Determinarile contextuale reprezinta subtipuri, ca si în alte terminologii (**matematica, filozofie** de ex.*: acvamarin brazilian, ~ sintetic, ~ siamez,* etc.) Are un numar mai limitat de termeni comuni cu alte stiinte (**chimia, artele plastice, simbolistica**) si, cel putin pentru ultimele doua, **mineralogia** este punctul de origine al termenilor interdisciplinari. În ciuda caracterului strict specializat al acestei terminologii, marcarea diastratica din dictionarele generale este deficitara.

Terminologia **filozofica** se caracterizeaza printr-un grad oarecare de ambiguitate, determinat de variatii de interpretare în functie de curente si tipuri de texte, dar si de contactele cu alte stiinte sau cu limba comuna. De aceea definitiile termenilor **filozofici** nu se pot limita la dictionare, fiind necesara analiza strategiilor argumentative si a figurilor textuale. Invers proportional cu aceasta necesitate de dezambiguizare, DEX-ul prezinta o marcare diastratica deficitara atât pentru termenii filozofici, cât si pentru celelalte terminologii cu care se stabilesc interdisciplinaritati, cum ar fi **matematica, lingvistica** si alte domenii **umaniste**. O buna parte a lexicului stiintific interdisciplinar are ca punct de plecare **filozofia**, al carei sens se pastreaza ca o medie semantica în majoritatea disciplinelor. Ca si în alte stiinte, determinarea contextuala exprima în general subtipuri (de ex. *sistem al stiintelor, ~ axiomatic, ~ filozofic*).

Terminologia **artelor plastice** prezinta aspecte paradoxale. Maniera de înregistrare si de definire echivoca, imprecisa a acestor termeni în dictionarele generale da impresia unui nespecialist de falsa accesibilitate, interpretare contrazisa categoric de definitiile precise, riguroase din dictionarele si textele specializate. Dependenta contextuala stricta a numerosi termeni din **artele plastice**, al caror sens specializat e conditionat de sintagmele fixe în care apare (de ex. *acord cromatic, compozitie de gen, semn plastic*) reprezinta o alta caracteristica a acestei terminologii. **Artele plastice** au un lexic stiintific interdisciplinar bogat, în care se remarca faptul ca sunt preluati cu unele modificari semantice (privind interesul pentru acest domeniu) termeni din alte stiinte, cum ar fi **chimia, mineralogia, matematica, fizica**. DEX-ul nu utilizeaza decât marcile diastratice (pictura), (sculptura) dispuse nesistematic si rar, ceea ce contribuie la o tratare deficitara a acestei terminologii.

Lexicul **stiintelor politice** prezinta, din prespectiva analizei întreprinse de noi, o serie de particularitati (unele asemanatoare cu **artele plastice**). Se remarca dependenta contextuala stricta a acestei terminologii, nici unul dintre termeni nefiind total liber contextual. Sensul specializat în **stiintele politice** se exprima, deci, aproape exclusiv pe cale sintagmatica, în contexte mai mult (*celula de criza, agregare de interese*, de ex.) sau mai putin fixe (diverse combinatii cu adjectivul **politic** în sintagme nominale: *capital politic, cartel ~, algoritm ~, contract ~, dialog ~, alternanta politica*). Preia (fara sa fie niciodata punct de plecare termeni din numeroase si variate stiinte: **economia, filozofia, dreptul,** dar si **lingvistica, biologia, medicina, geografia, fizica, psihologia, sportul**. În majoritatea acestor cazuri nu exista o motivare de continut stricta (dincolo de întrebuintarea metaforica), ceea ce determina, în mare parte, mai curând un lexic stiintific interferent

decât unul interdisciplinar. Poate si din cauza modificarilor continue si rapide din domeniul politicii, DEX-ul înregistreaza în mica masura termeni si sensuri din acest domeniu diastratic, ceea ce constitutie un dezavantaj în impunerea acestei terminologii.

3. Analiza lingvistica a limbajelor stiintifice (care ar putea fi extinsa) permite caracterizarea unor terminologii ca „*puternice*" (**matematica, mineralogia** de ex.), iar a altora mai „*slabe*" în diferite forme si grade (de ex. **stiintele politice, artele plastice**), cu dificultati mai mari de deschidere a codurilor în cazul primei categorii.

Delimitarea componentelor de sens relevante diferit în functie de vorbitori specializati si nespecializati ar putea constitui o baza obiectiva pentru rezolvarea mai eficienta a definitiilor alternative în dictionarele generale, foarte importante în „laicizarea" stiintelor necesara în grade diferite în epoca actuala. Exprimarea sensului specializat conditionat de dependentele contextuale mai mici (pentru terminologiile „puternice") sau mai mari (pentru terminologiile „slabe") constituie o caracterizare lingvistica relevanta. În schimb, în unele cazuri (ca pentru terminologia **politica**), determinarile contextuale sunt mai favorabile, „transparentei" semantice sau deschiderii codurilor specializate.

Analiza lexicului stiintific interdisciplinar (LSI) poate contribui si ea la determinarea specificului unor terminologii. Stiintele care constituie sursa, punctul de plecare pentru o parte a LSI îsi sustin, si pe aceasta cale, statutul de terminologie „puternica" (de ex. **matematica, fizica** si, din acest punct de vedere **filozofia**). Dimpotriva, atunci când punctul de plecare nu se poate stabili aproape niciodata la nivelul unor terminologii (**stiintele politice, artele plastice**), aceasta constituie o modalitate de determinare specifica. Diferentierea interdisciplinaritatilor (cu o motivare de continut determinata de considerarea referentului din diferite punte de vedere sau de un transfer conceptual) de simplele interferente (mai putin sau deloc motivate, cu modificari de sens ale termenilor, multe metaforice) se bazeaza pe aprecierea distantei semantice, verificata obiectiv.

Dat fiind rolul dictionarelor generale în impunerea si extinderea terminologiilor stiintifice, de interes pentru diferite categorii de vorbitori, carentele constatate în tratarea sensului si în marcarea lor diastratica riguroasa conduc la concluzia necesitatii unei reconsiderarii si remedieri a manierei de tratare din perspectiva „laicizarii" stiintelor.

Referinte bibliografice

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

[2] DEX - Dictionar explicativ al limbii române, (1996) ed.a 2-a sub coord. acad- I. Coteanu si Dr. Lucretia Mares, Ed. Univers Encilopedic, Bucuresti 1996

[3] A. Bidu-Vranceanu – coordonator (2000). *Lexic comun, lexic specializat,* Editura Universitatii din Bucuresti, 2000, cu colaboratorii: Alice Toma (**matematica**), Silvia Savulescu (**mineralogie**), Claudia Ene (**filozofie**), Alexandra Vrânceanu (**arte plastice**)

[4] A. Bidu-Vranceanu – coordonator (2001). *Lexic stiintific interdisciplinar*, Editura Universitatii din Bucuresti, 2001, cu colaboratorii: Silvia Savulescu (**stiinte politice si mineralogie**), Alice Toma (**matematica**),Claudia Ene (**filozofie**), Alexandra Vrânceanu (**arte plastice**)

[5] Em. Vasiliu (1980). Sens si definitie lexicografica „Studii si cercetari lingvistice", an XXXI, 465, 1980

6] Em. Vasiliu (1982/1983). Adevar analitic si definitie lexicografica „Analele stiintifice ale Universitatii „Al. I Cuza" din Iasi", sectiunea III, tom XXVIII/XXIX, 1982/1983

Gramaticile generative nontransformationale

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

Acest articol este o prezentare generala a gramaticilor generative nontransformationale (GNT) si a prezentei lor în cercetarea lingvistica din România. În prima sectiune a articolului este descrisa geneza acestor gramatici. În sectiunea a doua, sunt prezentate pe scurt caracteristicile lor, în timp ce în partea treia si a patra se mentioneaza principalele realizari stiintifice si formele de existenta institutionala ale curentului. Partea a cincea este consacrata initiativelor si pasilor care au dus la patrunderea acestor gramatici în mediile stiintifice de la noi. Concluziile articolul se vor a fi o pledoarie în sprijinul eforturilor de dezvoltare a acestei directii în cultura stiintifica românească.

1. Gramaticile generative nontransformationale: aparitia lor

Gramaticile generative nontransformationale reprezinta, în interiorul lingvisticii formale contemporane, o directie extrem de influenta si de un remarcabil dinamism. Istoria acestei directii este, desigur, mai recenta decât istoria generativismului din care face parte. Este însa o istorie deja bogata si diversa. Printre altele, diversitatea se exprima si prin faptul ca suntem obligati sa vorbim despre *gramatici* si nu despre o gramatica nontransformationala, pur si simplu.

Putem plasa începuturile acestei istorii la cumpana dintre anii ’70 si ’80. Sunt anii când programul gramaticii universale al lui Noam Chomsky este pe punctul sa depaseasca starea de impas atinsa prin faza denumita de istoricii miscarii “teoria standard”. Privita din perspectiva prezentului, lucrarea din 1981 a lui Chomsky (“Lectures on Government and Binding”) tocmai acest lucru îl subliniaza: depasirea crizei prin propunerea unui model nou de gramatica universala.

Punctele în care gramatica universala este reformulata în cadrul modelului “Government and Binding” (GB) nu sunt putine si nici neînsemnate. Dar cea mai importanta modificare a fost operata într-una din componentele care nascuse initial cele mai mari sperante: componenta transformarilor. Formulata succint, regândirea conceptului de transformare în cadrul modelului GB înseamna doua lucruri: simplificare si îngradire.

Simplificare, deoarece marea varietate de transformari se reduce acum la o singura operatie: deplasarea unui constituent oarecare α. Si îngradire, pentru ca deplasarea nu se poate produce oricum, ci numai în conditiile în care anumite reguli foarte generale, numite *principii*, sunt respectate.

Nu toti adeptii generativismului au fost însa multumiti cu noua propunere. Ceea ce s-a reprosat a fost ca transformarile ramâneau mai departe mecanisme prea puternice - în ciuda îngradirilor si a simplificarilor – deoarece ele operau pe un domeniu prea larg: cel al structurilor sintactice. O alta obiectie viza temeiurile mentale ale operatiei de deplasare: în ciuda plauzibilitatii aparente a acestei ipoteze, nu exista dovezi - sustineau criticii - ca mintea implicata în utilizarea limbajului ar face uz de o astfel de operatie. În sfârsit, existau cercetatori care considerau ca noul model de gramatica universala era greoi din punct de vedere computational, tocmai din cauza operatiei de deplasare: anume, pentru fiecare deplasare de constituenti, este necesara o verificare a compatibilitatii dintre principii si deplasarea constituentului.

În ansamblu, divergentele legate de conceptul de transformare au pregatit cea mai mare ruptura pe care a cunoscut-o în istoria sa curentul gramaticii universale. Criticii radicali ai conceptului de transformare au propus renuntarea la acest mecanism, propunere pe care Chomsky si cei ce l-au urmat nu au acceptat-o niciodata. Începând cu anul 1981, ruptura se oficializeaza. Apar pe rând Gramatica Lexico-Functionala (LFG - Bresnan si Kaplan), Gramatica Sintagmatica Generalizata (GPSG - Gazdar, Klein Pullum si Sag), Gramatica Arborilor Adaugati (TAG - Joshi), Gramatica Centrilor de Sintagma (HPSG – Pollard si Sag), Gramaticile Categoriale de Unificare (CUG- Uzkoreit)

2. Caracteristicile GNT

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

- Exploateaza în mod generalizat reprezentarile în termeni de trasaturi
- Fac recurs la mecanismul unificarii
- Se bazeaza pe constrângeri
- Sunt gramatici lexicaliste
- Au adecvare computationala

2.1. Reprezentari: structurile de trasaturi

Reprezentarile în termeni de trasaturi sunt bine cunoscute în lingvistica moderna, datorita fonologiei si semanticii structurale. GNT au meritul de a fi generalizat aceasta notatie la scara întregii teorii lingvistice. Prin perechea trasatura (atribut)–valoare, orice fel de informatie lingvistica – fonologica, morfologica, sintactica semantica, pragmatica – îsi gaseste o reprezentare adecvata. Câteva exemple: notatia [P(arte de)V(orbire): nume] spune ca o anumita entitate lingvistica este un nume. Reprezentarea [F(orma)V(erbala):

gerunziu] precizeaza ca avem a face cu un verb la gerunziu; reprezentarea [RAM(ura): v(aloare)n(on)v(ida)] spune ca obiectul lingvistic în chestiune are structura interna si este prin urmare o sintagma. Este usor de remarcat ca notatia atribut-valoare aplica principiul general al functiilor: unui anumit atribut îi corespunde o anumita valoare, întocmai cum unui argument dat îi corespunde o anumita valoare, datorita unei legi specifice de corespondenta. Reprezentarile de care se face uz în GNT sunt denumite *structuri de trasaturi.*

2.2 Unificarea

GNT se mai numesc si gramatici de unificare. Unificarea are drept obiect structurile de trasaturi. Unificarea a doua structuri de trasaturi A si B (notata A ∪ B) este structura minimala de trasaturi care cuprinde în acelasi timp si pe A si pe B. Daca o astfel de structura nu exista, unificarea "esueaza" (ceea ce e notat cu ⊥). Unificarea verifica asadar compatibilitatea dintre doua structuri de trasaturi si produce o structura rezultanta care contine toata informatia din structurile supuse unificarii. Iata câteva exemple:

(1) [CAT: det] ∪ [CAT: nume] = ⊥ (esec)

(2) [CAT: det] ∪ [ACORD: [NUM: sing]] = CAT: det, ACORD: [NUM: sing]
CAT: nume

(3) CAT:nume ∪ [ACORD: [GEN: masc]] = ACORD: GEN: masc
ACORD: [NUM: sing] NUM: sing

Operatia de unificare din primul exemplu esueaza pentru ca structura rezultanta ar trebui sa contina atributul CAT cu doua valori diferite (determinator si nume). Unificarea se realizeaza normal în (2) si (3), si produce o structura mai complexa.

Se poate remarca faptul ca rolul unificarii este acela de a explica tot ceea ce este corect în variate compartimente de limba. Daca are loc o unificare de informatii fonologice, aceasta explica un aspect al corectitudinii fonologice pentru o limba data. O unificare de informatii morfologice da seama de un aspect al corectitudinii morfologice, s.am.d. Nu e însa exclusa nici unificarea de informatii diferite, de exemplu, semantice si morfologice, semantice si sintactice etc.

2.3. Constrângeri

În exemplul (1) din paragraful precedent, unificarea esueaza deoarece nici o structura de trasaturi nu poate avea valori diferite pentru acelasi atribut. Aceasta este o "lege" inerenta unificarii, tot astfel cum în logica bivalenta o "lege inerenta" este tertiul exclus. Se poate spune ca (1) defineste o limita a unificarii si implicit o constrângere asupra acestei operatii. Constrângerea este de natura formala, pentru ca deriva din natura însasi a unificarii. Dar pentru scopurile unei teorii lingvistice, astfel de constrângeri nu pot fi

suficiente. Polona, de pilda, face la verbele de persoana I deosebirea între verbele folosite de un barbat si cele folosite de o femeie. Verbul are asadar gen în polona, dar nu si în româna. Pentru a face aceasta diferenta între cele doua limbi trebuie sa se admita ca unificarea informatiei de gen cu cea de verb se poate face în polona dar nu se poate face si în româna. Numai ca de aceasta data constrângerea privind unificarile nu mai are temei formal. Nu se poate spune ca în mod necesar verbul are sau nu gen. Unificarile acestor informatii sunt prin urmare "contingente", sau cu un alt termen, "empirice", tocmai pentru ca ele nu deriva din natura însasi a operatiei. Gramatica unei limbi se descrie mai ales în termenii unificarilor "contingente".

2.4 Lexicalism

În teoriile contemporane ale gramaticii, lexicalismul este o optiune privitoare la modul în care este conceputa structura cuvintelor în relatia lor cu sintaxa. Exista teorii, precum GB, care considera ca procesul de constituire morfologica a cuvintelor are loc în sintaxa. În acest sens, GB este o morfosintaxa deoarece generalizeaza operatia de deplasare la nivelul morfologiei însesi, prin mecanismul numit "deplasare centru-centru" (engl, "Head to Head Movement"). Gramaticile de unificare adopta o strategie distincta: ele considera ca procesele de constituire morfologica a cuvintelor sunt independente de sintaxa. În aceasta perspectiva, rezultatul proceselor morfologice furnizeaza sintaxei inputul necesar: cuvintele gata formate. Modularizarea celor doua componente ale gramaticii se dovedeste preferabila mai ales în cazul limbilor cu morfologie bogata.

Un alt aspect al lexicalismului asumat de GNT este ilustrat de modul în care sunt construite explicatiile de gramaticalitate. Explicatiile în GNT se sprijina în masura posibilului (dar într-o masura mult mai mare decât în alte teorii) pe proprietatile cuvintelor. În istoria generativismului, pasivul, de pilda, a fost considerat multa vreme o structura explicabila *sintactic*, adica o constructie rezultata din transformari ale unei alte structuri sintactice. GNT afirma însa ca nu e nevoie sa se recurga la structuri sintactice anumite, deoarece toate elementele de care e nevoie pentru a explica o constructie pasiva pot fi codificate la nivelul cuvintelor[62]. Un tratament asemanator poate fi observat în cazul dependentelor la distanta, sau în cel al constructiilor de ridicare (engl. "raising"), unde rolul unitatilor lexicale în determinarea proprietatilor acestor constructii este de asemenea semnificativ.

[62] *Preferinta aceasta pentru un compartiment de limba în defavoarea altui compartiment, atunci când se pune problema mecanismelor care justifica o anumita constructie nu e înteleasa înca nici azi de unii lingvisti. Este vorba de aceia care cred ca a avansa o explicatie lexicalista atunci când exista deja una sintactica pentru un fenomen oarecare înseamna doar a propune variatiuni pe aceeasi tema. Diferentele sunt în realitate cruciale si privesc mecanismele cognitive angajate în utilizarea limbajului. Este deja cunoscut ca procesarea unitatilor lexicale este mai usor de efectuat decât unele dintre procesarile structurilor sintactice. Acest fapt ofera un criteriu valoros de judecare a plauzibilitatii unei gramatici privite din unghi cognitiv.*

2.5 Adecvare computationala

În lingvistica, o teorie este considerata adecvata, daca teoria acopera domeniul de fapte pentru care este construita ca o explicatie. O morfologie a unei limbi, de pilda, este adecvata daca prin regulile propuse da seama de constructiile morfologic corecte ale limbii supuse analizei.

Acest principiu foarte general a fost nuantat de catre Chomsky. Nuantarea este deja celebra: pornind de la ideea ca utilizarea limbajului este o proprietate a mintii omenesti, Chomsky a sustinut ca o teorie trebuie socotita adecvata nu doar pentru ca produce explicatii ale cazurilor de corectitudine, ci si pentru ca mecanismele utilizate sunt dovedite (sau cel putin presupuse) a fi însusite de catre mintea omeneasca. Quine afirmase ca daca avem doua gramatici care cu mijloace diferite explica aceeasi realitate lingvistica, nu exista criterii suplimentare de alegere a uneia dintre ele. Chomsky a replicat ca un astfel de criteriu exista totusi, el fiind masura în care fiecare dintre aceste gramatici se foloseste de operatii cunoscute ca apartinând mintii în procesele ei cognitive.

Criteriul suplimentar formulat de Chomsky în evaluarea teoriilor lingvistice a apropiat comunitatea generativistilor de cea a psihologilor si a impulsionat cercetarile de psiholingvistica. S-au obtinut rezultate interesante si s-au construit ipoteze neasteptate. De pilda, regulile de constituenti sînt socotite astazi niste operatiuni cu mare probabilitate de a fi folosite de inteligenta umana. Recursivitatea este si ea considerata a fi o proprietate de care inteligenta umana face uz în utilizarea limbajului.

Criteriul lui Chomsky a condus însa si la cercetari cu rezultate greu de judecat. De pilda, despre realitatea psihologica a *urmelor*, concept cardinal al teoriei GB, s-a argumentat si pro si contra, si este foarte dificil chiar si azi sa se poata lua o pozitie.

Un lucru este cert totusi în evolutia raporturilor dintre teoria lingvistica si realitatea ei psihologica: comparativ cu faza de început, interesul psihologilor si al psiholingvistilor fata de ipotezele venite din comunitatea "chomskyenilor" a scazut semnificativ. A crescut însa interesul psiholingvistilor pentru ipotezele venite din lumea inteligentei artificiale. Este celebra în acest sens ipoteza de organizare a cunostintelor lexicale a lui Quillian, care a atras atentia în mod special colectivitatii de psihologi si de psiholingvsti. Un al treilea factor intra astfel în joc, rezultatul fiind ca unele teorii lingvistice au devenit atente la operatiile si mecanismele utilizate de inteligenta artificala. Erau exact teoriile generative netransformationale. Consecinta principala a acestei deplasari de interes a fost ca teoriile în cauza au devenit accesibile utilizarii automate. Cu alte cuvinte - si spre deosebire de gramaticile lui Chomsky - ele pot fi implementate computational.

Vom numi adecvarea unei teorii la domeniul de fapte pe care îl abordeaza *adecvare lingvistica*. Masura în care o teorie lingvistica apartine (sau poate fi presupusa a apartine) mintii omenesti defineste *adecvarea* ei *psihologica*. Iar gradul în care ea este livrabila inteligentei artificiale indica *adecvarea* ei *computationala*. Directia actuala a curentului de idei pare sa fie urmatoarea: legaturile si dialogul dintre psihologia cognitiva si inteligenta artificiala sunt într-o continua crestere, astfel încât adecvarea computationala a

unei teorii lingvistice are sanse mari sa-i confere si adecvare psihologica. Pe aceasta directie sunt plasate gramaticile generative netransformationale.

3. Realizari

Una dintre cele mai importante realizari ale gramaticilor nontransformationale îl reprezinta numarul mare de aplicatii. O enumerare a limbilor supuse analizelor nu este posibila aici, dar se poate preciza ca aproximativ doua treimi din familiile de limbi (considerate in esantioanele lor reprezentative) au fost analizate din perspectiva netransformationala. Este caracteristic acestor analize faptul ca refuza deosebirea chomskyana centru-periferie („core-periphery"). Ele se concentreaza asupra varietatii de date oferite de corpusuri.

Ceea ce este însa cel mai important sub aspectul realizarilor este faptul ca GNT au reusit sa produca replici viabile la analizele paradigmei dominante, cea chomskyana. O serie de fenomene gramaticale – privite de obicei ca fiind de la sine caracterizabile prin mecanismul deplasarii constituentilor – au primit in cadrul GNT analize alternative. Asa s-a întâmplat cu constructiile pasive, cu fenomenul de ridicare (si mai general cu fenomenele de depedenta limitata), cu constructiile nonlocale (precum topicalizarile, structurile relative si interogative). In aceasta privinta, GNT au continuat traditia fireasca, inaugurata de structuralism, traditie constând in regândirea fenomenelor de limba odata cu fiecare noua scoala lingvistica.

4. Forme institutionale de sustinere

GNT sunt bine reprezentate institutional. Ele si-au facut loc în primul rând în programele curiculare ale unor universitati 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 scolile de vara. O prestigioasa scoala de acest fel („European Summer School in Logic Language and Information" – ESSLLI) este organizata anual din 1989, cu rolul de diseminare a evolutiilor si curentelor formate în interiorul gramaticilor netransformationale. Este apoi de semnalat, în aceeasi linie a „didacticii" gramaticilor nontransformationale, nou înfiintata scoala de vara de la Konstaz (Germania).

În planul congreselor stiintifice, HPSG si LFG au de multa vreme propriile lor conferinte anuale. Iar un congres tinut o data la doi ani - cel de gramatici formale - urmareste sa adune sub acelasi acoperis toate scolile aceleiasi familii.

Pâna de curând, gramaticile nontransformationale nu au avut o revista proprie. Lucrarile însa au fost si sunt publicate in reviste de prestigiu, precum „Computational

Linguistics" „Natural Language and Linguistic Theory", „Journal of Linguistics", „Language" sau „Langages". O revista orientata explicit spre aceste gramatici este editata de putina vreme la cunoscuta editura olandeza Kluwer. Este vorba despre revista „Grammars". De asemenea, pe lânga Centrul de Studii asupra Limbajului si Informatiei de la Universitatea Stanford exista de mai multa vreme o deja celebra editura care publica lucrarile esentiale ale domeniului.

5. Gramaticile nontransformationale în România

Prezenta GNT în Romania poate fi discutata având în vedere doua coordonate: cea a contributiilor stiintifice si cea a programelor curiculare.

Din primul punct de vedere, întâia contributie (dupa cunostinta noastra, cel putin) a venit din partea Adrianei Costachescu ([14]). Adriana Costachescu este autorul unui studiu, din perspectiva GPSG (teorie care a precedat si inspirat HPSG), asupra relatiei dintre coordonarea adversativa si subordonarea concesiva. Studiul a fost elaborat in 1993 si publicat in 1996.

Lucrari de prezentare generala a diferitelor forme de GNT sau, dimpotriva, de prezentare a trunchiului comun – unificarea – au fost publicate în ultimii sase ani de Adrian Atanasiu, Verginica Barbu, Ana-Maria Barbu, Florentina Hristea, Emil Ionescu si Rodica Tatar.

Printre „pionierii" aplicatiilor acestor gramatici la limba româna trebuie mentionati Liviu Ciortuz si cercetatoarea italiana Paola Monachesi. Amândoi au folosit teoria HPSG. Rolul lui Monachesi în stimularea aplicatiilor de acest tip la limba româna trebuie în mod special subliniat. Studiile sale asupra cliticelor pronominale din româna au determinat o „mobilizare" a energiilor câtorva cercetatori români. Este vorba despre Ana-Maria Barbu, Emil Ionescu si Amalia Todirascu.

Ana-Maria Barbu a aplicat HPSG în analiza elementelor gravitând în jurul verbului – adverbul de negatie, semiadverbele, auxiliarele – si a ajuns la concluzia ca acestea sunt mai apropiate de afixe decât de cuvinte. Concluzia analizei se întâlneste cu concluzia exprimata în lucrarea Valeriei Gutu Romalo, „Morfologie structurala a limbii române", în care formele compuse ale verbelor sunt considerate forme cu afix mobil.

O alta contributie a Anei-Maria Barbu priveste ordinea constituentilor in grupul nominal. Valorificând sugestiile de analiza ale lui Valerio Allegranza, Ana-Maria Barbu a propus o clasificare a constituentilor grupului nominal, care este relevanta pentru problema ordinii acestora. Analiza produce astfel solutii clare si eficiente într-o problema complicata de gramatica a limbii române.

Semnalând unele neajunsuri în analiza GB a fenomenului de anticipare clitica a complementului direct nominal în româna, Verginica Barbu si Emil Ionescu propun o abordare alternativa HPSG. Analiza poate fi extinsa si la alte limbi care prezinta fenomenul în cauza. Analiza sustine ca pronumele neaccentuate nu au un comportament uniform,

proprietatile lor depinzând de faptul daca participa sau nu la structuri de dublare. Noutatea abordarii vine din faptul ca fenomenul anticiparii obiectului direct este în mod ultim justificat prin proprietatile lexicale ale verbului tranzitiv.

Un fenomen care, în aparenta cel putin, implica recursul la mecanismul deplasarii – este vorba de prezenta pronumelor neaccentuate în acuzativ în contexte în care ele nu sunt subordonate fata de vreun element din acel context – este tratat într-un alt studiu asupra cliticelor pronominale românesti[63] (). Studiul arata ca ipoteza deplasarii constituentilor nu este necesara în analiza fenomenului. Este propusa în alternativa o analiza fara deplasari care capteaza toate proprietatile fenomenului.

O analiza HPSG este propusa de asemenea pentru fenomenul negatiei duble si multiple în româna (). În sfârsit, Amalia Todirascu abordeaza într-unul din studiile sale asupra limbii române, o categorie de dependente limitate (asa-numitele *tough-constructions*), din aceeasi perspectiva HPSG.

În aceeasi linie a contributiilor stiintifice, merita amintita o intiativa institutionala: acreditarea de catre CNCSIS, în anul 2001, a Centrului de Lingvistica Computationala de pe lânga Facultatea de Litere. Centrul este perechea universitara a Centrului de Studii Avansate în Inteligenta Artificiala. Aparitia sa a fost semnalata în buletinul european ELSNEWS. Unul dintre programele de cercetare pe anul 2002 ale centrului are în vedere dezvoltarea aplicatiilor de gramatici netransformationale la limba româna.

În planul programelor curiculare, GNT si-au facut loc mai greu, si au fost întâmpinate uneori nu doar cu neîncredere, ci si cu ostilitate. A existat însa din fericire un sprijin substantial si constant al factorilor de decizie. Ne referim la decanul Facultatii de Litere, acad. prof. Dan Horia Mazilu, la rectorul Universitatii Bucuresti, prof. dr. Ioan Mihailescu, la prorectorul aceleiasi institutii, prof. dr. Ioan Pânzariu, si la acad. Dan Ioan Tufis, directorul Centrului de Studii Avansate în Inteligenta Artificala al Academiei Române, carora autorul acestor rânduri le exprima via si profunda sa gratitudine, pentru sustinerea pe care a simtit-o mereu în initiativele sale. Multumita acestui sprijin, au devenit realitate câteva proiecte care pot fi considerate succese:

- În programa cursurilor optionale de limba pentru anul al IV-lea al Facultatii de Litere a fost introdus în 1996 un curs introductiv de GPSG, iar din 1997 pîna în 2001 s-a tinut un curs introductiv de gramatici de unificare cu referire speciala la HPSG.
- Din 1999, se preda la Facultatea de Matematica a Universitatii din Bucuresti un curs optional de un an de prelucrare automata a limbii naturale, în care un loc important îl ocupa gramaticile de unificare.
- Din 1997 pîna în prezent masteratul de lingvistica teoretica al Facultatii de Litere din cadrul aceleiasi universitati gazduieste un curs de un semestru de teorie HPSG aplicata la limba româna.

---

[63] *În engleza, fenomenul este cunoscut sub numele de „clitic climbing", si este ilustrat în româna de structuri de tipul* ***Nu-l pot suferi pe Ion****.*

- Din 1999, acelasi masterat ofera un seminar de gramatici cu implementare computationala.
- În anul 2000, un proiect de dezvoltare a componentei de lingvistica computationala în cadrul masteratului de lingvistica teoretica a primit sprijin de finantare din partea Bancii Mondiale si a Guvernului României, sprijin care a facut posibile printre altele organizarea unor cicluri de conferinte pe teme de GNT (în special HPSG) la Facultatea de Litere a Universitatii Bucuresti. Au conferentiat Ivan Sag (Universitatea Stanford), Anne Abeille si Daniele Godard (Universitatea Paris 7), Stefan Müller (Universitatea din Jena), Robert Malouf (Universitatea Groningen), Howard Gregory (King's College, Londra), Erhard Hinrichs (Universitatea Tübingen), toti fiind personalitati recunoscute ale domeniului. Multumita aceluiasi program, cercetatorii români au putut petrece stagii de specializare la universitatile din Lille si Stanford, sau au putut participa la manifestari reprezentative, cum ar fi colocviul UNESCO asupra spatiilor virtuale si multilingvismului de la Paris (aprilie 2001), colocviul de gramatici bazate pe constrângeri Trondheim (august 2001), sau congresul de prelucrare automata a limbilor naturale de la Tokyo, (noiembrie, 2001). Cea mai importanta realizare legata de acest program, a constat însa în posibilitatea unor mobilitati studentesti, concretizate în vizitele de studiu ale studentilor masteratului de lingvistica teoretica, la universitatile din Darmstadt, Tübingen, Paris 7 si Sienna.

6. Concluzii

Desi GNT au patruns în mediile stiintifice din România mai târziu decât în alte tari, faptul ca ele sunt prezente la noi este un lucru încurajator. Exista tentatia de a privi aceste eforturi de sincronizare cu miscarea de idei din domeniul lingvisticii formale drept tentative mimetice si superficiale. Este o greseala grava. Diversele comunitati de lingvisti pot desigur ignora un curent, precum cel prezentat mai sus, dar aceasta este o atitudine, pentru a spune asa, pe proprie raspundere. GNT si teoria lingvistica pe care ele au inspirat-o si-au facut deja loc în lingvistica zilelor noastre si au devenit una din paradigmele majore. În plus, dubla deschidere a acestor gramatici catre psihologia cognitiva, pe de-o parte, si catre inteligenta artificiala, pe de alta parte, recomanda aceasta paradigma drept cadrul *privilegiat* de dialog interdisciplinar din stiintele umaniste ale contemporaneitatii. Din acest triunghi, sunt asteptate sa apara noi aplicatii – unele au si aparut deja - care vor extinde într-un mod neasteptat conceptul de lingvistica aplicata. Pentru toate aceste motive, tentativele de a pastra un contact viu si de perspectiva cu comunitatea stiintifica a GNT reprezinta o investitie sigura pe temen lung.

Bibliografie

[1] Abeillé, A. *Les nouvelles syntaxes. Grammaires d'unification et analyse du français*, Armand 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) "Informal 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*) Informal 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, München, 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*, München, 1996, p. 121-134

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

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

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

[22] Idem,, *Accusative Clitic Climbing in Romanian*, "Cahiers de Linguistique Théorique et Appliquée", 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", vol. 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 Placement in the Romanian Verbal Complex,* Gerlach and Grijzenhout (2000), p. 255-294.

[30] Pollard, C. si 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. *Inteligenţă artificială*, Editura Albastră, Cluj, 2001

[34] Todiraşcu, A. *Romanian **Tough**-Constructions*, Ciortuz, L, P. Monachesi, şi H. Uszkoreit (editori*) Informal 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

Catre o teorie X-bar functionala

Neculai CURTEANU
Institutul de Informatica Teoretica, Academia Româna, Filiala Iasi
curteanu@iit.tuiasi.ro

1. Teorii X-bar mai vechi si mai noi

Scopul prezentei lucrari este dublu: (***a***) de a propune o noua X-bar schema, numita X-bar schema *functionala* si *recursiva* (pe scurt, FX-bar schema), mai generala si mai adecvata decât cele existente, care sa satisfaca cerintele unei abordari functionale a limbajului natural (LN), în particular, ale strategiei lingvistice SCD (Segmentare-Coeziune-Dependenta) [1], [2], si (***b***) de a pune în evidenta faptul ca teoria FX-bar propusa poate reprezenta o posibila (si necesara) solutie la urmatoarea problema ridicata de Noam Chomsky în teoria Minimalist Program [3]: în doua capitole diferite, Chomsky afirma (în doua abordari diferite, aparent contradictorii, asupra structurii sintactice a LN) atât importanta crescânda a teoriei X-bar cât si posibilitatea ca teoria X-bar standard sa fie "*largely eliminated in favor of bare essentials*" (vezi sectiunea 5).

1.1. Teoria X-bar clasica

Printre (sub)teoriile care reprezinta substanta majora pentru câteva teorii formale importante asupra sintaxei (LN), un rol fundamental este jucat de catre asa-numita teorie X-bar. X-bar schemele propuse sunt de obicei însotite de definitii, ipoteze, restrictii, principii si alte (sub)teorii gramaticale care specifica într-o cât mai mare masura modul concret în care X-bar schemele sunt utilizate pentru a construi structurile sintactice de baza ale LN. În general, teoria X-bar stabileste categoriile gramaticale principale, proiectiile lor lingvistice (minimale si maximale), relatiile de dominare dintre categorii în cadrul acestor proiectii, sub–, co–, sau supra-ordonarea lor. Toate aceste aspecte asigura numai coloana vertebrala (infrastructura) consistenta a structurii sintactice în reprezentarea LN. Un capitol de o importanta deosebita este relatia dintre teoria X-bar si alte sub(teorii) sintactice si semantice care formeaza întregul corpus al unei anumita teorii lingvistice.

Prima forma a X-bar teoriei este propusa de catre Noam Chomsky în lucrarea *Remarks on Nominalizations* (1970) [4]. Chomsky scoate în evidenta diferentele reale existente în urmatoarele sintagme nominale:

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

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

în special datorita sablonului verbal (similar cu al verbului "*criticize*") rezultat din gerunziul nominal (pentru engleza) "*criticizing*", în comparatie cu forma nominala derivata "*criticism*".

Teoria X-bar originala propusa de Chomsky identifica trei categorii lexicale *primitive*, N [Eng: *noun*], V [Eng: *verb*] si A [Eng: *adjective*], fiecare dintre ele cu câte doua categorii sintagmatice corespunzatoare. Mai exact, utilizând notatia X = N, V, A, categoria gramaticala X se întâlneste ca *nucleu* [Eng: *head*] într-o categorie intermediara X' (sau X1, sau X^1), traditional numita X-bar, precum si într-o categorie maximala X" (sau X2, sau X^2), traditional numita XP, reprezentând *proiectia maximala* a categoriei gramaticale X (lexicala sau nelexicala). Categoria X este numita *nucleul sintagmelor* X' (sau X1) si X" (sau X2) care o contin. Sa mai notam ca prescurtarea pentru categoria *prepozitionala* este P.

Ulterior au fost considerate *patru* categorii lexicale, bazate pe urmatoarele combinatii ale celor doua trasaturi N si V (considerate ca fiind generice pentru categoriile lexicale):

N	este o categorie X cu trasaturile [+N, –V];
V	este o categorie X cu trasaturile [–N, +V];
A	este o categorie X cu trasaturile [+N, +V];
P	este o categorie X cu trasaturile [–N, –V].

Teoria X-bar poate fi înteleasa si ca o specificare a modalitatii în care unele categorii gramaticale sunt *dominate* de catre altele, deci ca o teoriei a dominantei gramaticale (sau, asa cum spune Chomsky, a "*guvernarii*"), care arata cum un nucleu (sau o categorie lingvistica) X se proiecteaza (se extinde) catre categoriile mai complexe (structurile sintagmatice) X' (sau X1) si X" (sau X2, sau XP). Structurile sintactice X1 su X2 devin categorii esentiale ale organizarii si reprezentarii textului în LN.

Deci, X-bar teoria clasica considera ca X, împreuna cu o secventa de *complemente* (sau *argumente*, notate Arg_i) este imediat dominata de X1, în timp ce X1 împreuna cu o secventa de *specificatori* (notata $Spec_j$) este imediat dominata de catre X2 (sau XP). Utilizând binecunoscutele notatii din domeniul teoriilor lingvistice formale, (X' = X1, X" = X2 = XP), categoriile lexicale si gramaticale ale teoriei X-bar clasice a lui Chomsky sunt urmatoarele:

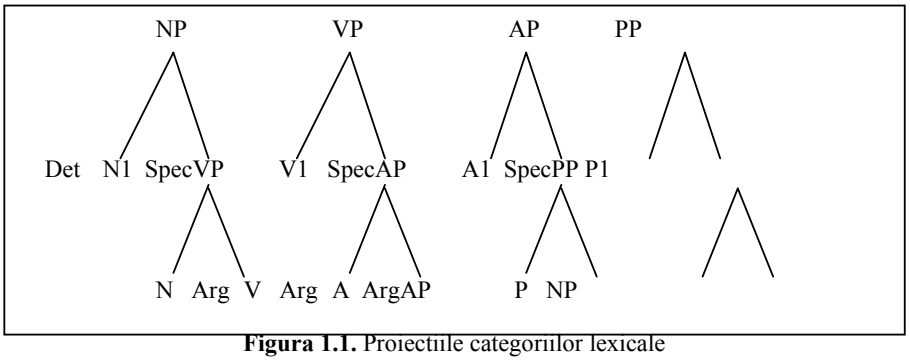


Figura 1.1. Proiectiile categoriilor lexicale
din teoria X-bar clasica

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

Stowell [5] propune ca teoria X-bar clasica sa fie extinsa la categorii nelexicale sau *functionale*. În particular, categoria gramaticala S [Eng: *sentence*; Rom: *fraza*], care corespunde uneia sau mai multor propozitii gramaticale (clauze), este vazuta ca I2 sau IP, deci ca proiectia maximala a categoriei nelexicale "I", sau INFL [Eng: *Inflectional*]. Nucleul nelexical I (INFL) reprezinta multimea de trasaturi de flexionare atribuite nucleului lexical al clauzei-matrice (propozitia principala, sau chiar una regenta) dintr-o fraza, asa cum sunt timpul, aspectul etc. în clauza a unei fraze. Remarcam *categoria* S, care introduce un anumit grad de ambiguitate în analiza gramaticala, atât în engleza cât si în româna. Termenul adecvat pentru realitatea lingvistica codificata de categoria S ar trebui sa fie acela de "*clauza gramaticala*" pentru engleza [Eng: (*grammatical*) *clause*], si de "*propozitie gramaticala*" pentru limba româna, cu doua sorturi principale: *clauza finita*, prescurtata CLF sau mai simplu CL, si *clauza infinita*, prescurtata CLI.

Astfel în extensia nelexicala a teoriei X-bar, S este proiectia (lingvistica) maximala a categoriei virtuale (nelexicale) I, în timp ce S1 este vazuta ca fiind C2, sau CP, unde nucleul C este un *complementizator*, o categorie gramaticala ce corespunde unei expresii (unui delimitator) sau unei sintagme care introduce o clauza subordonata, *e.g.* pronume relativ, conjunctie, locutiune conjunctionala etc. Teoria *X-bar extinsa* acrediteaza urmatoarele structuri:

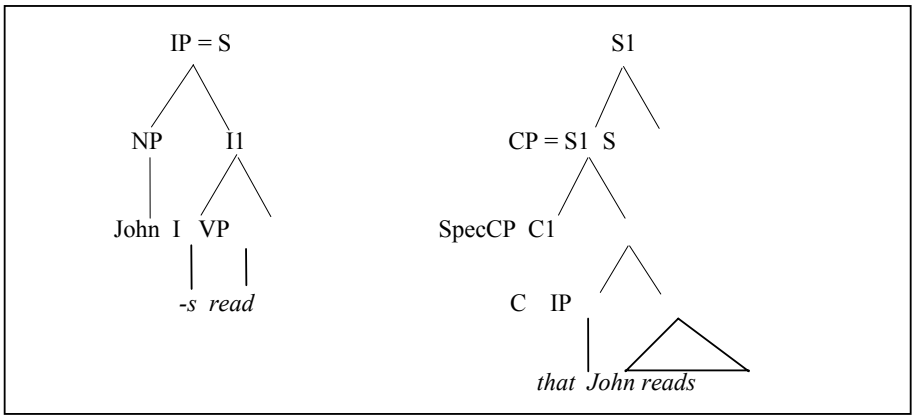


Figura 1.2. Teoria X-bar extinsa la categorii nelexicale

Sunt necesare câteva remarci:

(a) Teoria X-bar extinsa utilizeaza terminologia de "categorii nelexicale (sau functionale)", prin care Stowell, Chomsky si alti lingvisti definesc *noile nuclee* ale structurilor sintactice considerate. Categoria virtuala "I" este, desigur, una nelexicala, si sustine o anumita functionalitate depinzând de categoria lexicala careia îi este atribuita. Categoria C nu este, de obicei, nelexicala (exceptând situatia, posibila, când ea lipseste) deoarece C corespunde unor categorii gramaticale lexical nevide. În ceea ce priveste functionalitatea lui C, suntem de acord ca C corespunde într-adevar unor functii si relatii sintactice si semantice importante pe care le numim *marcheri de propozitie* (subordonate) [1], [2], [6], uneori incluse în clase mai largi cum sunt cea a *marcherilor de discurs* [7], reprezentând în acelasi timp si un element (deci o relatie) de co-referinta în cadrul fenomenului de *legare*, si/sau o "*bariera*" [8] în cadrul *teoriei limitarii* [9]. Aceste aspecte multi-functionale ale categoriei C nu sunt contradictorii ci doar complementare, întregind un tablou complex al functionalitatii lexical-semantice pentru o categorie lingvistica atât de speciala cum este C.

(b) A doua observatie este dedicata rolului unor categorii nelexicale în cadrul X-bar schemelor extinse. Din Fig. 2. reiese ca subiectul NP are rolul (nesigur) al unui specificator pentru S = IP, în timp ce VP reprezinta complementul categoriei virtuale I. De asemenea, S1 = CP se considera a fi proiectia maximala a categoriei C, în timp ce complementul sintagmei CP este IP. Admitând ca în engleza, din punct de vedere sintactic, aceasta supozitie are sens deoarece categoria C reprezinta nucleul acestor sintagme, în alte limbaje, inclusiv româna, acest lucru este nedecis, în special din perspective semantice si functionale. Unele abordari functionale ale acestor probleme sunt discutate în mai multe lucrari, dar ne vom restrânge sa mentionam aici la a mentiona solutiile oferite de catre *teoria gramaticii functionale* [10] si *strategia lingvistica* SCD [1], [2], [6]. Un interes

special prezinta abordarea *lexicala* (inclusiv *functionala*) a teoriei X-bar ca subteorie de baza în cadrul teoriei sintactice HPSG [Eng: *Head-driven Phrase Structure Grammar*] [11]. O analiza comparativa cu *FX-bar schema* propusa în aceasta lucrare va fi facuta 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 urmatoarele:

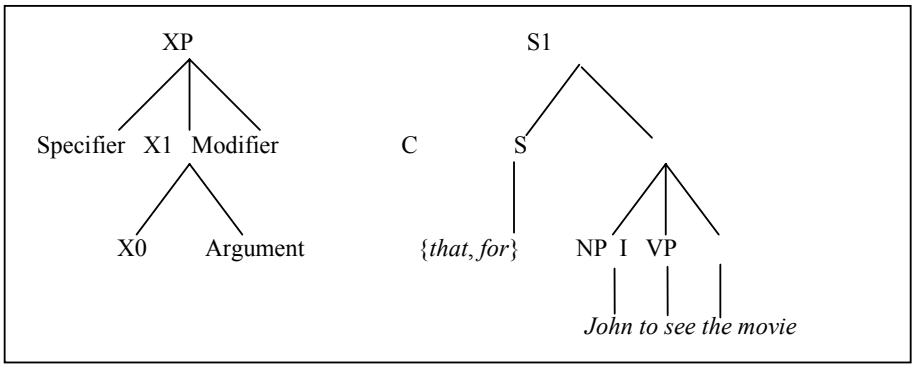


Figura 1.3. X-bar schema generala din GB, X = N, V, A, P, S

În teoria GB exista urmatoarele X-bar *echivalente* pentru proiectiile categoriilor gramaticale (lexicale si nelexicale).

Tabelul 1.3

Proiectii ale categoriilor lingvistice în GB

X	X1	X2
N	N1	NP
V	V1	VP
A	A1	AP
P	P1	PP
I	S	S1

În lucrarile GB [5] si cele care urmeaza, Chomsky considera categoria I ca fiind nucleul lui S, iar complementizatorul C ca fiind nucleul lui S1. În subsectiunea urmatoare, teoria sintactica GPSG a lui G. Gazdar [12] face un important pas înainte catre lexicalitate si catre utilizarea explicita a trasaturilor lingvistice atribuite categoriilor gramaticale.

1.4. Teoria X-bar în GPSG

În teoria lingvistica GPSG [Eng: *Generalized Phrase Structure Grammar*] [12], [13] etc., (sub)teoria X-bar joaca de asemenea un rol central, o sintagma a LN fiind definita ca proiectia *trasaturilor lingvistice* atribuite *nucleului* [Eng: *head*] acelei sintagme. Informatia cuprinsa în trasaturile nucleului determina caracteristicile principale ale comportamentului sintactic al sintagmelor LN. Reamintim ca o categorie sintactica în GPSG se reprezinta ca o multime de perechi <*trasatura*, *valoare*>. De exemplu, eticheta NP [Eng: *noun phrase*] (sau N2), prin care se noteaza o sintagma nominala, reprezinta o abreviere pentru multimea {<N, +>, <V, –>, <BAR, 2>}, unde BAR este *numele trasaturii* ce codifica *nivelul de proiectie* a categoriei sintactice N = {<N, +>, <V, –>}. Trasatura BAR poate lua valorile 0, 1, 2. Teoria GPSG considera N, V, A si 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 catre teoria GPSG ca având întotdeauna o *valoare* pentru *trasatura* BAR. Valoarea BAR pentru categoriile minore nu este definita niciodata în GPSG.

Teoria sintactica a GPSG aduce câteva elemente noi si interesante comparativ cu teoria GB: **(a)** X-bar schemele au, ca si în GB, trei nivele de proiectie (valorile trasaturii BAR); **(b)** Pentru economia reprezentarii, GPSG propune ca în X-bar schemele de baza, nivelul proiectiei lingvistice sa fie conservat când se trece de la nucleu catre expresiile subcategorizate, mai putin în cazul în care acest lucru se face prin (alte) reguli explicite; **(c)** Printr-un mecanism de *mostenire implicita*, nivelele BAR de proiectie a nodului-radacina si ale nodurilor-fiice ramân aceleasi, mai putin în cazul în care exista o indicatie contrara expresa.

O alta caracteristica este aceea ca 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 exista un nivel de proiectie pe care ele sa fie reprezentate (sub nivelul lexical BAR = 0). Consecinta este aceea ca, în GPSG, S este proiectia unei categorii V. Mai exact, proiectiile maximale ale lui V sunt VP, S, si S1, depinzând de urmatoarele valori luate de catre trasaturile SUBJ si COMP (= complementizator = C):

V[BAR 2][SUBJ –][COMP NIL] = VP;

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

V[BAR 2][SUBJ +][COMP α] = S1; unde $\alpha \in$ {*that*, *for*, *whether*, *if*}.

În sfârsit, trebuie sa remarcam ca GPSG trebuie sa rezolve problemele întâlnite în mod obisnuit în formalismele gramaticale bazate pe unificarea lingvistica (si/sau logica), de exemplu PATR-II [14], HPSG [15], [16] etc. O astfel de problema este, în particular, transmiterea informatiei despre *timpul verbului* între forma flexionara codificata de verb si nodul S. Pentru teoriile lingvistice care permit inserarea în arborele de derivare a cuvintelor flexionate, asa cum este cazul cu GPSG, HPSG etc., informatia despre forma flexionara trebuie sa poata fi mutata în ambele directii pe nivelele X-bar schemei. Din aceasta deriva,

în GPSG, conditia ca V sa fie *nucleul structurii clauzale* care corespunde categoriei S. Pe de alta parte, în GB, informatia asupra timpului unui verb poate fi transmisa dinspre nodul I catre proiectia sa în S înainte ca I sa fie combinat cu forma flexionata a verbului din S. Aceasta situatie poate produce potentiale dificultati procedurale si de reprezentare.

Este important de mentionat ca proiectiile categoriilor din Tabelul 4 ramân aceleasi pentru GPSG si LFG [Eng: *Lexical Functional Grammar*] (vezi de exemplu [13]), cu diferenta notabila ca prima celula din ultima linie a Tabelului 4 este goala, deoarece în aceste doua teorii lingvistice (ca si în altele), categoria virtuala I lipseste.

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

Vom propune în aceasta subsectiune o *formulare recursiva* a teoriei X-bar avându-si originea în *teoria barierelor* (TBarr) [8], [17] si fiind compatibila cu teoria sintactica a *Programului Minimalist* (MinP) [4] si cu modelul sau gramatical din *Principii si Parametri* (P&P) [4]. În conformitate cu MinP si P&P, gramaticile concrete ale limbajelor naturale (LNs) reale pot fi modelate de multimi de parametri si valorile lor, care specifica principii si teorii lingvistice universal valabile. Pentru o asemenea setare (asignare) a valorilor parametrilor, relatiile de precedenta (de ordonare liniara) dintre categoriile gramatice sunt obtinute din proprietati ca marcarea cazuala, atribuiri de roluri tematice ((θ-roluri si θ-marcheri), împreuna cu alte relatii si marcheri ce se aplica la nivelul sintagmelor, clauzelor, si discursului. Din acest motiv relatiile de precedenta pentru X-bar schemele propuse pot fi utilizate independent pe arborii sintactici considerati, informatia de ordonare (liniara) a categoriilor fiind data de urmatorii *parametri de precedenta*.

(**OrdPar**) *Un anumit parametru (depinzând de limbaj) specifica daca secventa de specificatori precede sau succede nucleul, iar un alt parametru (depinzând de limbaj) precizeaza când secventa complementelor precede sau succede nucleul din X-bar schema.*

De exemplu, în *engleza*, specificatorii preced de obicei nucleele lor nominale, în timp ce în *româna*, în mod normal, ei succed nucleele lor. În general, complementele (argumentele) succed nucleele lor si în engleza si în româna. Un caz special al argumentului este *subiectul* (sintactic). Aceasta exprimare a (OrdPar) poate fi înca particularizata în functie de categoriile lexicale concrete, din LNs concrete. De exemplu, atât în româna cât si în engleza, când o sintagma adjectivala (adverbiala) este *predicational activa*, fiind urmata de anumite argumente (complemente sau adjuncti), atunci este obligatoriu ca ea sa succeada propriul nucleu si nu sa îl preceada.

Consecinta principala a parametrizarii dependenta de limbaj a precedentei categoriilor lingvistice este ca în exprimarea teoriilor lingvistice se pot utiliza arbori neordonati, iar principiile propuse de teoria X-bar primesc un puternic caracter de independenta relativ la regulile structurilor sintagmatice. Este important faptul ca X-bar schemele obtinute în cadrul teoriei X-bar considerate sa asigure proiectii adecvate ale categoriilor lexicale, permitând inserarea *adjunctilor*, obtinerea categoriilor de proiectie

maximala, si acceptarea faptului ca unele proiectii minimale sau maximale din *structura de adâncime* pot fi vide (deci noduri care sa domine *categorii vide*), conform [9], [8], [17].

Fiind stabilit *principiul* (OrdPar), teoriile GB si Tbarr considera urmatoarele trei nivele ale proiectiei din teoria X-bar, sintetizate de urmatoarele *reguli* (*principii*) si de *X-bar schemele* corespunzatoare:

(PX0) Fiecare nod X0 dintr-o schema X-bar este fie vid, neavând nici o trasatura, fie este nodul-mama al unui element lexical a carei categorie gramaticala si trasaturi sunt specificate la nivelul lexiconului.

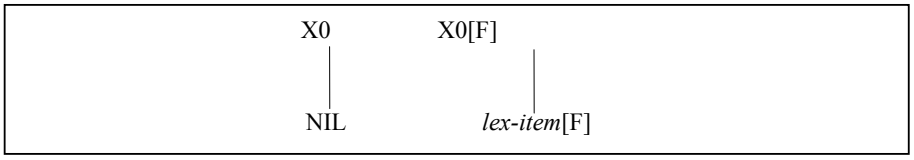


Figura 1.5.1. Nodul X0 în TBarr

(PX1) Fiecare nod X1 (X' sau X^1) având trasaturile lexicale F este fie nodul-radacina al exact unui nod X (care este *nucleu*) cu trasaturile F si al unei secvente de noduri XP (care sunt *complemente*, sau *argumente*), fie este radacina unui nod identic X1 împreuna cu exact un nod XP (care este *adjunct*).

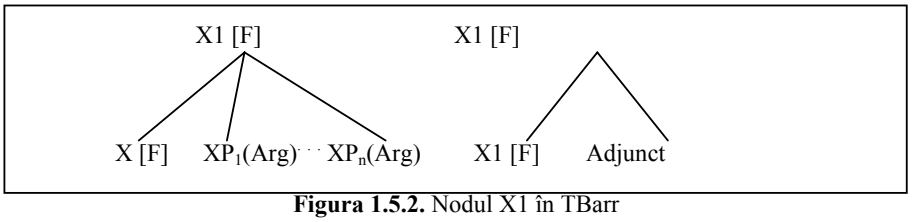


Figura 1.5.2. Nodul X1 în TBarr

(PX2) Fiecare nod XP care are trasaturile lexicale F trebuie sa satisfaca una si numai una din urmatoarele conditii: **(i)** XP este un nod-frunza (nu mai are nici un nod-fiica) si multimea F este vida; **(ii)** XP este radacina unei secvente de XPs (*specificatori*) si a exact unui nod X1 mostenind trasaturile F; **(iii)** XP este radacina unei secvente de XPs (*complemente*, sau *argumente*) si a exact unui nod X cu trasaturile F; **(iv)** XP este radacina unui alt nod XP mostenind trasaturile F si a exact unui nod XP.

O *observatie importanta* este aceea ca unele dintre *secventele* XP specificate în regulile (PX1) si (PX2) pot fi *vide*.

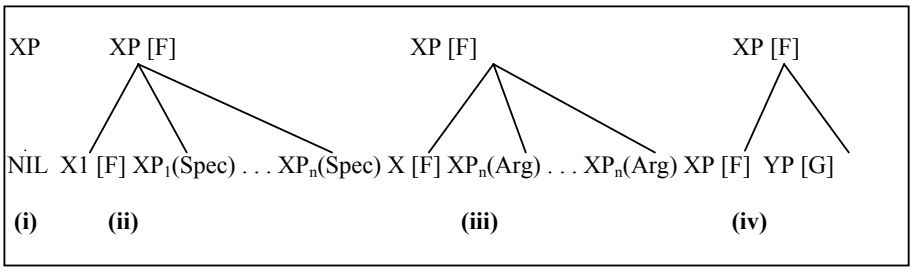


Figura 1.5.3. Nodul X2 în teoria TBarr

Combinând recursiv X-bar schemele rezultate din *regulile* (XP0)-(XP1)-(XP2) se pot obtine toate structurile sintactice întâlnite în *X-bar teoria clasica* si *extinsa*:

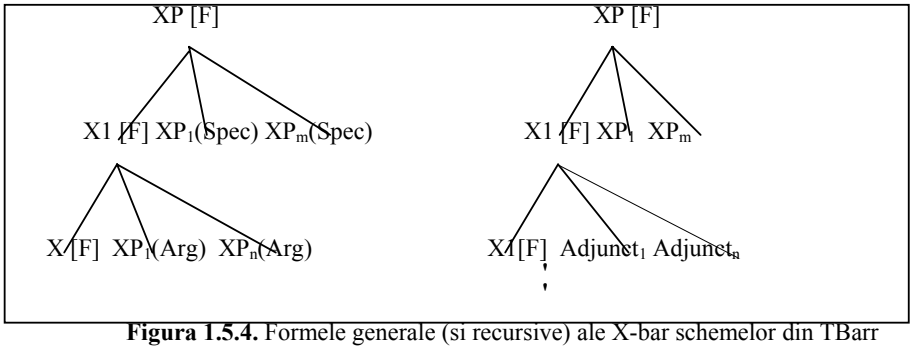


Figura 1.5.4. Formele generale (si recursive) ale X-bar schemelor din TBarr

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

2.1 Sistemul Chomskyan al gramaticii universale

Aceasta subsectiune contureaza câteva aspecte implicate de catre teoria X-bar în cadrul teoriilor MinP (*Minimalist Program*) si P&P (*Principles and Parameters*) [3]. Pentru a întelege contextul, este necesar sa schitam *teoria* lui *Chomsky* a gramaticii universale UG [Eng: *Universal Grammar*] si a relatiilor sale cu abordarea MinP bazata pe P&P [3]. Sunt introduse urmatoarele concepte de UG.

Capacitatea utilizarii si întelegerii LN se bazeaza în esenta 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 si a expresiilor pe care le genereaza ele reprezinta *Gramatica Universala* (UG).

Se considera ca UG specifica anumite *nivele lingvistice*, sau sisteme de reprezentare a informatiei lingvistice. UG a lui Chomsky [3] presupune ca fiecare SD este o secventa (δ, σ, π, λ) de *patru* reprezentari pe urmatoarele nivele, respectiv: *structura de adâncime* (D-structura), *structura de suprafata* (S-structura), *forma fonetica* (PF) si *forma logica* (LF). O ipoteza constructiva pentru UG este aceea ca limbajul este scufundat în *sisteme de*

performanta care permit ca exprimari în LN sa fie folosite pentru articulare, interpretare, referire, interogare, reflectie si alte actiuni, în timp ce SDs devin un complex de instructiuni pentru aceste sisteme de performanta.

O alta ipoteza standard pentru constructia UG este aceea ca un LN este format din doua componente: un *lexicon* si un *sistem computational*. Aceasta constructie este o inovatie esentiala comparativ cu teoria GB, care pretinde independenta sa fata de orice aspecte computationale sau de implementare. Lexiconul specifica elementele de intrare pentru sistemul computational, în timp ce acesta foloseste intrarile de lexicon pentru a genera derivari si SDs. Derivarea unei exprimari lingvistice particulare implica alegerea elementelor din lexicon si evaluarea construind perechea pe doua nivele de performanta, numite si *reprezentari de interfata*. Una din ipotezele de baza ale teoriei lui Chomsky *Minimalist Program* este aceea ca în constructia SD, utilizând lexiconul si sistemul de evaluare, sunt luate în considerare *numai doua* nivele de interfata, corespunzând lui PF (forma fonetica) si lui LF (forma logica), împreuna cu multimile de perechi (π, λ) rezultate din cele doua forme.

În abordarea P&P a teoriei lingvistice MinP, UG asigura un *sistem de principii* fixat, asociat cu un tablou finit de *parametri evaluati* (pe un numar finit de valori). Regulile pentru un LN particular se reduc la alegerea valorilor pentru acesti parametri. Notiunea de constructie gramaticala este eliminata, împreuna cu regulile particulare de constructie, specifice gramaticilor generative. Constructii ca VP, clauza relativa, pasivul etc. devin doar elemente ale unei taxonomii generale, sau colectii de fenomene explicate prin interactiunea principiilor de UG, legate (setate) cu anumite valori fixate ale parametrilor.

În sistemul computational al UG exista un set de *principii invariante*, fiecare cu un domeniu de *optiuni* restrânse la elementele functionale si proprietatile generale ale lexiconului. O *selectie* Σ printre aceste optiuni determina LN concret. În schimb, un limbaj determina o multime infinita de SDs lingvistice, fiecare pereche (π, λ) fiind obtinuta din nivelele de interfata (PF, LF), respectiv. *Achizitia de limbaj* implica fixarea multimii Σ, în timp ce *gramatica* limbajului se reduce la specificarea lui Σ. În fine, un *sistem de parsare* care este invariant si neantrenat (cum adesea se presupune) poate fi vazut ca o transformare a perechii (Σ, π) într-o schema structurata similara cu o SD. Conditiile asupra reprezentarilor LN impuse pentru diferite principii si (sub)teorii, cum ar fi teoria *legarii*, teoria *cazurilor*, θ-teoria etc., sunt satisfacute pe nivelele de interfata ale sistemelor de performanta. Toate aceste ipoteze fac parte din teoria MinP a lui Chomsky si din constructia sa pentru UG.

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

Sistemul computational al unui LN concret preia reprezentarile unei forme date si le modifica, în timp ce UG trebuie sa furnizeze mijloacele de a reprezenta o multime de elemente din *lexicon* într-o forma care sa poata fi accesata si procesata de catre sistemul computational. Forma sub care este accesat lexiconul de catre sistemul computational poate fi considerata ca fiind o anumita *versiune* a teoriei X-bar. Schemele X-bar pot fi asociate în mod natural cu *structuri de trasaturi lingvistice* [18], ca un *tip de date lingvistice* standard si invariant pentru a reprezenta si a procesa LN eficient. În strategia SCD, *schemele X-bar*

augmentate [19] considerate pâna acum nu sunt doar tipuri de reprezentare a datelor la nivelul lexiconului ci ele pot asigura structurile invariante fundamentale pentru a reprezenta si a procesa textul în LN la nivel sintactic [1], [2], [6].

În teoria *Minimalist Program* si modelarea P&P a UG, proprietatile si relatiile esentiale sunt formulate în termenii simpli si elementari ai *teoriei X-bar*. Astfel, o *structura* X-bar este compusa din *proiectiile* lingvistice ale *nucleelor* selectate din lexicon. În *schema X-bar* a teoriei MinP reprezentata în Fig. 2.2.1. sunt prezente doua relatii locale: relatia *Specificator-Nucleu* de la ZP la X, si relatia *Nucleu-Complement* de la X si YP (ordinea categoriilor nu este esentiala, fiind stabilita de catre parametri P&P adecvati de ordonare). Relatia Nucleu-Complement (Nucleu-Argument) nu este numai "locala" ci si fundamentala deoarece este asociata *(θ-)relatiilor tematice*.

Daca, pentru moment, nu este luata în considerare *relatia de adjunctie*, sau adjunctii se considera a se afla printre argumentele-complemente, X-bar structurile pot fi reduse la X-bar schema din Fig. 2.2.1, cu urmatoarele specificari: **(a)** Sunt considerate numai relatiile locale (deci nici o relatie de proiectie între X si vre-o sintagma inclusa în proiectiile maximale YP sau ZP); **(b)** Relatia *Nucleu-Complement* reprezinta *relatia locala de nucleu* [Eng: *core relation*]; **(c)** O relatie locala *admisibila* a schemei X-bar din MinP este cea *Nucleu-Nucleu*. De exemplu, relatia unui verb predicativ cu nucleul predi- cational (deverbal) al unei sintagme nominale pe care o subcate-gorizeaza; **(d)** O alta relatie în X-bar schema din MinP este *legatura de lant* [Eng: *chain link*], corespunzând unui *lant de dominare* sau de *guvernare*.

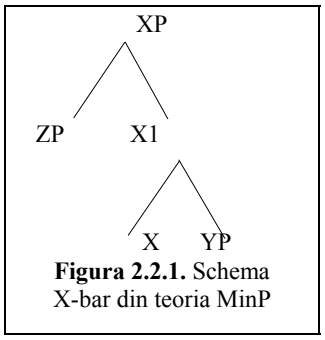


Figura 2.2.1. Schema X-bar din teoria MinP

Guvernarea realizata de nucleu joaca un rol central în toate componentele teoriei MinP asupra UG. Una dintre problemele-cheie este asignarea corecta a trasaturilor nucleului. În HPSG si SCD, de exemplu, acest lucru este realizat la nivel de lexical (BAR = 0), dupa aplicarea flexionarii, cât si la nivel de lexicon (nivel de proiectie notat conventional cu BAR = –1) pentru clasa categoriilor lingvistice cu *proprietati functionale* (*predicationale, relationale*), fie ele verbe, substantive, adjective, marcheri de sintagma, marcheri de discurs etc. care antreneaza un comportament sintactic functional [2], [6]. În particular, pentru teoria MinP, subteorii ca *θ-guvernarea* si *guvernarea de caz*, corespunzând *θ-marcarii* si *Caz-marcarii*, sunt cele mai importante forme de dominare. Un studiu comparativ al guvernarii categoriilor (dependenta, dominare), relatie prezenta 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 catre (sub)teoriile (de asemenea complementare) continute în MinP si P&P, si care expliciteaza fenomenele de *guvernare*, *legare*, *limitare* etc. ce s-au dovedit a fi importante pentru orice teorie

lingvistica deoarece ele asigura reguli pentru organizarea *lexiconului* si a *sistemului computational* care genereaza si recunoaste SDs.

De exemplu, în functionarea *teoriei cazurilor* în contextul schemelor X-bar din MinP, ipoteza standard din MinP este aceea ca, într-o fraza (propozitie), relatia *Specificator-Nucleu* atrage dupa sine *cazul structural* pentru *pozitia de subiect*, în timp ce *pozitia de obiect* primeste cazul sub guvernarea nucleului V, incluzând constructii în care obiectul marcat cazual de catre un verb nu este complementul sau ci doar un adjunct (asa-numita *marcare de caz exceptionala*).

În continuare este prezentata structura X-bar de baza a *clauzei* în teoria MinP, cu urmatoarele notatii uzuale: C = COMP = Complementizator, T = Timpul, Agr_S = acordul subiectului; Agr_O = acordul obiectului etc.

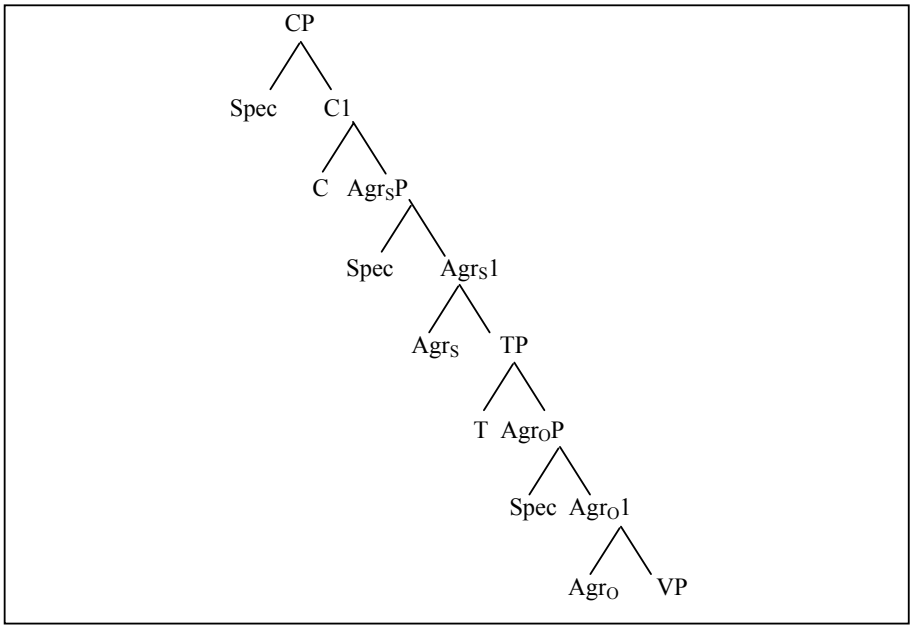


Figura 2.2.2. X-bar structura clauzei în teoria MinP

Schemele X-bar clauzale clasice din Fig. 1.2. si Fig. 1.3. sunt expandate în Fig. 2.2.2., cu urmatoarea posibila interpretare functionala: X-bar schema MinP are ca nucleu VP, care îsi selecteaza sintagma-Obiect (sau argument, mai general) prin acord si marcare, afectata apoi de Specificator. Un timp finit T aplicat sintagmei Verb-Obiect genereaza sintagma TP [Eng: *tensed phrase*], careia i se aplica apoi aceleasi functii de selectie a subiectului (acord, marcare, specificare), generând sintagma Verb-Obiect-Subiect, care este

de fapt clauza finita simpla (notata S). În fine, prin aplicarea asupra lui S (vazuta ca sintagma Agr_SP) a unui complementizator C (sau marcher clauzal, marcher de discurs etc.) se obtine o clauza "completa" ce poate, prin recursie, sa ne orice fraza [Eng: *sentence*].

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

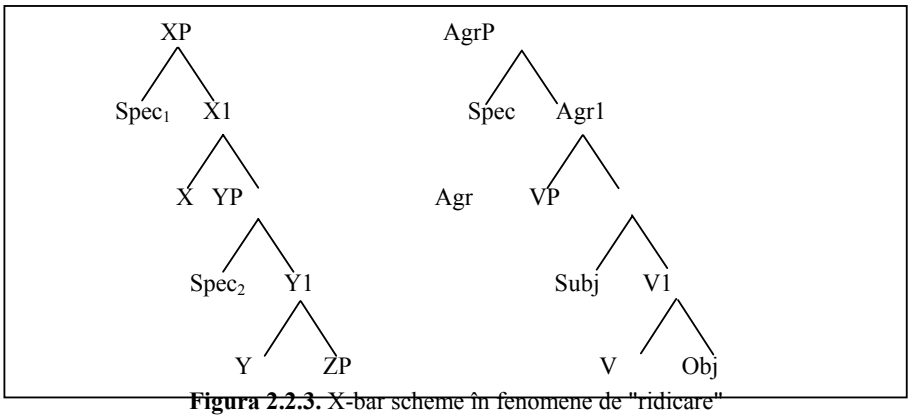


Figura 2.2.3. X-bar scheme în fenomene de "ridicare" la nivel de Spec în MinP

Concluzia este aceea ca teoria X-bar din MinP sintetizeaza relatiile fundamentale de dependenta, descrise de X-bar schemele propuse, si implicate în procesele de organizare a lexiconului si a sistemului computational din UG. X-bar teoria în abordarea MinP reflecta în principal *aspectele statice* întâlnite în fenomenele de *guvernare* (**c**-comanda, **m**-comanda, bariere, categorii de blocare etc.), în *teoria legarii* si în procesele de *referinta-coreferinta*, în stabilirea dependentelor la mare distanta (extra-clauzale) etc. Nu vrem sa intram în detalii (oricum complicate) si sa explicitam mecanismele de lucru ale X-bar schemelor considerate, ci mai curând sa atragem atentia asupra *teoriei X-bar* ca o *componenta fundamentala* a unei teorii lingvistice noi si elaborata cum este MinP si modelul sau P&P [3].

Teoriile MinP si P&P nu reprezinta un punct-terminus pentru evolutia teoriei X-bar. Dimpotriva, asigura o baza de pornire pentru o strategie radical diferita în care Chomsky examineaza cele mai serioase argumente pentru a abandona teoria X-bar [3; Cap. *Categorii si transformari*]! Aceasta alternativa si consecintele sale sunt discutate în sectiunea 5, si ar trebui sa reprezinte una dintre cele mai importante provocari prezente pentru domeniul analizei si proiectarii teoriilor lingvistice [21].

Unul dintre principalele scopuri ale sectiunii care urmeaza este de a introduce propunerea noastra de *scheme X-bar functionale* (*scheme FX-bar*) în cadrul strategiei lingvistice SCD. Propunerea noastra o consideram a fi o pozitie pragmatica si echilibrata în *directia* teoriei X-bar, atragând atentia asupra adevaratului sau rol si oportunitatilor computationale din lingvistica reala. Întelegerea corecta a aspectelor *statice* si *dinamice* ale acestei versiuni a teoriei X-bar ar trebui sa fie de asemenea o consecinta a unei privire cuprinzatoare a întregului context al teoriilor lingvistice care stabilesc principiile de dependenta, clasele de marcheri, categoriile si ierarhiile, regulile de referire si structurare, în strânsa relatie cu formele si regulile de constructie ale schemelor FX-bar.

3. Scheme X-bar functionale si strategia lingvistica SCD

În [19], în contextul *strategiei lingvistice SCD* (*Segmentare-Coeziune-Dependenta*) [22], [19], [1], [2], [6], este definita o clasa de *scheme X-bar augmentate (scheme AX-bar)*, scheme destinate a reprezenta *invarianti sintactici generali* de reprezentare si operare cu structurile gramaticale ale LN, în particular pentru limba româna, ca solutie la problemele de analiza si generare automata a LN. *Schemele FX-bar (functionale)* propuse aici completeaza si extind *schemele AX-bar* [19], si 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 informatiei lingvistice în structuri de trasaturi lingvistice, standardizate si tipizate; (2) din punct de vedere *dinamic*, schemele FX-bar pot codifica informatia lingvistica în forma procedurala ca *functii si relatii standard* ce sunt (recursiv) apelate în cadrul proceselor de analiza si generare a LN; (3) schema FX-bar generala poate fi de asemenea interpretata si utilizata ca un automat pe baza caruia sa se realizeze o analiza *on-line* a textului unei fraze, cuvânt cu cuvânt.

3.1. Câteva preliminarii asupra SCD

Sunt necesare unele precizari asupra notiunilor si notatiilor cu care lucreaza *strategia lingvistica* SCD. Unul dintre elementele importante este ca nivelul 2 (BAR = 1) în X-bar schema clasica joaca un *rol-cheie* în SCD pentru constructia structurilor sintactice, si 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 *proiectiei lexicale* X1, cu X = N, V, A, si *clauzei minimale* CL0, în X-bar schema fundamentala propusa în Fig. 5.7.2.1.

Sa mentionam ca orice XG (X1) este un XP (X2), dar nu si invers, deoarece nivelul proiectiei categoriei X în cazul XG lucreaza numai pentru nivelul BAR ≤ 1. SCD face de asemenea distinctie între câteva tipuri de NGs (NGs elementare, predicationale, non-predicationale, etc.), VGs (VGs la un timp finit si la un timp non-finit) etc.

O alta trasatura esentiala si specifica a SCD este un tratament adecvat al *proprietatilor functionale* ale categoriilor lingvistice, ca si al tuturor *categoriilor relationale* si sintagmelor (expresiilor) de *discurs*. Mecanismul utilizat pentru a obtine acest lucru se bazeaza pe *clase de marcheri lingvistici* si *ierarhiile* lor [1], [2], [6]. Câteva observatii se impun:

(a) *Marcherii* din SCD, numiti *marcheri de structuri sintagmatice* (PS-Ms) [Eng: *phrase-structure markers*], sunt cu totul diferiti de ceea ce teoria lui Chomsky numeste formal "*marcheri de sintagma*" [Eng: *phrase-markers*] în [17], sau *T(ree)-marcheri* în [3]. *Marcherii Chomsky* sunt definiti ca "*taieturi orizontale*" (sau "factorizari") în cadrul unui arbore de derivare, sau ca fiind arborele însusi. Mult mai apropiati de ceea ce sunt PS-Ms în HPSG [16], *marcherii de structuri sintagmatice* (PS-Ms) din SCD sunt acele categorii lexicale si nelexicale care se aplica cuvintelor si structurilor sintagmatice (PSs) cu scopul de *evidentia*, de a *marca*, anumite functii si relatii sintactice si semantice pe care PSs

respective le joaca în cadrul unei exprimari. Punerea în evidenta a anumitor functii care se aplica PSs se refera la (cel putin) câteva elemente: *tipul functiei* (sintactic, semantic, relational, logic, pragmatic, discursiv etc.), *locul*, în text, *unde începe aplicarea* functiei sau relatiei, si *domeniul (domeniile, conexe sau nu)* de aplicare a functiei sau relatiei (limitele textuale între care se aplica).

Exemple tipice de PS-Ms din SCD sunt: (*a*) trasaturile *predicative* generate de catre *categoriile predicationale* (de fapt, verbe, substantive, adjective si adverbe predicationale); (*b*) acele *mijloace gramaticale* prin care sunt introduse *noi* NGs (grupuri nominale în limbajul SCD), VGs, AGs (*Caz*-marcarea, acordul, gradele de comparatie, etc.); (*c*) acele *categorii* si *expresii* (numite si *marcheri de discurs*) care introduc *noi* clauze; (*d*) PS-Ms care introduc *proprietati relationale* asupra PSs si clauzale (de exemplu de *marcheri* de tip logic cum sunt structurile *daca-atunci-altfel*, *deoarece,* etc., dar si marcheri de tip sintactico-semantic cum sunt aceia care introduc *categorii* si *clauze subordonate* etc.)

(b) SCD se aseamana din unele puncte de vedere cu abordarea [16] a HPSG si, partial, cu [15], care exploateaza, pentru prima oara în clasa teoriilor lingvistice bazate pe *gramatici* de PSs (PS-Gs), într-o mult mai mare masura, categoria lingvistica a marcherilor PS-Ms. În [16], Pollard & Sag "postuleaza o noua parte a *marcherilor* de discurs,… ce se remarca … printr-un nou atribut al categoriilor (în plus fata de NUCLEU si SUBCAT) numita MARKING, cu valori din sortul *marking*". Teoria HPSG enunta PRINCIPIUL MARCARII [16, p. 400] dupa cum urmeaza:

"Într-o sintagma cu nucleu, valoarea trasaturii MARKING este lexical-identica cu cea a trasaturii MARKER-DAUGHTER daca aceasta exista, si cu cea a trasaturii HEAD-DAUGHTER în caz contrar.

Modul în care HPSG [16] pune la lucru PS-Ms reprezinta un bun si esential pas înainte, desi credem ca nu exploateaza îndeajuns potentialul functional si relational al diferitelor clase de marcheri si ierarhiile acestora (asa cum face strategia SCD, vezi si [7]).

(c) Continuând si extinzând constructia limbajului, ca o expresie de convergenta între gramatica categoriala si *Minimalist Program*, Chomsky [3] considera *transformarile generalizate* (GTs) si concepe un demers de înlocuire a X-bar teoriei, ce explica în Programul Minimalist structura constituentilor (sintagmatici) complecsi, prin GT *Merge* care construieste obiecte sintactice pornind de la obiecte sintactice simple (de exemplu, "*speaks*" si "*French*" sunt "reunite" într-un nou obiect sintactic "*speaks French*" etc.). Mai multe formalizari ale acestui nou curent al ideilor lui Chomsky pot fi gasite în cadrul gramaticilor logice multi-modale si de tipuri categoriale, *e.g.* [21], [23], [24] etc. (vezi si sectiunea 5).

(d) Dintr-o perspectiva diferita dar oarecum similara, *gramatica functionala* (FG) [25] a lui Simon Dik, orientata functional si semantic, încearca sa faca aceleasi lucruri. Ca si în SCD, FG gaseste *patru tipuri ierarhice* de baza ale categoriilor relationale, aceste tipuri corespunzând într-o buna masura cu *clasele de marcheri* PS-Ms si *ierarhiile* lor stabilite în

SCD [7], [2], [6]. PS-Ms reprezinta acele mijloace lingvistice de "suprafata" pe care le utilizeaza un limbaj natural pentru a organiza sintactic si semantic structurile codificate în constructii gramaticale. Se impune în viitor o analiza comparativa între cele *patru nivele* sau "*straturi*" din organizarea formala si semantica furnizata de FG [25], si cele *patru nivele* de proiectie lingvistica, împreuna cu clasele de marcheri corespunzatoare, din SCD: (1) *cuvântul* (lexical); (2) *sintagma* XG (X = N, V, A) subclauzala; (3) *clauza* (finita si infinita); (4) *discursul* (una sau mai multe fraze, care sa formeze un *segment* de discurs).

(e) În fine, privitor la utilizarea intensiva a *caracterului predicational* pe care categoriile lexicale majore (N, V, A) îl poarta (proprietate mostenita sau dobândita apoi de alte categorii gramaticale), *strategia lingvistica* SCD este comparabila în special cu FG, cu accentul particular pe ierarhiile de delimitare si marcare aplicate structurilor sintactico-semantice. SCD porneste de la *lexicon* si stabileste la acest nivel o *taxonomie predicationala initiala* pentru categoriile lexicale majore. Un exemplu simplu al acestei taxonomii predicative este dat de catre cele doua categorii importante de substantive comune: *substantive existentiale* sau *obiectuale*, a caror *predicationalitate* (*trasatura* PRED) este EXIST (*e.g.* [Eng: *student, table*; Rom: *elev-student, masa*]) si a caror reprezentare functionala reflecta categorii individuale sau personale, de exemplu predicatul uni-variabil *student*(X), *masa*(X) etc., si substantive de tip-predicational, a caror predicationalitate (*trasatura* PRED) are valoarea ACT, *e.g.* [Rom: *întâlnire, invidie, marcare* etc.], si ale caror reprezentari functionale depind de mai multe variabile, de exemplu *întâlnire*(X, Y,…), *invidie*(X, Y,…), *marcare*(X, Y) etc. Substantivele proprii si/sau personificarile sunt codificate prin constante ale variabilelor din predicatele de mai sus. Câteva din remarcile anterioare vor fi aprofundate în concluziile finale ale lucrarii.

Schemele FX-bar, ca si precursoarele lor *schemele AX-bar* [19], reflecta pentru SCD faptul ca un XPG (grupul sintagmatic de nucleu X), sau mai simplu XG, contine un *nucleu*, reprezentat printr-o categorie lexicala (nevida) sau printr-o categorie virtuala (vida), înconjurat (prin relatii de *coeziune*) de specificatori si/sau modificatori de tipul A (adjectival-adverbial). Este esential sa facem urmatoarea specificare: un XG din SCD nu include nici un complement (argument obligatoriu) sau adjunct. Complementele si adjunctii, împreuna cu *nucleele* de nivel BAR = 1 formeaza nivelul BAR = 2 în FX-bar schema propusa în Fig. 3.2.1. Pentru un anumit nivel de specificare semantica, FX-bar schemele nu fac o distinctie clara între complemente (argumente obligatorii) si adjuncti, considerând toate structurile subcategorizate ca fiind argumente sintactice; clasificari ulterioare (suplimentare) sunt facute pe baza sabloanelor verbale si restrictiilor sintactice, semantice, si pragmatice asupra componentelor sablonului, la nivel de lexicon.

O problema a carei solutie poate influenta în mod special si teoria X-bar este aceea a asignarii corecte a complementelor si adjunctilor, în particular, a stabilirii corecte a dependentelor dintre grupurile nominale (NGs). Solutia acestei probleme nu se poate obtine la nivel sintactic, iar o solutie completa nu se poate obtine uneori nici chiar în contextul unui nivel semantic minimal (vezi [26], [27]). Chomsky remarca realitatea ca "**…** *the distinction between modifiers and arguments is notoriously difficult in certain cases*" [9, p. 44]. Exemple simple ilustreaza aceasta problema: în TBarr [8], sintagmele "*the students of physics* " este

vazuta ca un NP cu un *argument* PP, în timp ce sintagma "*the students in the yard* " este considerata a fi un NP cu un *adjunct modificator* PP. De fapt, în numeroase LNs, inclusiv engleza, se pot aduce multiple argumente serioase pentru ca cele doua sintagme sa poata fi la fel de bine interpretate fie într-un fel, fie în celalalt.

Solutia SCD pentru acest exemplu foarte particular este urmatoarea (schitând si solutia problemei generale): substantivul "*students*" este obiectual, adica *nu are* o natura *predicationala* prin el însusi, astfel ca ambele sintagme nominale care îl succed sunt considerate de catre SCD ca fiind *modificatori* pentru NG "*students*". Natura acestor modificatori poate fi diferita deoarece "*physics*" este introdus de marcherul de *caz* (genitiv) "*of*", în timp ce "*the yard*" este introdus de marcherul *prepozitional* "*in*". În general, când nucleul lui NG poseda o trasatura *predicationala*, atunci NG care urmeaza nucleului predicational asigura o distributie sintactica ce satisface un anumit sablon (verbal) al predicatului (verbului) corespunzator.

Clasele din PS-Ms si ierarhiile lor din SCD [7] sunt responsabile pentru delimitarea structurilor sintagmatice propuse de schemele FX-bar, si pentru stabilirea dependentelor sintactico-semantice. Diferitele tipuri de marcheri sunt adesea aplicate simultan (deci multiplu) asupra acelorasi categorii gramaticale, în cadrul anumitor nivele de structurare (proiectii pe BAR-nivel). Similar cu unele teorii lingvistice (LFG, FG, si partial HPSG) dar contrar altora (GB, GPSG etc.), SCD nu considera *prepozitia* (X = P) ca fiind o categorie lexicala majora. În SCD, P primeste rolul unui marcher (functional), având atât proprietati de marcher de caz cât si de complementizator. Categoriile HPSG PP[+PRD] sau PP[–PRD] (vezi [16]) sunt irelevante pentru SCD deoarece trasatura +PRD în HPSG este atribuita numai lui PP subcategorizat de un V, în timp ce trasatura (*predicationala*) PRED din SCD poate fi în mod egal atribuita lui V, N, sau A (la nivelul lexiconului, cel putin) dar nu si lui P.

În S-C-D proprietatile de subcategorizare sunt exploatate *ab initio*, la nivelul de organizare al lexiconului, pe baza *trasaturii functionale* PRED de *predicationalitate*, asignata sau nu, unora din categoriile sintactice majore N, V, A. Observatii lingvistice empirice ne-au convins, înca de la începuturile cristalizarii SCD [22], ca o *taxonomie* functionala si predicativa adecvata ar trebui sa reprezinte punctul de plecare al oricarei teorii lingvistice, atât din motive teoretice cât si pragmatice, si ca multiple abordarile actuale (cum ar fi [27]-[32]) aduc o sustinere puternica pentru multe din ideile esentiale din SCD, în special folosirea intensiva a predicativitatii si functionalitatii descrierilor lexical-semantice ale categoriilor lingvistice atât în *procesarea automata* a LN cât si în cadrul *bazelor de cunostinte lexicale*.

[19] propune urmatoarea specificare a *Principiului Proiectiei Maximale* (PMP) [Eng: *Principle of Maximal Projection*], ca un pas important catre folosirea intensiva a trasaturilor predicationale (functionale) ale categoriilor lexicale majore în SCD. Propunem aici

O specificare a PMP (forma actualizata):
Proprietatile de subcategorizare ale categoriilor sintactice majore N, V, A depind de trasatura lor lexical-semantica PRED(icativity), **cu valorile ACT si EXIST, si de trasatura lor morfo-semantica TENS**(e), **cu valorile FINI**(te) **si INFI**(nite).

Trasatura PRED, atribuita categoriilor majore N, V, A la nivel de lexicon, primeste doua valori: valoarea ACT, pentru acele categorii care au *proprietati predicationale* (în literatura este folosit adesea termenul "*deverbale*"), si valoarea EXIST, pentru acele categorii N, V, A cu caracter *existential*, obiectual, non-predicational. Trasatura TENS primeste valorile FINI(te) pentru acele forme ale categoriei V care poseda un timp sau aspect finit, personal, si valoarea INFI(nite) pentru toate celelalte categorii. Exemple:

[Eng: *boy, pencil*; Rom: *baiat, pix*] PRED:= EXIST; si TENS:= INFI;

[Eng: *attempt, showing, proved*; Rom: *încercare, aratând, demonstrat*]
PRED:= ACT; si TENS:= INFI;

[Eng: *are*; Rom: *sunt*] PRED:= EXIST; si TENS:= FINI;

[Eng: *gives*; Rom: *da*] PRED:= ACT; si TENS:= FINI.

Într-un grup verbal VG reprezentând un compus la un timp finit, valorile "pozitive" de trasaturi, cum sunt ACT sau FINI sunt *mostenite* de la nucleul V al VG de catre întreaga sintagma VG, sau pot fi obtinute *cumulativ* prin *proiectia* morfo-sintactica.

Specificarea PMP de mai a functiei *proiectiei maximale* este necesara în SCD deoarece în multe LNs, inclusiv în româna, *calitatea deverbala* (predicationala, deci functionala) a categoriilor lexicale traditional *non-verbale* cum ar fi N si A trebuie descoperita cât mai devreme posibil si asignata la nivel de lexicon. De exemplu, în engleza, desi pentru substantivele care '*verbalizeaza*' în "*-ing*" valoarea trasaturii lor TENS este INFI, aceste substantive poseda, pentru trasatura PRED, aceeasi valoare ACT sau EXIST pe care o au verbele din care provin substantivele (sau gerunziile) în "*-ing*", si astfel poseda *aceleasi* proprietati de subcategorizare ca ale verbului de origine.

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

Continuând ideile de baza ale schemelor AX-bar din [19], propunem, pentru SCD, FX-bar schema generala din Fig. 3.2.1. Muchiile din stânga contin *noduri* cu rol *functional* sau *relational*: marcheri, cuantificatori, specificatori, modificatori (eventual adjuncti). Pentru a obtine reprezentari sintactice si semantice corecte, nodurile functionale se aplica (recursiv) *nucleelor* Xk si CLk, k = 0, 1, 2, iar nucleele, cu rol functional (predicational, X1) sau relational (eventual X2), au ca argumente clauze infinite (complemente, X1) sau finite (X2). Precizam ca la acest nivel nu se poate face distinctia dintre complemente $COMPL_i$ (argumente obligatorii) si adjuncti $ADJCT_i$ (argumente optionale). În mod normal, în Fig. 3.2.1., $ADJCT_i$ sunt "amestecati" printre ARG_j , la nivel sintactic nefiind discernabili de complementele obligatorii ale unui nucleu predicational. Pozitia functionala (la stânga nodului X1) a nodurilor ADJCT poate rezulta doar în urma unor calcule semantice si pragmatice suplimentare, din care se obtine rolul tematic al argumentelor ARGs ale lui X1.

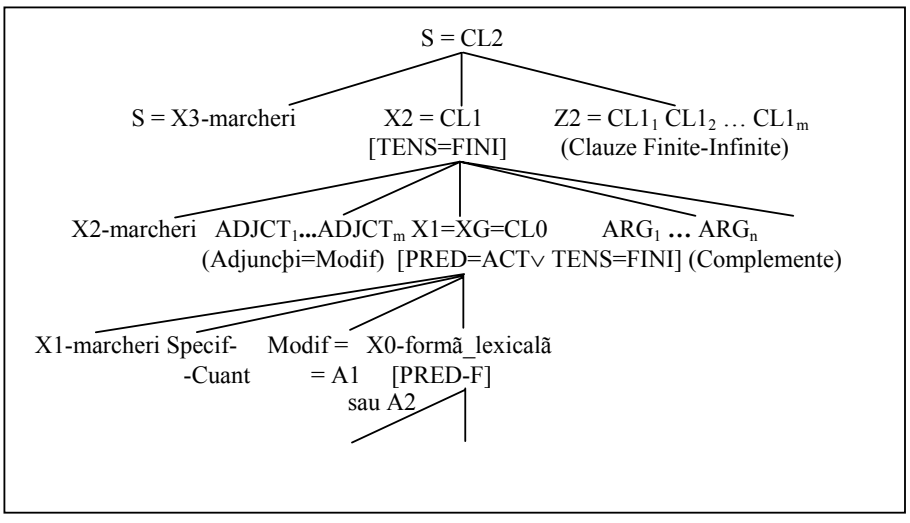


Figura 3.2.1. Schema (funcțională) FX-bar generală

(♣) ***Aspecte specifice ale schemei FX-bar propuse***: (♣1) Sunt permise un numar arbitrar de *argumente* (sau *sateliti* în sensul [10], [31]), toate notate cu ARGs. În SCD, ARGs sunt formate din *complemente* obligatorii (COMPLs) si din *adjuncti* (ADJCTs), sau *complemente optionale*. ADJCTs pot fi reprezentati la nivel sintactic tot ca argumente ale nucleului, însa la nivel semantic ADJCTs au rol de *modificatori* ai nucleului. Notatia "A-pozitie" din teoriile Chomskyene, care înseamna ARG-pozitie, nu trebuie confundata cu notatia noastra pentru categoria A = adjectiv-adverb. În teoriile si notatia lui Chomsky, COMPLs sunt în A-pozitie (ARG-pozitie), în timp ce ADJCTs nu. SCD se situeaza pe o pozitie sintactica similara cu HPSG [16], care utilizeaza lista SUBCAT pentru a codifica toate sintagmele pe care le subcategorizeaza un nucleu semantic, adica atât COMPLs cât si ADJCTs (sau ARGs din SCD). (♣2) Sintagmele AG = A0 sau A1, sau AP = A2 sunt *postulate* de catre SCD ca fiind de tipul categoriei *functionale* Modif, manifestate prin categoriile A (de nivel X0, si aplicabile la nivel X0), ADJCTs (de nivel X1, si aplicabile la nivel X0 si X1), si clauza relativa (de nivel X2, si aplicabila la nivel X0 si X1). (♣3) Categoria generica Specif (sau Spec), în care intra cuvintele si sintagmele ce desemneaza *cuantificatori* de toate tipurile (generalizati), determinatori (în particular), este postulata de catre SCD ca fiind o *categorie functionala* ce poarta trasaturi de natura *cuantificationala* la nivel lexical (în particular, *negatia* la nivelul X1), inclusiv articularea (hotarâta sau nu), suprapunându-se deci uneori peste X1-marcheri de trasaturi functionale cum este *acordul*. Relatiile (functionale) de *acord* sunt esentiale pentru *coeziunea locala* si *globala* în cadrul strategiei SCD: acordul dintre X0-Modif si X0-Specif cu nucleul X0 (la nivel X1), acordul Nucleu-Subj (sau chiar Nucleu-COMPL) si acordul COMPL-$\text{Pron}_{\text{Emfat}}$ (Pronume emfatic) (la nivel X2), o anumita *corespondenta* a timpurilor evenimentelor într-o clauza si între clauze. Aceste tipuri de relatii de acord, referinta si coreferinta, coeziune, coerenta, etc. sunt responsabile pentru o larga clasa de dependente locale si globale, inclusiv dependente la distanta mare si în extra-pozitie. Accentul în componenta de *coeziune* a strategiei SCD (Segmentare-Coeziune-Dependenta) cade pe mijloacele *sintactice* si de "*suprafata*", mai

curând decât pe cele semantice, încercând sa gasim, sa extragem, si sa utilizam într-o masura maximala informatii de ordin superior, cum ar fi informatia de discurs [34], pragmatica, semantica etc. (♣4) Sintagma traditionala PP din teoriile lingvistice clasice, iar în SCD, grupul prepozitional PG (format dintr-un grup nominal NG care este precedat de o prepozitie sau o locutiune prepozitionala) este întotdeauna considerata un ARG (COMPL sau ADJCT) în FX-bar schemele al caror nucleu (lexical nevid sau vid) este N, V, A. Aceasta ipoteza de baza asupra PG este justificata de SCD prin faptul ca P este considerata o categorie majora, adica o categorie de nivel X1 în schema FX-bar din Fig. 3.2.1. ci doar o categorie de nivel X0. Proprietatile de subcategorizare ale N, V, A (dar nu si P) pot fi asignate *ab initio*, la nivel de lexicon, începând cu trasatura lexicala PRED a categoriilor predicationale. Categoria P poate primi proprietati functionale, cel mai adesea ca *marcher de caz*, uneori proprietati *relationale* (de exemplu [Eng: *on*; Rom: *asupra*]), dar nu si proprietati de subcategorizare. (♣5) Subiectul (Subj) în SCD, lexical nevid sau vid (PRO), este considerat ca un argument special al proiectiilor maximale ale categoriilor X = N, V, A într-o clauza finita (de nivel X2) sau infinita (de nivel X1). (♣6) În ipotezele (♣5) si (♣2) de mai sus, categoria lingvistica traditionala VP este dizolvata într-un grup verbal VG (finit sau infinit), înconjurat (de cele mai multe ori urmat) ca nucleu de ARGs si formând o clauza finita, respectiv infinita. (♣7) *Teoria limitarii* si multe probleme majore legate de TBarr [8], [9], [17] sunt explicitate si rezolvate în cadrul realizat de SCD si schemele FX-bar, în principal datorita delimitarii clare a functiilor si relatiilor care se aplica cuvintelor si sintagmelor, a reprezentarii lor lexicale prin clasele de PS-Ms, si a specificarii domeniului lor de aplicare. Acest rol este realizat explicit în cadrul claselor si ierarhiilor de marcheri propuse si utilizate de SCD [2], [6], [7]. Trebuie sa remarcam ca lucrarile sale cele mai recente [34], [35], Chomsky adopta o tehnica similara de "limitare" a operatiilor de *construire* [Eng: *merge*] si *transformare* [Eng: *move*] doar la "domeniul" sintactic al unei "*faze*" [Eng: *phase*], o unitate textuala (care în general coincide cu clauza!) în care Chomsky propune urmatorul *principiu de impenetrabilitate* "*Într-o faza* (clauza *n.n.*) F *cu nucleul* H, *domeniul lui* H *nu este accesibil la operatii în exteriorul lui* F, *ci este accesibil numai* H si *muchia sa* (nodul sau ascendent)" [34]. Exact asa este construita si functioneaza schema FX-bar ! De asemenea, fenomene de *teoria legarii* [9], [8], [3], [16], *legaturile* [Eng: *linking*] din [27], mecanisme de *coeziune* (locala si globala) si *discurs* întâlnite în [36], [31], [33], etc. sunt mai usor de pus în evidenta si de rezolvat în cadrul oferit de strategia lingvistica SCD si teoria FX-bar.

(♦) ***Observatii asupra ipotezelor de lucru pentru schema FX-bar din Fig. 3.2.1.***: (♦1) Schema FX-bar este proiectata sa lucreze în asociere cu un parser care este capabil sa recunoasca clasele de PS-Ms si structurile sintagmatice considerate de strategia lingvistica SCD. Schema FX-bar este organizata pe *patru nivele* de proiectie BAR = 0÷3 (deasupra nivelului de lexicon, notat conventional BAR = –1); *trei nivele* **X0-X1-X2** corespund proiectiei dintre nivelul lexical (BAR = 0) si nivelul *clauzal*, al structurilor *uni-eveniment*; alte *trei nivele* **CL0**(=X1)**-CL1**(=X2)**-CL2** corespund proiectiei dintre nivelul *clauzal minimal* CL0 = X1 si nivelul *frazei*, al structurilor *multi-eveniment*. Nivelele uni-eveniment X0-X1-X2 exprima predicatia clauzei (propozitiei) simple în care sunt distribuite categoriile lexicale de baza si sintagmele pe care le genereaza, în timp ce nivelele CL0-CL1-CL2 exprima relatiile logice si predicationale (de ordinul doi) dintre clauzele simple. Schema FX-bar lucreaza într-o maniera recursiva (top-down sau bottom-up), atât în situatiile de analiza cât si în cele de

generare în care este antrenat parserul asociat, în strânsa cooperare cu strategia lingvistica SCD, cu clasele de PS-Ms si ierarhiile lor si, mai ales, pe baza *meta-algoritmilor* SCD de analiza-generare [1], [2], [6], [7]. Sa mai observam ca FX-bar schema din Fig. 3.2.1. poate fi utilizata independent de asa numita *ordine canonica* (sau *sistemica*) a cuvintelor si sintagmelor dintr-o clauza, specifica fiecarui LN [37], [38]. (♦2) Valoarea ACT de trasatura (functionala) pentru categoriile N si A (si implicit V) este atribuita acestor categorii la nivel de lexicon atunci când ele corespund unor evenimente cu actanti si/sau stari multiple. Valoarea EXIST este implicit sau explicit introdusa de formele si întelesurile verbelor existentiale (*a fi*), modale (*a trebui*), etc. (♦3) Trasatura (functionala) TENS este similara cu categoriile virtuale I (INFL) si T (Tense) din teoriile GB si TBarr ale lui Chomsky si din schemele S-bar corespunzatoare (Fig. 1.3. si Fig. 2.2.3.). Pentru un VG finit (TENS = FINI), structura V2 corespunzatoare devine clauza finita clasica. Daca sintagma XG (X1) este un grup a carei categorie-nucleu X poseda valorile de trasaturi PRED = ACT si TENS = INFI, atunci XG devine noul nucleu al unei clauze infinite ce face parte dintr-o structura de nivel X2 (XP). (♦4) Pozitia speciala a *subiectului sintactic* (Subj) este considerata de catre SCD atât o ARG-pozitie (asemanatoare, de fapt, cu o COMPL-pozitie) cât si o *Caz*-pozitie. În concordanta cu TBarr [8] si cu HPSG [16], Subj primeste pozitia speciala a *primului element* din lista SUBCAT [16]. Aceasta este în esenta o pozitie sintactica, iar Subj poate primi o functie tematica (θ-pozitie) autentica doar ca rezultat al unor calcule sintactice si semantice suplimentare. (♦5) Asa cum rezulta din schema FX-bar din Fig. 3.2.1., sintagmele AP si PP din teoriile lingvistice clasice sunt segmentate de catre marcherii SCD [7] în sintagme mai mici XG, X = N, V, A. Asa cum am precizat deja, SCD atribuie noilor sintagme urmatoarele roluri: AG = Modif, cu rol functional la nivelul de proiectie X1, si 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 (nesaturata) A(X), unde X = N, V, A este nucleul (existent, viitor, sau lipsind pur si simplu) al sintagmei de nivel X1 în care Modif = A. În mod similar, orice categorie Specif (determinator, cuantificator, etc.) joaca un rol similar, schema FX-bar impunând reprezentarea functionala Specif(X), unde X este nucleul sintagmei. (♦6) În ciuda anumitor asemanari (inerente) între schemele FX-bar si versiunea MinP a teoriei X-bar, exista diferente de baza în ce priveste organizarea si functionarea constructiva dintre schemele (F)X-bar din Fig. 3.2.1. si Fig. 2.2.1. De exemplu, în schema FX-bar, fiecare element lexical se proiecteaza într-o categorie obiectuala sau functionala (relationala), aceasta este (coeziv si recursiv) înconjurata de catre Specif si/sau Modif, iar daca valoarea ACT a trasaturii PRED a nucleului este prezenta, atunci aceasta valoare ACT este mostenita de catre întreaga sintagma al carei nucleu a fost specificat sau modificat. Aceasta sintagma cu nucleu predicational îsi subcategorizeaza complementele (argumentele obligatorii COMPLs) si adjunctii ADJCTs (care modifica sintagma-nucleu). În schema X-bar din Fig. 2.2.1., se întâmpla 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 (apart from adjunction, …*)" [3: p. 53]. (♦7) Desi schema FX-bar generala a fost proiectata având în vedere în primul rând limba româna, ea poate fi aplicata pentru a reprezenta, grafic si logic, structuri sintactico-semantice ale LNs cu valori ale parametrilor gramaticali foarte diferite, cum ar fi engleza-germana sau franceza-germana. Distributia complementelor (argumentelor) în româna (engleza, franceza) poate fi foarte

diferita de cea din germana; de exemplu, într-o clauza al carei verb principal din compusul sau verbal VG se afla în pozitie finala, sau pentru o categorie A (adjectiv-adverb) având valoarea de trasatura 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 Fußball / /zerbrochene / /Glass/*

Dupa cum am remarcat în (♦1), schema FX-bar poate fi utilizata independent de regulile structurilor sintagmatice si ordinea lor (din româna sau germana), aceasta deoarece principiile ramân aceleasi si difera numai anumiti parametri si valorile lor pentru LNs distincte: în româna (si engleza, franceza) argumentele succed o categorie A ce reprezinta un nucleu predicational, în timp de în germana ele îl pot (!) precede. Daca un nucleu V al unei clauze are valorile de trasaturi PRED = ACT si TENS = FINI, atunci distributia ARGs este similara cu cea din româna, cu posibile (si probabile) diferente impuse de *ordinea sistemica*, strict dependenta de LN, a ARGs (a se vedea [37] dar si [27]).

Daca se încearca utilizarea formei FX-bar ca "schelet" pentru un automat (sau gramatica formala) de analiza si generare a LN, un asemenea automat ar trebui sa mimeze atât forma generala a schemei FX-bar cât si regulile gramaticale de analiza-generare. Partea din automat care reflecta cele *patru* nivele de organizare a structurilor LN în schema FX-bar ar trebui sa fie independenta de limbaj (cel putin pentru o larga clasa de limbaje europene), în timp ce (sub)partea constituenta care recunoaste structurile lingvistice pe fiecare nivel individual Xk (k = 1, 2, 3) trebuie sa fie dependenta de limbaj (acest fapt este binecunoscut si parametrizat). Reprezentarea schemei FX-bar pentru Ex.3.2.2.G. este aceeasi cu reprezentarile FX-bar pentru Ex.3.2.2.R.-E., si similara cu figura pentru Ex.4.2.R.-E.

4. Exemple de aplicare a schemelor FX-bar

Vom expune câteva exemple de aplicare a schemelor FX-bar la reprezentarea sintagmelor, clauzelor si frazelor. În exemplele prezentate, categoriile gramaticale pentru care PRED = ACT sau TENS = FINI vor fi *subliniate*, iar PS-Ms care se aplica sintagmelor Xk (k = 0, 1, 2) sunt reprezentati grafic în text prin aparitia unuia sau mai multe semne 'slash' /. Sa notam ca schemele (augmentate) AX-bar din [19], desi oarecum asemanatoare în spirit sunt efectiv scufundate în schema FX-bar generala, diferentele substantiale constând în forma unitara a FX-bar schemei si în criteriile sintactice si logico-semantice mai clare, pe baza carora clasele de PS-Ms si ierarhiile lor sunt explicit propuse si aplicate în functionarea schemei FX-bar.

Care este relatia dintre exemplele de FX-bar scheme si formulele logice atasate dupa reprezentarea grafica? *Prima formula* este o reprezentare uzuala a LN, care foloseste limbajul logicii predicatelor, reprezentare mai apropiata de exprimarea în LN, continând toate variabilele ce exprima referintele-coreferintele, dar (pentru simplitate) fara cuantifcatorii corespunzatori. *A doua formula* este traducerea mai completa a primei formule în limbajul de programare logica Prolog, folosind tehnici clasice de reprezentare a

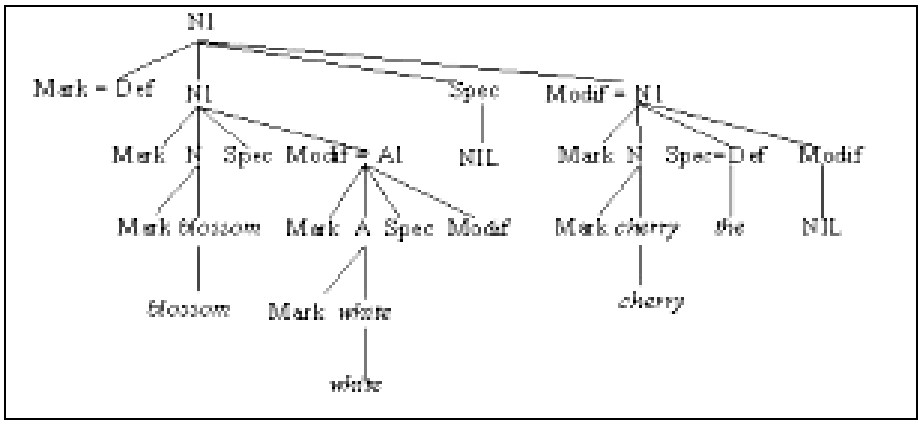
cunostintelor de LN în Prolog. Pe o scala ascendenta a masurii în care sintagmele LN ar fi analizate, *schema* FX-bar poate fi vazuta ca un prim rezultat al procesului de parsare (analiza), *prima formula* ar urma procesului de parsare, incorporând fenomenele de referinta (si coreferinta, rezolutie a anaforei, etc.), iar *a doua formula* ar reprezenta o rafinare a primei formule. Formulele de tipul doi reprezinta de asemenea atât un stadiu final al procesului de analiza a frazei cât si punctul de pornire în procesul de generare a unei fraze (conform cu abordarea [39], [6] a generarii automate a LN, însa diferita de [31], de exemplu).

Este important sa remarcam ca schema FX-bar propusa reflecta, în principal, relatiile de *dependenta* dintre diferitele categorii, sintagme, si clauze dintr-o fraza, împreuna cu marcherii corespunzatori care controleaza, în parte, si comportamentul lor distributional. Deoarece am vazut în ce masura schemele ordinea argumentelor este (parametric) dependenta de limbaj în schemele FX-bar, acestea pot codifica nu numai situatii în care argumentele succed (situatia obisnuita) sau în care ele preced nucleul lor semantic (Ex.3.2.2.), dar si în care argumentele aceluiasi nucleu sunt interschimbabile. Deci aceleiasi schema FX-bar i se pot atribui mai multe formule logice corespunzatoare "echivalente".

4.1. De la text la scheme FX-bar

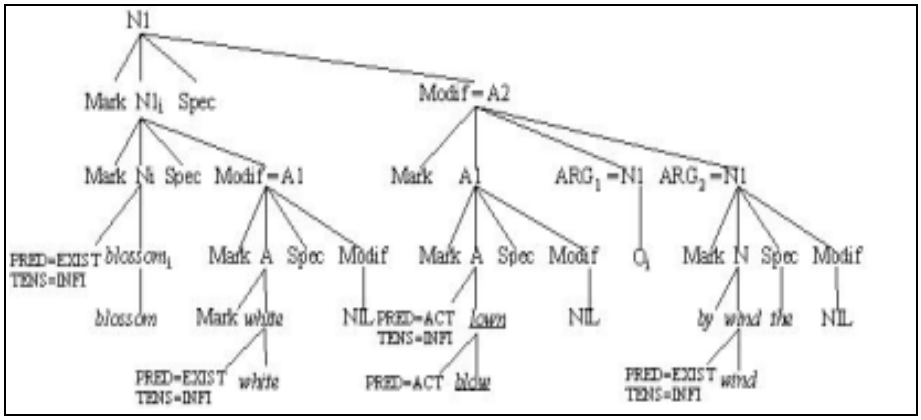
Strategia SCD propune urmatoarele scheme FX-bar pentru exemplele de mai jos. Desi muchiile ale caror noduri sunt Modif sau Specif sunt situate în dreapta nucleului corespunzator (pentru conveniente grafice), ele trebuie întelese ca având rol functional (situate la stânga si aplicându-se nucleului X1). La fel si cu unii adjuncti, la nivel X2. Diferentele dintre codificarea formei pentru engleza si cea pentru româna sunt nesemnificative (cu exceptia unor aspecte suplimentare de acord, care sunt puse în evidenta). Forma codificata a textului pentru limba engleza este un argument suplimentar pentru versatilitatea schemelor FX-bar propuse.

Ex. 4.1.1.R. */ floare albă / de cireş /*
Ex. 4.1.1.E. */ the cherry / white blossom /*


N1
Mark = Def
N1
Spec
Modif = N1
Mark N
Spec Modif = A1
NIL
Mark N
Spec=Def
Modif
Mark blossom
Mark A
Spec
Modif
Mark cherry
the
NIL
blossom
Mark white
cherry
white

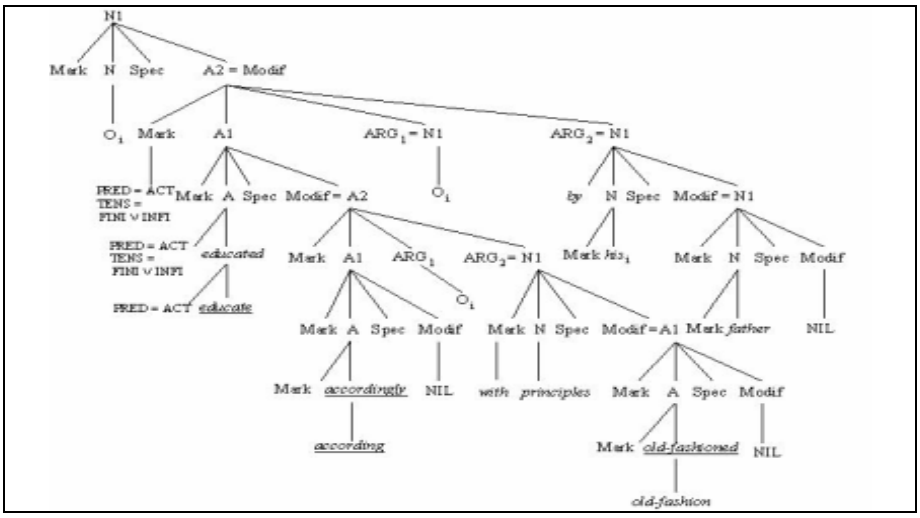
LR.4.1.1.R. de(cireş)(albă(floare(X)));
LR.4.1.1.E. quant(indef, X, white(blossom(X)), cherry(X)).

Ex. 4.1.2.R. */ floare albă /, / / <u>bătută</u> / / de vânt /*
Ex. 4.1.2.E. */ the white blossom /, / / <u>blown</u> / / by the wind /*
$object_i = O_i$; $event_j = e_j$



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

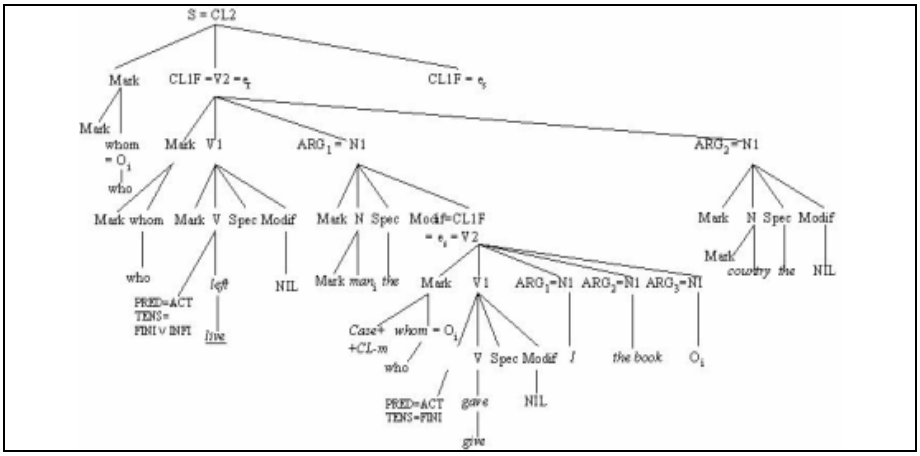
Ex. 4.1.3.R. */ / <u>educat</u> / / [de tatăl său] / / <u>corespunzător</u> / / cu vechile principii /*
Ex. 4.1.3.E. */ / <u>educated</u> / / [by his father] / / <u>accordingly</u> / / with old-fashioned principles /*



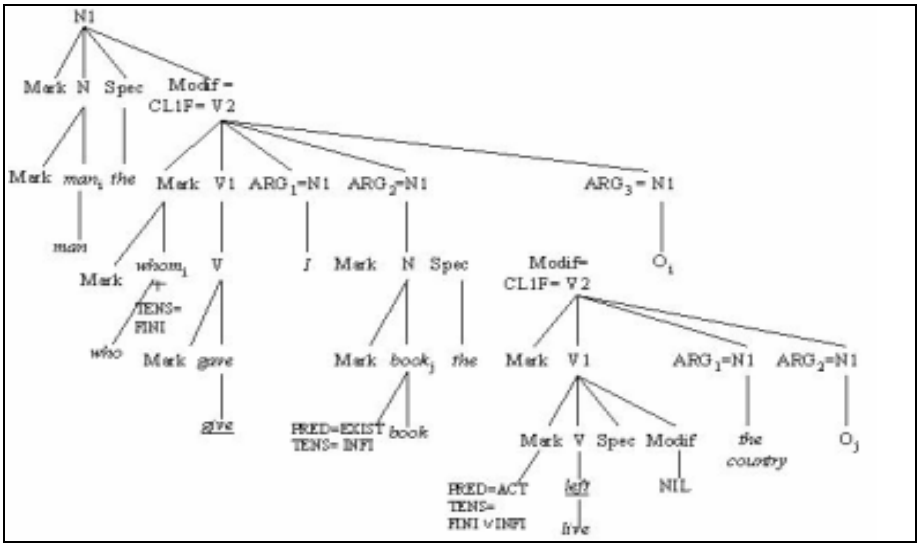
LF.4.1.3.R. corespunzător(X, cu(vechile(principii(Y))))
(educat(X, de(său(tatăl(Z)))));
LF.4.1.3.E. quant(indef, X, educated(X, by(his(father(Z)))), quant(indef, Y, with(old(principles(Y))), accordingly(X, Y))).

Ex.4.1.4.R. / / *Omul*$_i$ / / *căruia*$_i$ / / PRO$_j$ / i_i -*am dat* / / *cartea* / / PRO$_i$ *a părăsit* / / *ţara.* / /
Ex.4.1.4.E. / / *The man*$_i$ / / *whom*$_i$ / I_j / *gave* / / *the book* / / PRO$_i$ *left* / / *the country.* / /

4.1.4.E. Reading 1 (*left* = past_tense(*leave*))



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



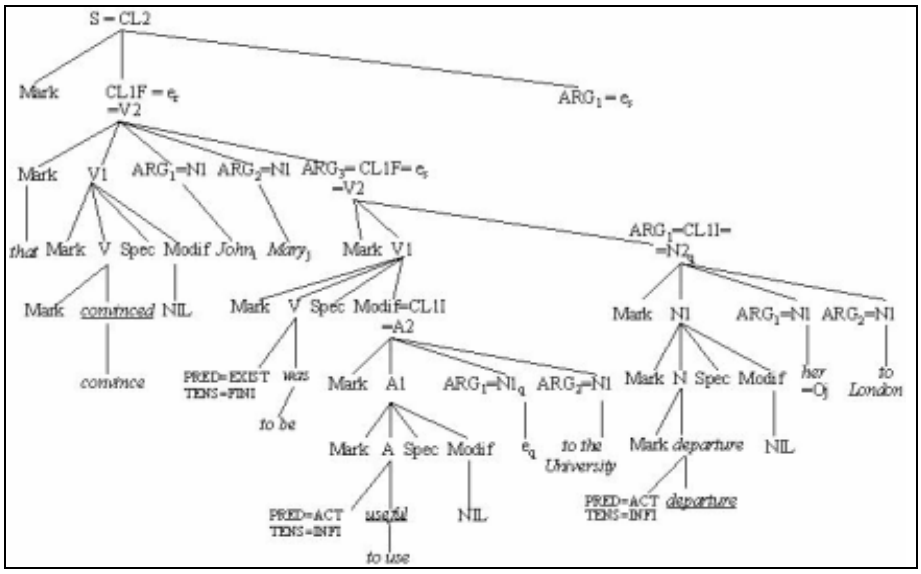
4.1.4.E. Reading3 ([Eng: *left*] = [Rom: *stânga*]) **. .**

. . .

LF.4.1.4.R. a-părăsit(omul(X) ∧ am-dat(Y, cartea(Z), X), țara(T);
LF.4.1.4.E. quant(def, X, and(man(X), quant(def, Y, I(Y), quant(def, Z, book(Z), gave(Y, Z, X)))), quant(def, T, country(T), left(X, T))).

Ex. 4.1.5.R. */ / Ion*$_i$ */ / a convins-o*$_j$ */ / pe Maria*$_j$ */ / că / / deplasarea*$_k$ */ / ei*$_j$ */ la Lodra / / a fost utilă / / e*$_k$ */ Universității . / /*
Ex. 4.1.5.E. */ / John*$_i$ */ / convinced / / Mary*$_j$ */ / that / / her*$_j$ */ / departure*$_k$ */ / to London / / was useful / / e*$_k$ */ to the University . / /*

object$_i$ = O$_i$; event$_j$ = e$_j$



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

4.2. Observatii generale

(♠1) Nu este scopul prezentei lucrari sa arate cum sunt obtinute reprezentarile FX-bar ale structurilor LN (într-o maniera mai mult sau mai putin algoritmica), ci doar sa propuna schema FX-bar generala ca un mecanism esential de reprezentare a informatiei lingvistice, sa sugereze cum lucreaza, si sa explice ratiunile introducerii acestui mecanism. Teoria FX-bar este integrata ca o componenta importanta a strategiei lingvistice SCD, însa ea poate utilizata si în alte contexte computationale, cu conditia de a include ingredientele necesare, si anume, clasele de PS-Ms, ierarhiile acestor clase, o taxonomie functionala (predicationala) si relationala a categoriilor majore si a marcherilor, un algoritm (în particular, algoritmii SCD) de obtinere a structurilor de dependenta, etc. Aspecte mai detaliate ale SCD au fost prezentate în [1], [2], [6], [7]. (♠2) Functionarea corecta a schemelor FX-bar expuse arata clar cât de necesara este utilizarea (intensiva) a trasaturilor *predicative* si *functional-relationale* pentru fiecare categorie lexicala. Din experienta noastra în ce priveste analiza si generarea automata a limbii române [6], consideram ca accentul pus pe trasaturile functionale ale categoriilor gramaticale, cuplat cu punerea în evidenta a PS-Ms, reprezinta elemente-cheie în utilizarea cu succes a teoriilor X-bar curente în procesarea automata a LN si în cadrul unor teorii lingvistice moderne (UG, FG,

HPSG, etc.). (♠3) Punerea în valoare a trasaturilor *functionale* (în particular, predicationale) ale categoriilor majore N, V, A, si a celor *relationale* ale claselor de *marcheri* (marcheri numiti în literatura si "*cue phrases"* [Rom: *sintagme indicatoare*] [28], [31], sau *conective* [29], [30], etc., desi esentiale, nu poate rezolva toate problemele. De exemplu, asignarea dependentelor corecte în juxtapunerea de NGs este o problema binecunoscut de dificila, imposibil de rezolvat complet doar la nivel sintactic. Exista însa în prezent un puternic curent catre acest tip de abordari, aceasta deoarece ele reflecta mult mai adecvat structura reala a textului de LN (cel putin pentru o clasa larga de LNs europene). Aceste abordari pot diferi substantial în instrumentele si tehnicile de parsare, însa principiile ramân foarte similare (de exemplu, [19], [29], [31], [33], etc. (♠4) PS-Ms (marcherii de structuri sintagmatice) joaca un rol fundamental în delimitarea structurilor sintactice si semantice, si stabilirea dependentelor corecte între aceste structuri. SCD a pus accentul înca de la începuturi pe acest aspect [22]. Se remarca în prezent o întreaga miscare catre reconsiderarea rolului esential al marcherilor, în special la nivel de *discurs* si în analize complexe ale marilor unitati textuale (regasirea informatiei, rezumare automata, planificare si generare automata de text, etc.). Strategia SCD, cu componenta ei de teorie FX-bar, încearca sa puna la lucru întreaga paleta de PS-Ms, de la nivel *lexical* si de *coeziune* (locala), pâna la nivel de *discurs* (*coeziune* si *coerenta* globala), punând accentul pe sintaxa (nivelul de "*suprafata*", [Eng: *shallow*]) si pe un nivel minimal de semantism. În functie de problema de LN ce trebuie rezolvata, aceste nivel poate fi amplificat în mod corespunzator. (♠5) Cuplarea schemelor FX-bar cu: (*a*) clasele de marcheri SCD si cu ierarhia lor ce corespunde celor *patru* nivele de proiectie lingvistica din FX-bar [7]; (*b*) o taxonomie bazata pe predicationalitate a categoriilor majore N, V, A; (*c*) exploatarea maximala a trasaturilor *functionale* (predicationale) si *relationale* a tuturor categoriilor lexicale si nelexicale (deci a PS-Ms); (*d*) o schema X-bar simpla si unica, apelata recursiv pe cele patru nivele ale sale, pornind de la *lexicon* (conventional, BAR = –1) si pâna la nivelul de *discurs* al frazei multi-eveniment (BAR = 3), aceste aspecte reprezinta principalele diferente (si noutati) dintre teoria FX-bar si teoriile X-bar precedente. (♠6) Schema FX-bar poate fi de asemenea asociata cu un automat dependent de limbaj (pentru o larga clasa de LNs), care începe sa lucreze pentru fiecare fraza, primeste *on-line* cuvânt cu cuvânt, si se opreste odata cu semnul de punctuatie final al frazei. Pentru valori adecvate ale parametrilor de LN cum sunt *ordinea cuvintelor (argumentelor)* si *directia proiectie lingvistice* pentru categoriile majore si pentru marcheri, schema FX-bar poate reprezenta corect dependentele structurilor lingvistice (inclusiv pentru Ex.3.2.2.G).

5. Problema X-bar teoriei actuale

Mai este necesara X-bar teoria sau nu? Este teoria X-bar pe moarte sau nu? Care este valoarea teoretica si, mai ales, practica a (sub)teoriei X-bar în teoriilor lingvistice si a tehnologiilor actuale ale LN? Cum trebuie sa percepem în mod corect X-bar teoria atunci când, în aceeasi carte a lui Chomsky, gasim urmatoarele doua pasaje:

(**Chomsky1**): *"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 ca aceste citate nu sunt extrase din text astfel încât sa nu aiba relevanta în context, intentia de a provoca confuzie. Dimpotriva! De asemenea, scopul nostru nu este de a cauta o posibila incoerenta ci de a pune în evidenta noua pozitie a lui Noam Chomsky, între 1992 si 1995. Încercam sa deschidem o discutie pe aceasta tema deoarece consideram ca exista o problema, si ca ea este de o reala importanta.

În aceasta sectiune urmarim patru obiective: **(A)** Sa enuntam problema X-bar teoriei. **(B)** Sa rezumam solutiile existente în momentul de fata. **(C)** Sa stabilim rolul X-bar teoriei în interiorul contextului teoriilor lingvistice si sa sugeram posibile dezvoltari. **(D)** Sa specificam pozitia FX-bar schemelor propuse privitor la dilema eliminarii complete a X-bar teoriei si, în special, relatia noii FX-bar teorii conturate în contextul strategiei lingvistice SCD. **(E)** Cateva concluzii si perspective.

(A) Sa consideram urmatoarea *problema*: reflecta *teoria X-bar* o *realitate lingvistica* a LNs, si daca da, prin ce *mijloace* aceasta realitate lingvistica ar putea fi *cel mai bine* reflectata? Proiectia categoriilor lingvistice este un fapt lingvistic de netagaduit. Chomsky si alti distinsi lingvisti nu au fost în completa eroare în ultimii 25-30 de ani? Credem ca nu. Problema este daca teoria X-bar poate înca sa mai fie un *bun model*, sau vehicul, care sa exprime acest *fapt*, si cu ce pret de utilitate. *Principiul Proiectiei Extinse* [3, p. 55] si *Principiul Proiectiei Maximale* (propus în [19] si sectiunea 3.1.) au ca scop sa stabileasca forma si marginile cele mai probabile ale unitatilor textuale obtinute în cadrul procesului de proiectie a categoriilor lingvistice.

(B) Ipoteza (Chomsky1) de mai sus da un raspuns afirmativ la aceasta întrebare în timp ce (Chomsky2) reprezinta, aparent, opusul acestui raspuns. Abordarea din [3, Cap. *Categories and Transformations*] pentru ipoteza (Chomsky2) este ca disolutia schemelor X-bar, deci a proiectiei categoriilor lingvistice, poate fi înlocuita cu succes prin folosirea proprietatilor de functionalitate, predicativitate, tipologie si transformare intrinseci acestor categorii, desi aceste proprietati sunt reprezentate în [3] cu acelasi aparat X-bar pe care îl combat! În cadrul unei teorii a "*structurii sintagmatice pure*", operatiile unui sistem computational al NL "*construiesc recursiv obiecte sintactice*", iar "*categoriile sunt constructii elementare rezultate din proprietatile elementelor lexicale*", cu conditia "*sa nu fie adaugate obiecte noi în cursul procesarii, înafara de rearanjari ale proprietatilor lexicale*" [3]. Rezultatul pare sa fie spectacular: dispar nivelele de proiectie (în sensul teoriei X-bar), astfel spus, nu se face nici o deosebire între elementele lexicale si nucleele proiectate din ele, în timp ce "*teoria structurilor sintagmatice poate fi eliminata în întregime, se pare, pe baza celor mai elementare ipoteze*" [3, p. 294].

Nu ar fi pentru întâia oara când teoria lingvistica încearca sa renunte la (sub)teoria X-bar. Chomsky sugereaza ca nivelele de proiectie lingvistica pot fi înlocuite de catre "*proprietatile* (functionale *n.n.*) *ale elementelor lexicale*". Acesta este chiar cazul *gramaticii functionale* (FG) [25] în care, formal, lipseste teoria X-bar. Dar chiar si în gramatica functionala a lui Dik, continutul ascuns al teoriei X-bar este scufundat de fapt în cele *patru nivele* de structuri ierarhice ale *functorilor* si *operatorilor* ce se aplica pe categoriile si structurile cu care FG lucreaza la fiecare nivel sintactic. O situatie speciala avem în SCD, unde nivelele de proiectie a categoriilor lingvistice sunt recuperate pe baza unei functionalitati ierarhice a elementelor lexicale, iar FX-bar schema propusa poate fi utilizata (recursiv) ca un invariant sintactic constructiv al structurilor sintagmatice în cadrul proceselor de analiza si generare automata a LN (limbii române).

Schema FX-bar propusa (Fig. 3.2.1.) poate fi considerata ca un compromis, o negociere, între (Chomsky1) si (Chomsky2), deoarece (Chomsky2) se prezinta fara mecanisme concrete pentru a-si sustine ipoteza: în timp ce teoriile X-bar clasice nu mai pot fi utilizate ca instrumente operationale pentru a reflecta o viziune exclusiv functionala (si relationala) asupra sintaxei, teoria FX-bar propusa poate face acest lucru.

(C) Pozitia noastra privind problema **(A)** asupra teoriei X-bar poate fi rezumata astfel: (C1) Proiectia categoriilor gramaticale este un fapt lingvistic. (C2) Acest fapt poate fi corect reflectat prin "*nuclee*" si "*nivele* (bar) *de proiectie*" în interiorul schemelor X-bar, dar si prin proprietatile functionale "*intrinseci*" ale categoriilor lexicale si gramaticale. (C3) Teoria X-bar include deci o componenta de adevarata constructie lingvistica, iar ingredientul sau de baza este confectionat din relatiile functionale stabilite între elementele lexicale (si nelexicale) continute în cadrul schemelor X-bar. (C4) Atunci când proprietatile functionale ale categoriilor lexicale nu sunt evaluate si exploatate corespunzator, teoria X-bar este inconsistenta si produce dificultati de calcul si rezultate incorecte. (C5) Acestea sunt consecintele unui aspect mult mai general, si anume ca teoria X-bar nu trebuie sa fie vazuta ca o teorie gramaticala singulara, construita pentru sine, ci ca un dispozitiv component al unui *mecanism lingvistic teoretic si computational mai general*, ale carui principii sa guverneze teoria X-bar. Axiomatica (bazele constructive ale) teoriei X-bar trebuie sa fie un *rezultat* al bunei ei functionari, pe fenomenele concrete de limbaj, si nu invers! (C6) *Ad limitum*, se poate concepe ca mecanismul lingvistic teoretic mentionat mai înainte poate functiona si fara includerea dispozitivului reprezentat de teoria X-bar, asa cum încearca teoria MinP sa propuna în [3, Cap. *Categories and Transformations*] (dar folosindu-se în explicare tot de aparatul de reprezentare al teoriei X-bar), precum si în cazul FG [25].

(D) Consideram ca schemele (functionale) FX-bar propuse furnizeaza un (sub)sistem necesar si folositor în cadrul oricarei teorii sintactice asupra LN, inclusiv (si în special) pentru strategia lingvistica SCD. O conditie esentiala pentru schemele FX-bar este ca ele sa reflecte corespunzator proprietatile *functionale* si *relationale* ale categoriilor tuturor *lexicale* si *gramaticale*. Exemplele 4.1.1.-4.1.5. arata cum sunt construite schemele FX-bar, cum se obtin (prin apel recursiv pe nivele) structurile sintagmatice complexe ale

LN, si cum acestea ramân închise la operatorul de compunere (adjunctie) pe baza principiilor si regulilor SCD.

Schimbând perspectiva, prin definirea teoriei FX-bar ca o componenta a strategiei lingvistice SCD, si parafrazând formalismul bine-cunoscut al gramaticilor TAG [Eng: *tree adjoining grammar*], strategia SCD poate fi vazuta si ca o teorie a evaluarii si adjunctiei de FX-bar scheme. Este doar o mostra a rolului important pe care teoria X-bar îl poate înca juca în cadrul teoriei si 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 (si nu numai): Schemele FX-bar pentru X = N, V, A, CL (CL = clauza) trebuie concepute ca un set de invarianti sintactici (dinamici) ce pot fi folositi **(1)** la *reprezentarea* informatiei lingvistice la nivel de lexicon (în mod similar cu structurile de trasaturi lingvistice [18], dar într-o maniera mai simpla si mai regulata), si **(2)** la *procesarea* (*analizarea si generarea*) *automata* de text în LN (inclusiv, si mai ales, limba româna), de la structurile sintagmatice simple pâna la cele de discurs.

Derivarea de *automate* si *gramatici formale* bazate pe schema FX-bar, pentru analiza LN, ar fi o consecinta normala si o provocare a prezentei propuneri. Modul *recursiv*, *ascendent* si *incremental* (prin apelul de functii si relatii cu rol lingvistic multiplu), dar si *descendent* (bazat pe satelitii nucleelor semantice), utilizarea la maximum a *contextualitatii* marcherilor de toate tipurile poate reprezenta o motivatie naturala pentru cercetarea relatiei dintre strategia SCD (cu componenta ei de *teorie* FX-bar), si modelele generative generoase oferite de catre *gramaticile contextuale Marcus* [41], [42], un formalism *context-dependent* puternic, destinat reprezentarii, parsarii, dar si analizei semantice si de discurs (articularea *topic-focus* [37]) a LN. *Gramaticile contextuale Marcus* apartin unei serii de formalisme care includ *gramatici* TAG [43], *gramatici orientate-nucleu* [15], [16], *gramatici indexate*, *gramatici* X-*bar*, *gramatici* context-free *marcate* [44] etc., formalisme ce realizeaza o modelare mai realista a comportamentului sintactic, semantic si discursiv al LN.

Referinte bibliografice

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

[2] N. Curteanu (1994). *From Morphology to Discourse Through Marker Structures in the SCD Parsing Strategy. A Marker-Hierarchy Based Approach*. Language and Cybernetics, 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 lingvistica SCD aplicata la analiza si generarea limbii române*. Limbaj si Tehnologie (Dan Tufis, Ed.), Academia Româna, Bucuresti, p. 169-176.

[7] N. Curteanu, C. Lintes (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 Functional Grammar*. Foris Publishers, Dordrecht.

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

[16] Carl 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. Todirascu, G. Holban (1997). *Teorii sintactice ale limbajului natural*. Raport de cercetare, Institutul de Informatica Teoretica, Academia Româna, Iasi, 66 p.

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

[22] Neculai Curteanu (1983). *Algoritmi de analiza sintactica a frazei si propozitiei*

românesti. *INFO-IASI'83*, p. 553-548.

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

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

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

[26] Robert Kasper (1993). *Adjuncts in the Mittelfeld*. In "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*. Computational Linguistics 19(3): 501-530.

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

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

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

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

[33] O. Popârda, N. Curteanu (2002). *L'évolution du discours juridique français analysé par la stratégie linguistique SCD*. In "Représentation du Sens Linguistique" (D. Bouchard, Ed.), ELCOM Studies in Theoretical Linguistics, ELCOM EUROPA.

[34] Noam Chomsky (2000). *Minimalist inquiries*: *the framework*. In 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*. In 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*. Computational Linguistics 17(1): 21-48.

[37] Eva Hajicova, H. Skoumalova, P. Sgall (1995). *An Automatic Procedure for Topic-Focus Identification*. Computational 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*. Computational Linguistics 16(1): 30-41.

[40] Steven Abney (1996). *Part-Of-Speech Tagging and Partial Parsing*. In "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*. In Cap. 5 (Vol. 2) din "The Handbook of Formal Languages", G. Rozenberg, A. Salomaa, Eds., Springer-Verlag, Berlin, 215-235.

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

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

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

Teoria HPSG. Studiu de caz: acordul încrucisat

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

1. Introducere

Oricât ar fi de mare entuziasmul creat de performantele realizate cu calculatorul, care cuprinde deopotriva si domeniul prelucrarii limbajului natural, rezultate temeinice nu se pot obtine daca acestea nu sunt fundamentate pe îndelungi si profunde analize teoretice. Nu putem aspira la obiective majore în ingineria ligvistica, precum analizarea si generarea de texte, construirea de verificatoare ortografice si gramaticale sau chiar de traducatoare automate, daca se ignora particularitatile inerente ale obiectului în studiu, anume ale limbajului natural în general, si a limbii de aplicatie, în special. Or aceste particularitati sunt oferite, sub un aspect sau altul, tocmai de teoriile gramaticale. Experienta a dovedit ca esecurile din ingineria lingvistica au avut ca posibile surse esecurile în descrierea corespunzatoare a fenomenelor de limba, dar si succesele, la rândul lor, s-au datorat în parte acuratetei, exactitatii, si nu în ultimul rând caracteristicilor computationale ale unui model gramatical teoretic.

Iata de ce alegerea unei teorii lingvistice adecvate, cu scopul de a scrie pe baza acesteia o gramatica computationala a unei limbi particulare, în speta a limbii române, este un act de prima însemnatate.

Dupa anii primelor dezvoltari ale gramaticii generative, sintaxa formala este, de aproape doua decenii, repusa în discutie ca obiect de studiu autonom distinct în acelasi timp de cel al lexicului si cel al sensului. Mai multe curente teoretice, cunoscute sub numele generice de „gramatici de unificare" sau „gramatici bazate pe constrângeri", s-au nascut din aceasta reconsiderare a sintaxei. Este vorba de modele recente (cele mai vechi datând de la începutul anilor '80), dezvoltate în cea mai mare parte în Statele Unite, si în general aproape necunoscute publicului român. Aceste modele se preteaza scrierii de gramatici pentru calculator, dar ambitia lor este mai întâi de a constitui teorii lingvistice de sine statatoare. Autorii lor se înscriu pe linia programului gramaticii generative chomskyene din 1957, de la care preiau grija pentru o formalizare operatorie a sintaxei, dar se disting suficient de modelul actual al Scolii de la Cambridge (numit Government and Binding) pentru a prezenta teorii alternative. Printre punctele comune ale gramaticilor de unificare, se afla pe de o parte atentia acordata unei articulari mai explicite a lexicului, sintaxei si semanticii, pe de alta parte accentul pus pe descrierile lingvistice si recurgerea la un stil de

analiza sintactica mai "concret", care limiteaza recurgerea la elemente "vide" (nerealizate concret) si care restrânge numarul etapelor intermediare în producerea unui enunt.

În acest articol vom prezenta pe scurt una dintre teoriile lingvistice amintite, anume „Gramatica sintagmatica ghidata de centru", denumita abreviat HPSG dupa numele sau din engleza „Head-driven Phrase Structure Grammar". Apoi vom ilustra modul în care poate fi aplicata aceasta teorie în reprezentarea unui fenomen mai special de limba româna prin aceea ca presupune dependente încrucisate de acord. Este vorba de structuri relative de tipul *baiatul a carui sora cânta* unde articolul genitival *a* se acorda cu substantivul *sora*, iar pronumele relativ *carui* se acorda cu substantivul *baiatul*.

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 Carl Pollard si Ivan Sag cu scopul de a permite o integrare mai explicita a diferitelor nivele de analiza lingvistica: fonetic, sintactic si semantic. El a luat nastere în principal din Gramatica Sintagmatica Generalizata (GPSG) si din lucrarile lui C. Pollard despre *Head Grammar* [1], dar autorii lor s-au inspirat deopotriva din numeroase alte teorii. Ei au preluat de la modelul chomskyan al Guvernarii si Anaforicitatii (GB) notiunea de modularitate si recurgerea la principii foarte generale (Principiul anaforicitatii, al controlului etc.). De la gramatica functionala de unificare FUG [2] au împrumutat reprezentarea uniforma a elementelor lexicale, a sintagmelor si regulilor gramaticale sub forma de structuri de trasaturi. S-au inspirat de la gramatica lexical functionala LFG pentru îmbogatirea cadrelor de subcategorizare si a notiunii de regula lexicala. Au luat de la gramaticile categoriale ideea de saturare progresiva a predicatelor si recurgerea la o ierarhie de functii gramaticale (cf. [3]). S-au inspirat, în sfârsit, dintr-un punct de vedere mai formal, din lucrari de logica si informatica asupra tipurilor si mostenirii.

Teoria este prezentata în cele doua lucrari ale lui C. Pollard si Ivan Sag: [4] si [5]. Majoritatea exemplelor privesc limba engleza si trateaza fenomene variate: fenomene de acord, constructii infinitivale, anafore, constructii relative si comparative. Fenomenele de control sunt totodata dezvoltate în [6], iar o analiza a anaforelor este propusa în [2]. Primele lucrari au conferit de la bun început o dimensiune multilinguala acestei teorii prin abordari privind germana ([8], [9]), catalana ([10]), japoneza ([11]), dar si coreana ([12]), franceza ([13]) si italiana ([14]).

C. Pollard si I. Sag preiau din modelul GPSG notiunea de gramatica sintagmatica, cu distinctia între o componenta ierarhica (schema DI –de dominanta imediata) si o componenta liniara (principii de precedenta liniara), precum si recurgerea la principii foarte generale de partaj si de propagare a trasaturilor. Totusi ei se separa de modelul original în câteva puncte. Structurile sintagmatice sunt în întregime exprimate în termeni de de structuri de trasaturi, cu introducerea unui atribut Ramuri. Structurile de trasaturi sunt la

rândul lor organizate în ierarhii de tipuri, comportând fiecare trasaturi predefinite. Modelul HPSG ofera astfel anumite simplificari în raport cu GPSG: întregul arsenal de reguli DI este redus la sase scheme de baza; metaregulile sunt eliminate în favoarea regulilor lexicale. S-a urmarit deosebirea clara între ceea ce tine de domeniul constrângerilor universale si ceea ce tine de descrierea unei limbi particulare. Principiile de coocurenta a trasaturilor din GPSG, care amesteca constrângerile universale si cele specifice unei limbi date, au fost suprimate.

2.2 Organizarea generala a HPSG

2.2.1 Caracteristici specifice gramaticilor de unificare

Se poate considera ca gramaticile de unificare, sau gramticile bazate pe constrângeri, reprezinta noile teorii sintactice ale anilor ‘80. Este vorba de modele care urmaresc o articulare explicita între lexic, sintaxa si semantica. Proprietatile lingvistice corespunzatoare sunt concepute ca “informatii” asociate morfemelor, sintagmelor sau constructiilor, combinate prin operatii variate, dintre care unificarea ocupa un rol central. Aceasta conceptie “integratoare” este unul dintre atuurile lor pentru tratarea automata a limbajelor naturale. Un alt avantaj este ca ele se bazeaza pe modele logice sau matematice (gramatici de constituenti, structuri de trasaturi), pentru care au fost definite metode de programare. Ele sunt în general rezultatul unui compromis între expresivitatea lingvistica (grija de a facilita exprimarea diferitor principii lingvistice adaugând u-se variante notationale sau operatori) si eficacitate (notatii concentrate, putine operatii).

Aici, ne vom rezuma sa punctam trasaturile lor comune cele mai pregnante, dintre care:

- reabilitarea descrierilor de suprafata;
- reînnoirea descrierilor sinatctice prin definirea de trasaturi complexe;
- definirea de principii generale de buna formare a enunturilor;
- integrarea lexicului, sintaxei si semanticii.

Gramaticile de unificare îmbogatesc aparatul formal al gramaticilor de constituenti cu un numar de notiuni importante. În acest capitol ne vom limita la prezentarea principalelor notiuni utilizate pe parcursul lucrarii, pentru detalii putând fi consultate S. Shieber 1986a sau H. Uszkoreit 1989.

2.2.1.1 Structuri de trasaturi

Structurile de trasaturi (engl. *feature structure*) sunt primitive ale teoriilor sintactice bazate pe unificare si reprezinta ansambluri de trasaturi, numite si complexe de trasaturi (engl. *feature complexes* sau *feature bundles*), care pot fi reprezentate sub forma de matrice. O **trasatura** este o pereche atribut-valoare, valorile putând fi simboluri atomice sau trasaturi. Trasaturile cu valoare non atomica conduc la structuri de trasaturi care prezinta îmbricari.

Spre exemplu, cuvintelor *acest* si *câine* li se asociaza o trasatura Cat cu valoare atomica (pentru categorie) si o trasatura complexa Acord care ia ca valoare conjunctia a doua trasaturi Num (pentru numar) si Gen:

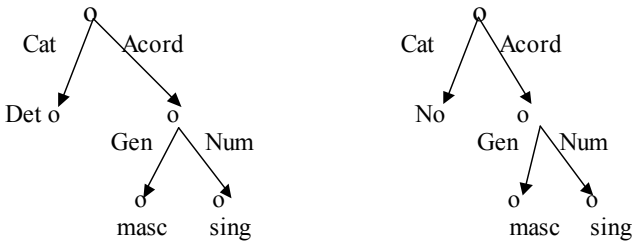
(1) acest câine

$$\begin{bmatrix} \text{Cat} = \text{Det} \\ \text{Acord} = \begin{bmatrix} \text{Gen} = \text{masc} \\ \text{Num} = \text{sing} \end{bmatrix} \end{bmatrix} \qquad \begin{bmatrix} \text{Cat} = \text{N} \\ \text{Acord} = \begin{bmatrix} \text{Gen} = \text{masc} \\ \text{Num} = \text{sing} \end{bmatrix} \end{bmatrix}$$

O structura este rau formata când contine de doua ori acelasi atribut (la acelasi nivel de îmbricare) cu o valoare diferita.

Si alte reprezentari de structuri de trasaturi (sau structuri atribut-valoare) sunt posibile, fiind echivalente formal. Cele mai utile, pentru implementarea informatica, sunt cele care utilizeaza grafuri orientate: arcuri care poarta nume de trasaturi si puncteaza spre noduri care sunt etichetate cu valoarea trasaturii (daca e vorba de trasaturi cu valoare atomica) sau sunt puncte de plecare pentru alte arce (pentru trasaturi cu valoare non atomica). De pilda, pentru exemplele de mai sus vom avea urmatoarele reprezentari:

(2)



În termeni de grafuri, echivalentul interdictiei ca un acelasi atribut sa apara de doua ori la acelasi nivel cu valori diferite este interdictia ca doua arcuri sa puncteze, plecând din acelasi nod, catre doua noduri diferite care poarta aceeasi eticheta (ceea ce e o restrictie generala asupra grafurilor ce corespund automatelor deterministe).

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

În lucrul cu structuri de trasaturi complexe se impun unele distinctii, de pilda, între structurile identice si **structurile cu valori partajate** (sau **reentrante**). Cele din urma sunt identice si vor ramâne astfel indiferent de modificarile suferite ulterior, ceea ce nu se întâmpla cu primele. În exemplul ce urmeaza structura de trasaturi A comporta doua

atribute cu valori identice Acord si Num. În structura B, cele doua atribute Acord sunt coindexate (prin indicele 1), ceea ce face ca ele sa partajeze în mod egal trasatura [Num = sing].

(3)

$$A: \begin{bmatrix} Det = [Acord = [Num = sing]] \\ Nume = [Acord = [Num = sing]] \end{bmatrix} \qquad B: \begin{bmatrix} Det = [Acord = |1| [Num = sing]] \\ Nume = [Acord = |1|] \end{bmatrix}$$

Daca se unifica fiecare din aceste structuri cu structura C de mai jos, rezultatul nu va fi acelasi:

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

(4)

$$C \cup A: \begin{bmatrix} Det = [Acord = [Num = sing, Gen = masc]] \\ Nume = [Acord = [Num = sing]] \end{bmatrix}$$

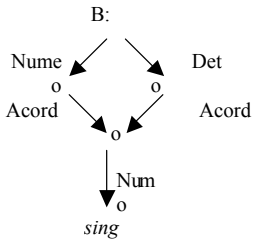
(5)

$$C \cup B: \begin{bmatrix} Det = [Acord = |1| [Num = sing, Gen = masc]] \\ Nume = [Acord = |1|] \end{bmatrix}$$

Dupa unificare, trasatura Acord îmbricata sub atributul Nume va avea si el o trasatura Gen specificata în cazul lui $C \cup B$, dar nu si în cazul $C \cup A$.

În termeni de grafuri, reprezentarea unei structuri reentrante ca B este urmatoarea:

(6)



2.2.1.2 Extensiune ?i unificare

Se defineste o relatie de **extensiune** între trasaturi dupa cum urmeaza:

O structura de trasaturi A este o **extensiune** a unei structuri de trasaturi B (notându-se A⊃B) daca si numai daca:

-- toate trasaturile cu valoare atomica prezente în B sunt prezente si în A cu aceeasi valoare,

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

De exemplu, structura de trasaturi asociata cuvântului *câine* în (1) este o extensiune a structurii din (7), dreapta, dar reciproca nu este adevarata pentru ca structura de mai jos nu are trasatura [Num = sing] prezenta în cea a cuvântului *câine*:

(7) $$\begin{bmatrix} \text{Cat} = \text{N} \\ \text{Acord} = \begin{bmatrix} \text{Gen} = \text{masc} \\ \text{Num} = \text{sing} \end{bmatrix} \end{bmatrix} \supset \begin{bmatrix} \text{Cat} = \text{N} \\ \text{Acord} = [\text{Gen} = \text{masc}] \end{bmatrix}$$

Daca numarul de atribute nu este limitat se pot obtine o infinitate de structuri care sunt extensii ale unei structuri date. Relatia inversa a extensiei se numeste **subsumare** , A subsuma B daca si numai daca B este o extensie a lui A.

Pe baza acestei relatii de ordine putem defini o structura de stiva, cu o limita superioara si o limita inferioara. Este de notat ca aici nu exista o relatie de ordine stricta pentru ca orice structura este o extensie a ei însesi (A⊃A). Structura care le subsumeaza pe toate celelalte (pentru care toate celelalte sunt extensiuni) este structura vida (notata $\mathcal{T}$), pe care o putem interpreta ca disjunctia tuturor cuplurilor atribut-valoare ale gramaticii. Daca dorim sa plasam o limita superioara, structura care va fi o extensie a tuturor celorlalte (care este subsumata de toate celelalte) va fi cea care contine conjunctia tuturor cuplurilor atribut-valoare posibile (notata ⊥) adica o structura "falsa" sau rau formata.

Aceasta relatie de ordine e folosita pentru a defini unificarea. Aceasta operatie a luat nastere din cercetarile în logica si informatica (limbajul Prolog). Definita la început ca procedura de rezolvare pentru logica predicatelor de ordinul întâi, cf. [15], ea a fost introdusa în lingvistica de A. Colmerauer, [16], apoi de M. Kay, [17], pentru a testa, fuziona si propaga trasaturi sintactice. Ea este definita în felul urmator:

Unificarea a doua structuri de trasaturi A si B (notata A ∪ B) este structura minimala care este în acelasi timp o extensiune a lui A si a lui B. Daca o astfel de structura nu exista, unificarea "esueaza" (ceea ce e notat cu ⊥).

Altfel spus, unificarea verifica compatibilitatea dintre doua structuri de trasaturi si produce o structura rezultanta care este cea mai mica structura ce contine toata informatia din prima structura si toata informatia din a doua structura.

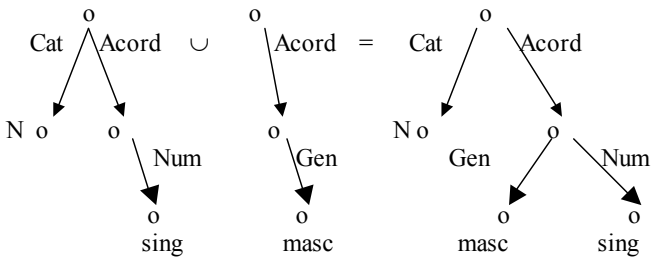
Unificarea este o operatie idempotenta (A ∪ A = A), comutativa (A ∪ B = B ∪ A) si asociativa (A ∪ (B ∪ C) = (A ∪ B) ∪ C), spunem de asemenea ca este declarativa (daca A = A' si B = B' atunci A ∪ B = A' ∪ B') si monotona (A ∪ B ⊃ A si A ∪ B ⊃ B; daca A

$\supset$ B atunci $\forall$C A $\cup$ C $\supset$ B $\cup$ C), ceea ce vrea sa spuna ca relatiile de extensiune sunt conservate prin unificare. Colocvial spus, unificarea adauga informatie, fara sa o scada.

În termeni de grafuri, echivalentul operatiei de unificare este fuziunea definita pentru automatele cu numar finit de stari. Pentru exemplul din (8a) se obtine reprezentarea grafica din (8b):

(8) a. $\begin{bmatrix} \text{Cat} = \text{N} \\ \text{Acord} = [\text{Num} = \text{sing}] \end{bmatrix} \cup \left[\text{Acord} = [\text{Gen} = \text{masc}]\right] = \begin{bmatrix} \text{Cat} = \text{N} \\ \text{Acord} = \begin{bmatrix} \text{Gen} = \text{masc} \\ \text{Num} = \text{sing} \end{bmatrix} \end{bmatrix}$

b.



Anumiti operatori pot fi adaugati structurilor de trasaturi (cf. L. Karttunen 1984), cei mai utili fiind negatia (notata ~ sau ≠ pentru trasaturi cu valoare atomica) si disjunctia (notata prin acolade sau semnul /). Folosirea negatiei permite sa se renunte la anumite disjunctii. Exista de exemplu echivalenta între urmatoarele doua ecuatii, daca consideram ca atributul Mod are 8 valori posibile în româna (indicativ, conjunctiv, imperativ, prezumtiv, infinitiv, gerunziu, supin, participiu):

[Mod ≠ inf] ⇔ [Mod = ind/conj/prez/imp/ger/sup/part].

În sectiunea urmatoare vom trece la descrierea caracteristicilor specifice ale teoriei HPSG care o fac distincta de toate celelalte teorii bazate pe unificare. Trebuie spus de la bun început ca autorii modelului HPSG au preluat o multime de caracteristici ale teoriilor aparute anterior, inclusiv de la gramatica generativa, tocmai din dorinta de a aduna într-un singur formalism tot ce e mai adecvat pentru reprezenatrea lingvistica în general. Pentru o paralela detaliata între HPSG si alte teorii bazate pe constrângeri a se vedea [18].

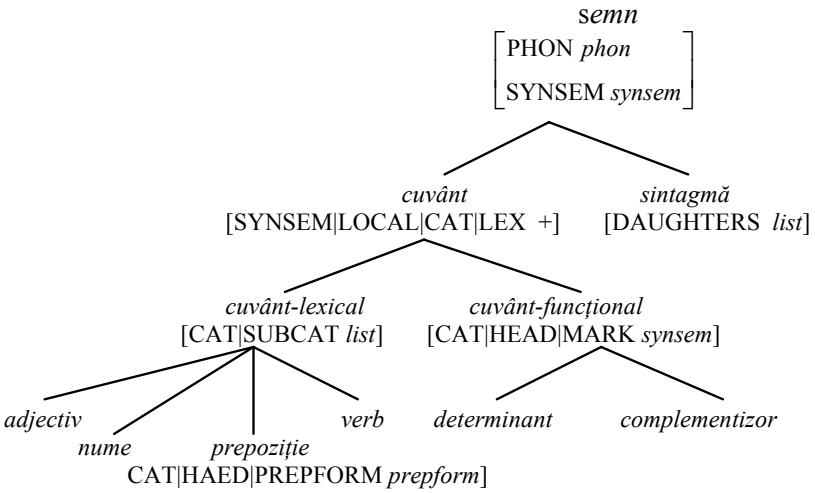
2.2 Caracteristici specifice HPSG

În HPSG, structurile de trasaturi, utilizate în LFG pentru reprezentarea functiilor gramaticale, iar în GPSG pentru reprezentarea categoriilor, sunt sistematizate pentru a include atât structurile de constituenti cât si regulile gramaticale. Ele corespund la ceea ce se numeste un **semn** lingvistic, adica un cuvânt, o sintagma sau o regula, continând

informatii fonetice, sintactice, semantice si discursive. Structurile de trasaturi sunt cât se poate de adecvate pentru organizarea într-o notatie comuna a informatiilor lingvistice eterogene.

Spre deosebire de celelalte teorii lingvistice bazate pe unificare, HPSG utilizeaza ierarhizarea tipologica. Fiecare structura de trasaturi este încadrata într-un anumit tip pentru care sunt predefinite anumite constrângeri si care îsi are locul într-o ierarhie de tipuri. În cadrul ierarhiei functioneaza relatia de mostenire a constrângerilor tipurilor superioare asupra descendentilor lor. Un exemplu de ierarhie de tipuri este data în (9).

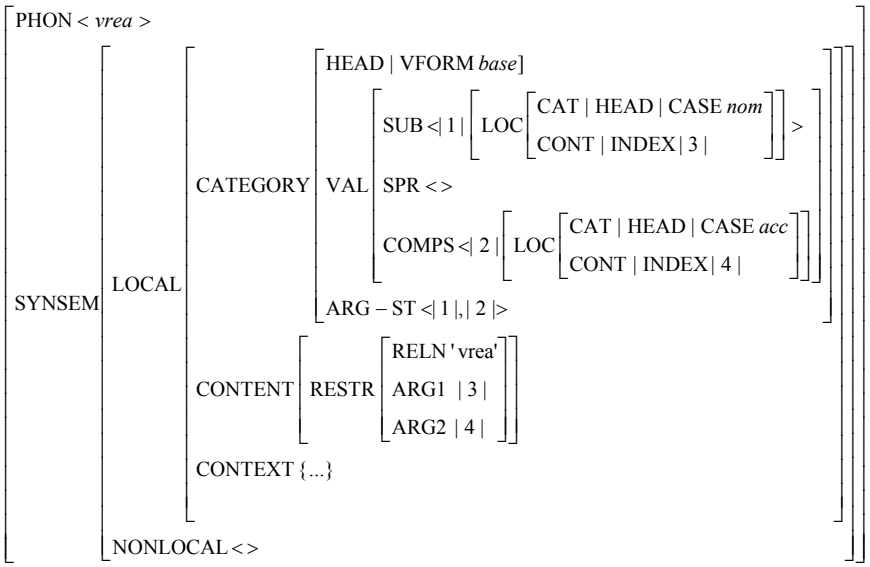
(9)



Pentru fiecare tip sunt definite anumite trasaturi specifice (sau anumite constrângeri) care se adauga constrângerilor mostenite de la tipurile din care descind. Trebuie adaugat ca într-o ierarhie de tipuri sunt permise mosteniri multiple, adica sunt permise tipuri care au mai multi parinti.

Cel mai general tip în HPSG este „semnul" (în engleza *sign*). El contine informatie fonologica (prin trasatura PHON) si informatie sintactico-semnatica (prin trasatura SYNSEM). Semnul, la rândul lui, poate fi un cuvânt sau o sintagma, dupa cum se vede în (9), mai sus. Sintagma are spre deosebire de cuvânt o trasatura în plus, numita DAUGHTERS (adica ramuri-surori) care are ca valoare o lista cu semnele combinate în sintagma. Un exemplu de semn lexical împreuna cu descrierea trasaturilor specifice acestuia este data în (10) pentru verbul *a vrea*.

(10)



Combinarea cuvintelor în sintagme se face pe baza unor reguli exprimate la rândul lor sub forma de structuri de trasaturi tipizate, purtând numele de scheme de Dominanta Imediata (scheme DI). Asupra regulilor actioneaza suplimentar principiile, care, la rândul lor, sunt exprimate prin constrângeri asupra anumitor trasaturi. În cele ce urmeaza vom prezenta pe scurt principiile si schemele DI de baza. Dintre principii, ne rezumam prezentarea la urmatoarele:

a. Principiul Trasaturilor Centrale
b. Principiul de Subcategorizare
c. Principiul Semantic

a. Principiul Trasaturilor Centrale

Pentru majoritatea sintagmelor se defineste un atribut HEAD („centru"), inclus în trasatura CATEGORY (CAT), a carui valoare trebuie sa fie partajata cu cea a atributului HEAD din semnul ramurii-centru HEAD-DTR a sintagmei. Principiul Trasaturilor Centrale poate fi exprimat prin descrierea urmatoare (notând valoarea partajata prin indicele [1]):

(11)
$$\begin{bmatrix} \text{SYNSEM} \mid \text{CAT} \quad [1] \\ \text{DAUGHTERS} \mid \text{HEAD - DTR} \mid \text{SYNSEM} \mid \text{CAT} \mid \text{HEAD} \quad [1] \end{bmatrix}$$

Semnul HEAD-DTR poate fi sintagmatic sau lexical.

b. Principiul de Subcategorizare

Atributul SUBCAT are ca valoare o lista care este actualizata progresiv, pe masura ce sintagma se "satureaza", în sensul ca atunci când complementele sunt realizate, ele sunt eliminate din lista SUBCAT a sintagmei respective. O sintagma se numeste saturata (sau completa) când valoarea listei SUBCAT este vida. Principiul de Subcategorizare poate fi enuntat astfel:

Valoarea listei SUBCAT a ramurii HEAD-DTR a unei sintagme trebuie sa corespunda concatenarii listei L1ca valoare a atributului SUBCAT al sintagmei si a listei L2 a semnelor ce apartin ramurii de complemente COMPS-DTR (sau, mai precis, nu lista semnelor, ci a trasaturilor SYNSEM a acestor semne).

Acesta poate fi reprezentat prin structura de trasaturi urmatoare (notând prin simbolul $\oplus$ concatenarea listelor):

(12)
$$\begin{bmatrix} \text{SYNSEM} \mid \text{CATEGORY} \mid \text{SUBCAT} \quad \text{L1} \\ \text{DAUGHTERS} \begin{bmatrix} \text{HEAD - DTR} \mid \text{SYNSEM} \mid \text{CAT} \mid \text{SUBCAT} \quad \text{L1} \oplus \text{L2} \\ \text{COMPS - DTR} \quad \text{L2} \end{bmatrix} \end{bmatrix}$$

Tinând seama de Principiul de Subcategorizare pot fi descrise urmatoarele doua scheme DI:

1. Schema DI pentru o sintagma saturata cu ramura Complemente: ***head-compl*** sau ***head-subject***

(13)
$$\begin{bmatrix} \text{SYNSEM} \mid \text{CATEGORY} \mid \text{SUBCAT} \quad <> \\ \text{DAUGHTERS} \begin{bmatrix} \text{HEAD - DTR} \mid \text{SYNSEM} \mid \text{CAT} \mid \text{SUBCAT} \quad < \text{X} > \\ \text{COMPS - DTR} \quad < \text{X} > \end{bmatrix} \end{bmatrix}$$

2. Schema DI pentru o sintagma non saturata cu ramura Complemente: ***head-compl***

(14)
$$\begin{bmatrix} \text{SYNSEM} \mid \text{CATEGORY} \mid \text{SUBCAT} \quad < \text{X} > \\ \text{DAUGHTERS} \begin{bmatrix} \text{HEAD- DTR} \mid \text{SYNSEM} \mid \text{CAT} \mid \text{SUBCAT} \quad < \text{X}, \text{Y1}, \text{Y2} \ldots \text{Yn} > \\ \text{COMPS- DTR} \quad < \text{Y1}, \text{Y2} \ldots \text{Yn} > \end{bmatrix} \end{bmatrix}$$

3. Schema DI pentru o sintagma cu ramura Adjunct: ***head-adjunct***

Modificatorii (adjective atributive, adverbe, complemente circumstantiale) sunt introdusi într-o ramura speciala numita ramura Adjunct (sau ADJCT-DTR). Modificatorii selectioneaza categoria pe care o modifica (N' pentru adjective, V sau GV pentru adverbe). Aceasta selectie se face printr-un atribut MODIF, care are ca valoare o structura de trasaturi SYNSEM. Pentru o sintagma centru-adjunct bine formata trebuie sa aiba loc unificarea valorii trasaturii MODIF a adjunctului cu valoarea trasaturii SYNSEM a centrului. Astfel adjectivele pot selectiona numele pentru care sunt atribute, iar adverbele pot selectiona verbele respective, adica se poate preciza în intrarea lor lexicala trasaturile Categorie, Continut, Index etc. ale numelui sau verbului asteptat. Descrierea unei sintagme cu Adjunct este urmatoarea:

(15)
$$\left[\text{DAUGHTERS}\left[\begin{array}{l}\text{HEAD - DTR} \mid \text{SYNSEM} \mid 1 \mid \\ \text{ADJCT - DTR} \mid \text{SYNSEM} \mid \text{CAT} \mid \text{HEAD} \mid \text{MODIF} \mid 1 \mid\end{array}\right]\right]$$

c. Principiul de Semantic

Principiul semantic reglementeaza propagarea trasaturilor semantice, adica cele doua trasaturi CONTENT si CONTEXT. Se urmareste pe de o parte ca sintagmele sa partajeze valoarea trasaturii CONTENT din ramura centrului cu trasatura proprie CONTENT, iar pe de alta parte sa determine "ridicarea" la nivelul sintagmelor superioare a eventualilor cuantificatori si a variabilelor care le pot corespunde.

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

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

O alta schema DI, *head-functor*, propusa de Allegranza în [19], reprezinta o modificare a schemei *head-adjunct* cu scopul de a satisface exigentele de reprezentare a determinatorilor într-un grup nominal. Determinatorii sunt tratati ca functori aplicati centrului. Ei selecteaza centrul prin atributul ARG-SLOT si marcheaza sintagma rezultata cu anumite trasaturi specifice determinatorului respectiv prin partajarea valorii atributului MARKER între ramura Functor si nodul mama. Descrierea acestei scheme este data mai jos.

4. Schema DI pentru o sintagma cu ramura Functor: ***head-functor***

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

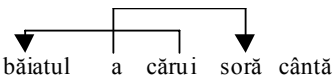
Cu aparatul formal oferit de HPSG, în sectiunea care urmeaza, dam spre exemplificare analizarea unei structuri concrete din limba româna. Structura propusa contine un centru nominal modificat de o propozitie relativa al carei element de relatie este în cazul genitiv precedat de articolul genitival. Aceasta structura este interesanta prin faptul ca prezinta un fenomen, acela de acord încrucisat, care pare sa scape reprezentarilor gramaticilor independente de context. Avantajul teoriei lingvistice discutate aici, însa, ofera o solutie pe cât de unitara, pe atât de eleganta, dupa cum speram sa reiasa din cele ce urmeaza.

3. Structuri relative cu acord încrucişat

În limba româna structurile care prezinta acord încrucisat sunt propozitiile relative în care pronumele relativ este precedat de articolul genitival, ca în exemplul de mai jos.

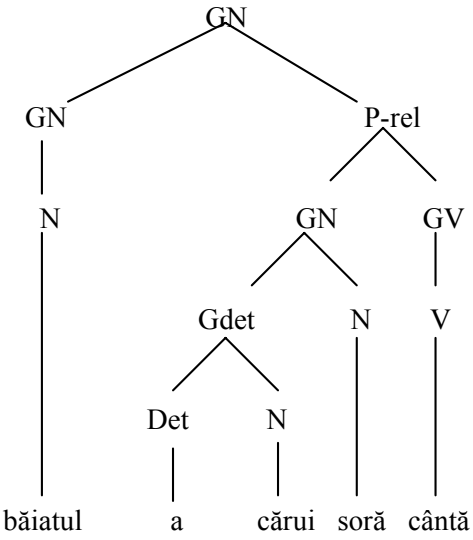
(17) baiatul a carui sora cânta

Acordul este încrucisat prin aceea ca pronumele relativ propriu-zis se acorda cu substantivul determinat de propozitia relativa, *baiatul*, iar articolul genitival *al* se acorda cu subiectul relativei, *sora*, dupa urmatoarea schema:

(18)


Structura interna a acestui grup nominal este reprezentata în arborele de mai jos.

(19)

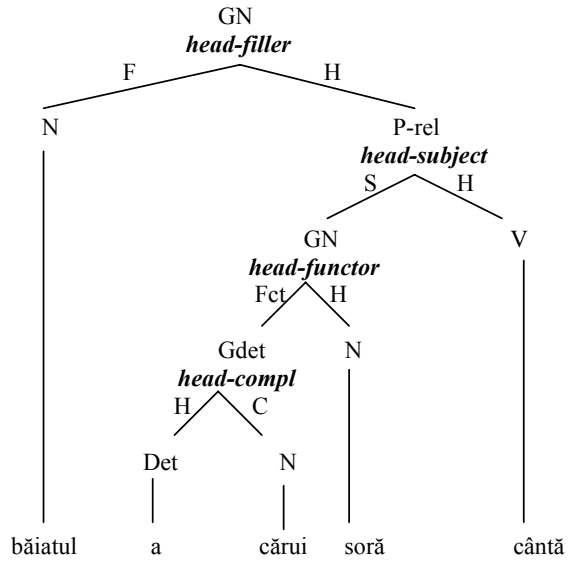


Dupa cum se vede în acest arbore, exemplul din (17) este format dintr-un substantiv centru, *baiatul*, modificat de o propozitie relativa al carei subiect, *a carui sora*, cuprinde elementul de relatie care face legatura dintre numele amintit si propozitia relativa.

Daca ne-am limita descrierea la regulile independente de context sugerate în arbore, nu am putea da seama de fenomenul de acord încrucisat pe care-l discutam aici. Acest lucru este însa posibil daca folosim o gramatica HPSG, beneficiind de avantajele oferite de mecanismul unificarii si de reprezentarile prin structuri de trasaturi.

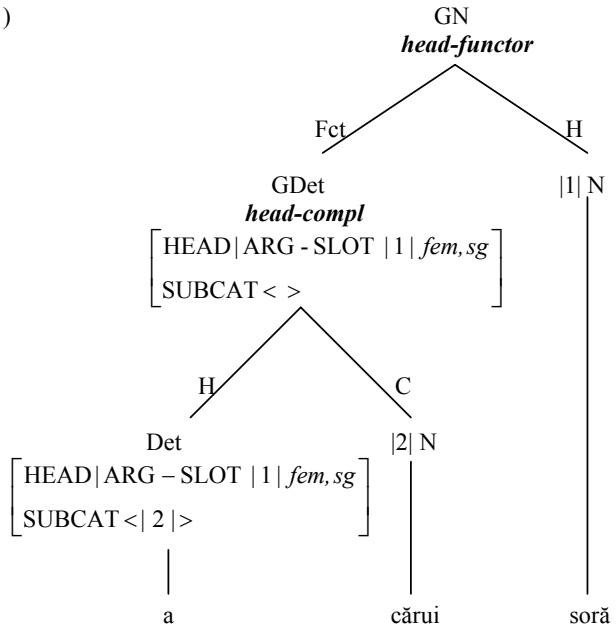
Aplicând schemele de dominanta imediata si principiile specifice teoriei HPSG, arborele de mai sus poate fi adnotat cu regulile HPSG aplicate, în felul urmator (unde am folosit ca notatii functionale H=centrul sintagmei, C=complement, Fct=functor, F=filler).

(20)



Fenomenul de acord încrucisat presupune pe de o parte acordul determinatorului *a* cu substantivul *sora*, iar pe de alta parte acordul pronumelui relativ *carui* cu substantivul *baiatul*. Primul acord amintit se face relativ banal. Intrarea lexicala a determinatorului *a*, în calitatea sa de functor, specifica în valoarea atributului sau central ARG-SLOT ce trasaturi de acord trebuie sa aiba substantivul pe care urmeaza sa-l modifice. Când detrminatorul *a* se combina cu complementul sau *carui*, principiul trasaturilor centrale face ca aceasta informatie sa fie percolata la nodul mama GDet. Mai departe, schema DI head-functor verifica daca trasaturile de acord ale GDet unifica cu cele ale centrului sau nominal. Acest mecanism este ilustrat în arborele de mai jos.

(21)



A doilea tip de acord, în schimb, ridica anumite dificultati prin faptul ca nu se realizeaza într-un arbore local, adica nu se realizeaza între ramurile surori ale aceluiasi nod. Prin urmare, trasaturile de acord ale pronumelui relativ trebuiesc percolate pâna la nivelul nodului P-rel (din (19))pentru a putea fi controlate prin unificare de regula head-filler cu trasaturile de acord corespunzatoare substantivului determinat.

Mecanismul din teoria HPSG care da seama de propagarea la distanta a anumitor trasaturi se numeste mecanismul dependentelor la distanta si se aplica fenomenelor de limba precum interogatiile, topicalizarile si, cum este cazul nostru, constructiile relative. Aici ne vom ocupa numai de tratarea relativelor, pentru celelalte fenomene a se vedea [5].

Ideea principala a acestui mecanism este ca pronumele relative poarta în intrarile lor lexicale informatii despre numele la care se refera. Intrarea lexicala a pronumelui relativ din exemplul nostru va contine, prin urmare, informatiile date în (22).

(22)
$$\left[\begin{array}{l}\text{PHONOLOGY} < \textit{carui} > \\ \text{SYNSEM}\left[\begin{array}{ll}\text{LOCAL} & \left[\begin{array}{l}\text{CATEGORY}\left[\begin{array}{l}\text{HEAD}\ \textit{noun}[\ \text{genitive}] \\ \text{SUBCAT} < >\end{array}\right] \\ \text{CONTENT}\ \textit{npro}[\text{INDEX}\ |\ 1\ |[\ \text{masculine}, \text{singular}\ ] \\ \text{CONTEXT}\,|\,\text{BACKGROUND}\left\{\left[\begin{array}{l}\text{RELN}\ \textit{human} \\ \text{INSTANCE}\,|\,1\,|\end{array}\right]\right\}\end{array}\right] \\ \text{NONLOCAL} & \left[\begin{array}{l}\text{INHERITED}\left[\begin{array}{l}\text{QUE}\ \{\} \\ \text{REL}\ \{|\ 1\ |\} \\ \text{SLASH}\ \{\}\end{array}\right] \\ \text{TO - BIND}\left[\begin{array}{l}\text{QUE}\ \{\} \\ \text{REL}\ \{\} \\ \text{SLASH}\ \{\}\end{array}\right]\end{array}\right]\end{array}\right]\end{array}\right]$$

Valoarea trasaturii NONLOCAL | INHERITED indica acele trasaturi care vor fi supuse Principiului Trasaturilor Nonlocale. Aceste trasaturi 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. Dupa cum se observa în (22), acest ultim atribut are în cazul de fata valoare non-vida, coindexata cu continutul semnatic de masculin-singular al pronumelui.

Potrivit Principiului Trasaturilor Nonlocale, formulat în (23), valoarea atributului nonlocal INHERITED („mostenit") este trecuta din nod în nod spre vârful arborelui pâna va întâlni o ramura sora ale carei trasaturi locale unifica cu cele mostenite.

(23) **Principiului Trasaturilor Nonlocale**

Pentru fiecare trasatura nonlocala, valoarea atributului INHERITED a nodului mama este egala cu reuniunea valorilor atributului INHERITED ale ramurilor fiice mai putin valoarea atributului TO-BIND a ramurii centru.

Atributul TO-BIND, practic, opreste propagarea trasaturilor mostenite în momentul în care se realizeaza elementul cautat, adica elementul care a facut necesara aceasta propagare. De exemplu, trasaturile de acord ale pronumelui relativ, în exemplul nostru *carui*, se propaga la nivelul propozitiei relative pâna când este realizat substantivul la care se refera acest pronume, adica *baiatul*.

Regula care asigneaza o valoare atributului TO-BIND în momentul în care are loc unificarea trasaturilor locale ale unui element cu trasaturile mostenite pe ramura centru este

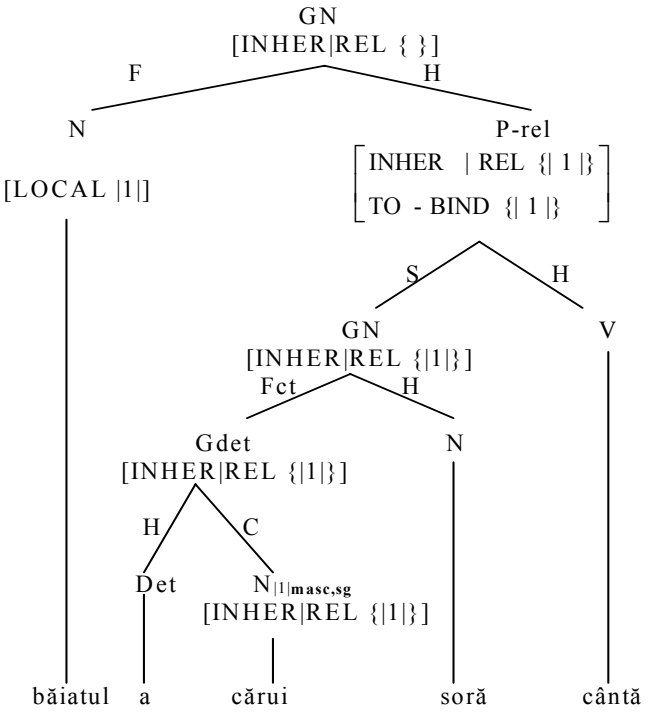
o schema de dominanta imediata numita *head-filler* (*filler* ar putea fi parafrazat drept „ceea ce vine sa completeze o lipsa") si este descrisa în (24).

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

$$\left[\begin{array}{l}\text{DAUGHTERS}\left[\begin{array}{l}\text{FILLER-DTR}\,|\,\text{SYNSEM}|\,\text{LOCAL}\,|\,1\,| \\ \text{HEAD-DTR}\left[\begin{array}{l}\text{SYNSEM}\left[\begin{array}{l}\text{LOCAL}|\,\text{CATEG}\left[\begin{array}{l}\text{HEAD}|\,\text{VFORM}\,\textit{finite} \\ \text{SUBCAT}<>\end{array}\right] \\ \text{NONLOCAL}\left[\begin{array}{l}\text{INHERITED}|\,\text{REL}\,\{|\,1\,|,\ldots\} \\ \text{TO-BIND}|\,\text{REL}\,\{|\,1\,|\}\end{array}\right]\end{array}\right]\end{array}\right]\end{array}\right]\end{array}\right]$$

În sfârsit, daca aplicam Principiul Trasaturilor Nonlocale si schema DI *head-filler*, acordul la distanta pe care îl avem în vedere se realizeaza în maniera ilustrata în arborele de mai jos.

(25)



În concluzie, acordul încrucisat avut în vedere presupune, pe de o parte, un acord local, cel dintre articolul genitival si substantivul determinat, în cazul nostru subiectul propozitiei relative, iar pe de alta parte un acord la distanta, cel dintre pronumele relativ si substantivul determinat, exterior propozitiei relative. Primul tip de acord se face pe baza Principiului Trasaturilor Centrale si a acordului banal dintre functor si centrul sau, pe când cel de al doilea tip de acord face uz de Principiul Trasaturilor Nonlocale si de schema de Dominanta Imediata *head-filler*.

4. Concluzii

Analiza oferita aici pune în lumina faptul ca un fenomen dificil precum acordul încrucisat poate fi tratat într-o maniera relativ simpla si eleganta cu ajutorul unei teorii lingvistice adecvate, cum este teoria Head-driven Phrase Structure Grammar.

Prin aparatul formal si adecvarea lingvistica pe care le ofera aceasta teorie, descrierea fenomenelor limbii române devine incontestabil mai unitara, mai explicita si mult mai riguroasa. Acestor avantaje li se adauga înca unul, extrem de important, acela al adecvarii teoriei pentru implementarea informatica. Este deschis astfel drumul pentru construirea de gramatici computationale ale limbii române si dezvoltarea componentei informatizate a acesteia.

Aplicatiile informatice ale teoriei HPSG sunt, de altfel, în plina dezvoltare si nu am dori sa încheiem înainte de a aminti câteva aspecte în acest sens.

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

Teoria HPSG a inspirat deopotriva noul formalism european ALEP, a carui implementare (în Prolog) presupune un mecanism de gestionare de gramatici si lexicoane, un analizor, un generator si un modul de transfer pentru traducerile automate. Este de altfel utilizat în mai multe centre de cercetare universitara (precum DFKI la Saarbrücken, *Center for Cognitive Science* în statul Ohio, CSLI la Stanford) sau industriale, în special la ATR în Japonia (pentru traducerea automata englezo-japoneza pentru stabilirea de întâlniri prin telefon).

O alta aplicatie informatica a acestei teorii, pe cât de recenta, pe atât de importanta este cea cuprinsa în proiectul Verbmobil, [26], care s-a ocupat cu traducerea bidirectionala, în timp real, a textelor vorbite în trei limbi (germana, engleza si japoneza).

Head-driven Phrase Structure Grammar este o teorie care s-a impus incontestabil în ligvistica moderna atât prin numeroasele sale aplicatii informatice, cât si prin

„generalitatea" aparatului sau care o face adecvata pentru numeroase limbi ale lumii, asa cum se poate vedea din impresionanta bibliografie electronica HPSG oferita de pagina www.dfki.de/lt/HPSG. Nu trebuie trecute cu vederea lucrarile de limba româna dezvoltate în acest cadru, dintre care le amintim pe cele ale lui Ionescu ([27]-[33]), Monachesi ([34]-[36]) si Barbu ([37]) la care s-ar cuveni sa se adauge multe altele spre afirmarea limbii române în lingvistica internationala.

Referinte bibliografice

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

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

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

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

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

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

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

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

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

[10] Balari, S. (1993) 'Feature structures, linguistic information and grammatical theory', Teza de doctorat, Universitatea Autonoma din Barcelona.

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

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

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

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

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

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

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

[18] Abeillé, A. (1993) *Les nouvelles syntaxes. Grammaires d'unification et analyse du français*, Armand Colin, Paris.

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

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

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

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

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

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

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

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

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

[28] Ionescu, E. (1995-1996), 'Accusative Weak Pronouns in Romanian', *Cahiers de Linguistique Théorique et Appliquée*, tomes XXXII-XXXIII, 1995-1996, 40-52

[29] Ionescu, E. (1995-1996), 'Accusative Clitic Doubling in Romanian', *Cahiers de Linguistique Théorique et Appliquée* tomes XXXII-XXXIII, 1995-1996, 53-73

[30] Ionescu, E. (1995-1996), 'Accusative Clitic Climbing in Romanian', *Cahiers de Linguistique Théorique et Appliquée*, tomes XXXII-XXXIII, 1995-1996, 74-87

[31] Ionescu, E. (1995-1996), '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. (1995-1996), *pro-Drop: An HPSG Account without Lexical Rules*, "Bucharest Working Papers in Linguistics", vol. I, nr.1, 1999, 117-124

[33] Ionescu, E. (1995-1996)*, 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. (1998) 'The morphosyntax of Romanian cliticization' în P.-A. Coppen, H. van Halteren, & L. Teunissen, eds., *Proceedingd of Computational Linguistics in The Netherlands 1997*, pp. 99-118, Amsterdam-Atlanta:Rodopi.

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

[36] Monachesi, P. (2000) '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

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

Dupa 10 ani de experienta terminografica: noul model de date terminologice al TermRom

Dan MATEI

A. Preambul

Din 1991 — când a fost înfiintata — Asociatia Româna de Terminologie (TermRom) a desfasurat o activitate terminografica materializata într-o baza de date proprie (accesibila, în parte, pe web la www.cimec.ro/tr/) si într-o serie de publicatii specifice. Formatul terminografic utilizat — descris în [Matei.1996] —, derivat din formatul standard MicroMATER (ISO 6156), se bazeaza pe un model de date (relativ) complex, serializat pe doua nivele: nivelul conceptului si nivelul termenului. Practica terminografica (ce se traduce prin prelucrarea unei mari diversitati de date terminologice) ne-a revelat o tensiune între complexitatea datelor reale si insuficienta complexitate a modelului folosit. În plus, necesitatea transferului de date între aplicatii diverse a scos la iveala utilitatea consemnarii cu o granularitate sporita a elementelor înregistrarii terminologice. Mai mult, „entuziasmul" cu care ISO revizuieste standardele terminologice în ultimii ani[64], cu alte cuvinte, relativa instabilitate a standardelor din acest domeniu, îndeamna la o si mai fina granularitate, pentru a spori sansele de compatibilitate cu normele de transfer viitoare. Pe de alta parte, pe masura acumularii experientei, era din ce în ce mai limpede ca modelul de date folosit ar trebui sa acomodeze o mai mare diversitate si complexitate de metadate bibliografice, ca si o fina si flexibila tratare a metadatelor „administrative", de gestionare a colectiei terminologice (vezi si [ISO 16642]).

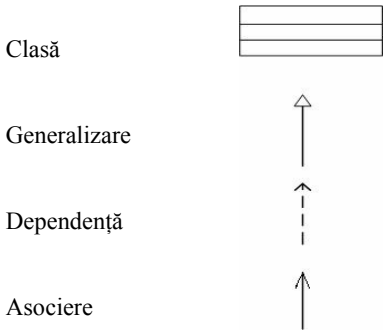
Aceste considerente au dus la elaborarea unui model de date obiectual, care, pe lânga cerintele expuse mai sus, sa fie si suficient de abstract ca sa permita o serializare convenabila (pentru transfer de date), — probabil bazata pe XML, de exemplu în formatul MARTIF [ISO 12200] — si sa nu ceara elaborarea de aplicatii informatice de o complexitate excesiva.

---

[64] *Atât [ISO 12200] cât si [ISO 12620] sunt în revizie (desi ambele dateaza doar din 1999), iar [ISO 16642], este înca nedefinitivat. Desigur, aceasta stare a lucrurilor probeaza si faptul ca domeniul nu este înca bine „asezat".*

B. Modelul

Modelul este prezentat în continuare, într-un formalism UML [Unified Modelling Language] (mult simplificat), folosind urmatoarele notatii:



Conventional, modelul este împartit în sectiuni („pachete" [packages], în terminologia UML). La nivelul cel mai de sus, se disting sectiunea (asa zis) functionala si sectiunea administrativa.

B.1. Sectiunea functionala

În fig. 1 se prezinta clasele functionale esentiale si asocierile lor. Practic, orice element al modelului este o 'înregistrare'. Cu alte cuvinte, 'înregistrare' este clasa generica. Existenta unei clase generice ofera — pe lânga gruparea proprietatilor comune tuturor elementelor — si posibilitatea de a avea un identificator unic pentru fiecare înregistrare din baza de date ce implementeaza acest model.

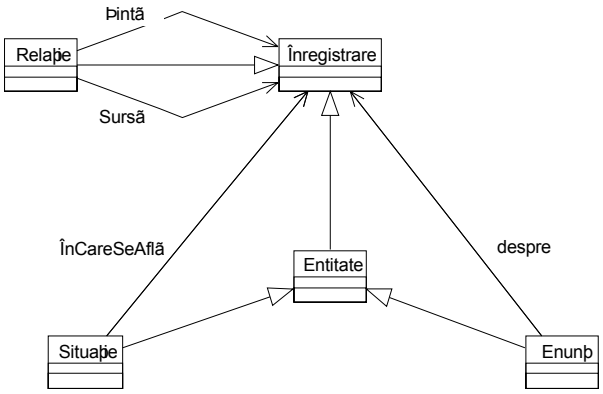


Figura 1 – Sectiunea functionala (generica)

Clasa 'înregistrare' are doua subclase: 'entitate' (care grupeaza elementele ce au o existenta autonoma) si 'relatie' (care grupeaza asocierile binare între înregistrari). Se observa ca sunt acceptabile chiar si relatiile binare între relatii, lucru folositor si în practica.

Reificarea relatiilor binare între înregistrari simplifica mult modelul si constituie o maniera flexibila de a consemna o mare varietate de asocieri între elementele modelului. O relatie R poate avea doua caracteristici fundamentale, utile în cadrul modelului:

a. simetria: daca x este în relatia R cu y, y este în relatia R cu x;
b. tranzitivitatea: daca x este în relatia R cu y si y este în relatia R cu z, x este în relatia R cu z.

Pentru fiecare instanta a clasei 'relatie', aceste caracteristici (sau lipsa lor) se consemneaza ca un atribut al tipului respectiv de relatii (nereprezentat grafic în model)[65]. Consemnarea acestor proprietati ale relatiilor poate fi foarte folositoare pentru programele care ar exploata baza de date.

Pentru a se rezolva (relativ) simplu si flexibil asocierile multiple între înregistrari, s-a introdus subclasa 'situatie' a clasei 'entitate'. Dupa cum se vede în figura, o instanta (sau mai multe) a clasei 'situatie' se asociaza cu o instanta a clasei 'înregistrare', iar obiectul 'situatie' este conectat cu oricâte alte elemente prin instante banale ale clasei 'relatie'. În practica, cele mai frecvente utilizari ale acestui tip de obiect sunt ca încarnari de contexte si evenimente. În fine, cea de-a doua subclasa a clasei 'entitate' este 'enunt'. Acest tip de obiect este destinat a consemna atribute ale unei înregistrari care n-au fost aprioric prevazute în model, cu alte cuvinte el gazduieste mentiuni pentru care se doreste un statut superior simplelor note, si anume care se doresc a fi colocabile si/sau indexabile.

În continuare se prezinta doar subsectiunile sectiunii functionale care sunt de interes în contextul acestui volum.

B.1.1. Sec?iunea terminologic?

Fig. 2 prezinta entitatile (i.e. subclasele clasei 'entitate') de natura terminologica.

[65] *O categorie de relatii — importanta în terminologie — este cea a relatiilor ierarhice, i.e. cele tranzitive si asimetrice.*

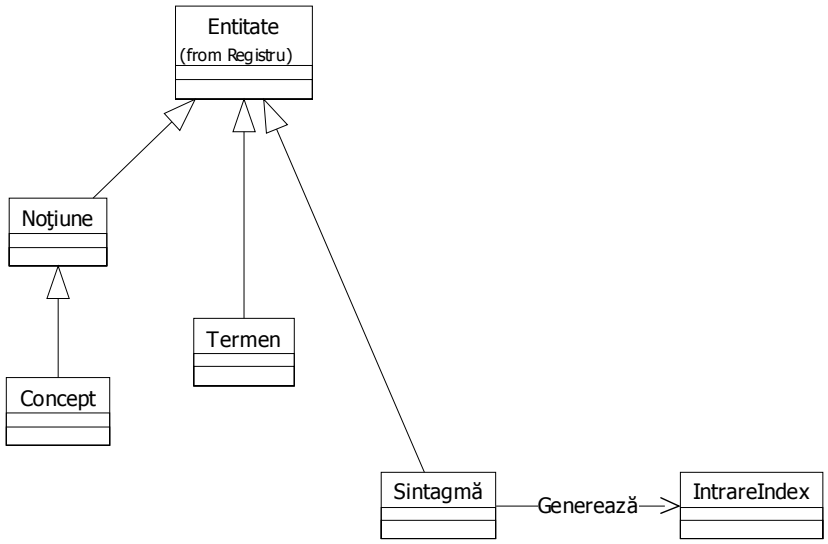


Figura 2 – Sectiunea terminologica

Principala clasa a acestei subsectiuni este 'notiune'. Instantele ei consemneaza notiunile vehiculate în baza de date terminologica, independent de limba. Din ratiuni practice, si anume din necesitatea de a cuprinde în baze de date terminologice si materialul organizat de obicei în tezaure terminologice, s-a decis sa se cuprinda în modelul de date nu doar conceptele pure, ci si unitati semantice mai largi, precum cele desemnate de termenii compusi într-un tezaur (sau ceea ce ISO 12620 numeste 'unitati frazeologice' [A.2.1.18]). Clasa acestor unitati conceptuale care cuprinde conceptele si unitatile semantice mai largi este clasa 'notiune'. Distinctie fina între 'notiune' si 'concept' este formulata în logica astfel [Chetan&Sommer.1978]:

Notiune: forma logica fundamentala care reflecta însusirile caracteristice necesare si generale ale unei clase de obiecte.

Concept: notiune care reflecta însusirile esentiale ale unei clase de obiecte[66].

Asadar, o notiune care nu e concept cuprinde mai multi factori semantici, deci poate fi factorizata[67].

[66] Exemple de noţiuni care nu sunt concepte: "bărbat blond", "scriitor important".

[67] *O regula simpla, pragmatica de a distinge o notiune care este concept de una care nu este, ni se pare: notiunea care e concept si-ar gasi locul într-un dictionar, pe când cea care nu e, nu.*

A doua subclasa a acestei sectiuni este 'termen'. Instantele ei consemneaza doar „denumirile" conceptelor (A.1. în ISO 12620). Cu alte cuvinte, consemneaza ceea ce au în comun o familie de expresii lingvistice ce designeaza un concept[68]. Expresiile lingvistice propriu-zise sunt consemnate în instantele clasei 'sintagma'[69]. Din pricina faptului ca un termen poate fi exprimat printr-un set de expresii lingvistice (flexiuni, variante ortografice etc.), s-a preferat separarea „termenului" de expresiile sale lingvistice, în felul acesta nu ne conformam strict definitiei pentru 'termen', din ISO 12620 (A.1): "a designation of a defined concept in a special language by a linguistic expression".

Se poate observa în figura faptul ca sintagmele genereaza intrari de index. În fapt, o sintagma poate genera — prin inversare/permutare — mai multe intrari de index, daca terminograful decide ca asta ar fi în folosul utilizatorilor, prin colocarea sintagmei la fiecare „factor" semnificativ. Exemple:

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

Clasa 'relatie' este vitala pentru consemnarea asocierilor între entitatile modelului. Pentru a ilustra modul în care se consemneaza informatia terminologica esentiala, în fig. 3 s-au reprezentat tipurile de relatii esentiale care asociaza, pe de o parte, conceptele cu termenii care le designeaza, iar pe de alta, termenii cu sintagmele care-i exprima. De asemenea, se vede cum o „situatie" (care — în aceasta ilustrare — implica (cel putin) un loc, o perioada si un agent) caracterizeaza designarea.

[68] *Exemple de "familie de expresii lingvistice" sunt: a) cladire, cladiri; b) expresiv, expresiva, expresivi, expresive.*

[69] *în acest context, 'sintagma' desemneaza — printr-un abuz de limbaj — atât sintagme cât si cuvinte.*

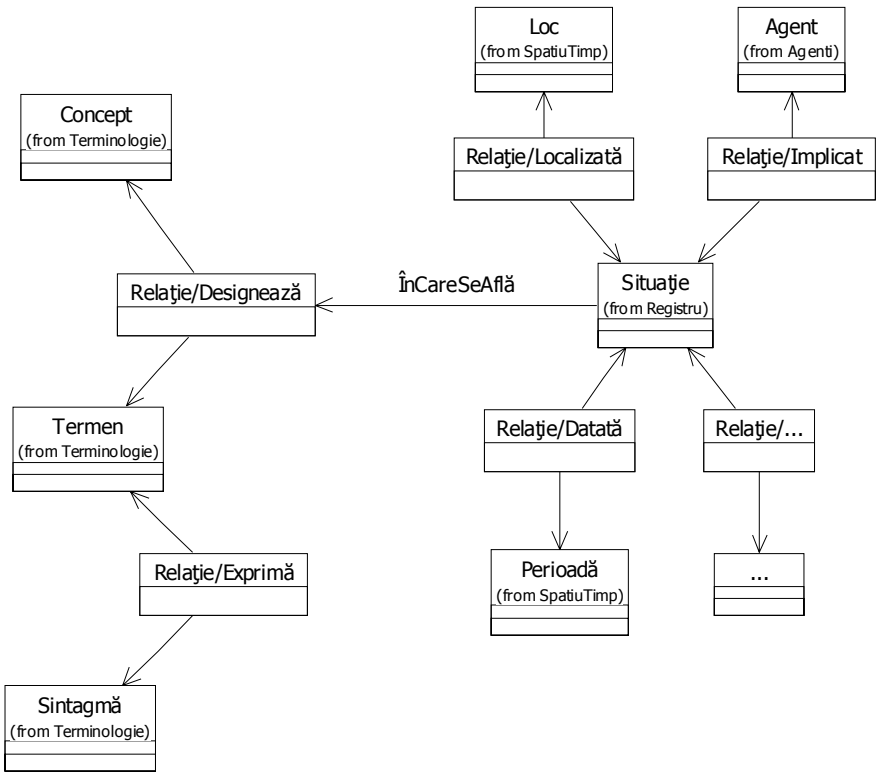


Figura 3 – Ilustrare a reprezentarii informatiei terminologice

Într-o astfel de schema, se pot reprezenta cu acuratete cazuri precum:

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

- Relatie/designeaza:
  Termen (stiintific) [latina]:
  Relatie/exprima:
  Sintagma: *Kalmia latifolia*
- Relatie/designeaza:
  Situatie/context:
  Relatie/localizeaza:
  Loc: *nordul Statelor Unite*
  Termen [engleza]:

Relatie/exprima:
Sintagma: *mountain laurel*

- Relatie/designeaza:
  Situatie/context:
  Relatie/localizeaza:
  Loc: *sudul Statelor Unite*
  Termen [engleza]:
  Relatie/exprima:
  Sintagma: *calico bush*
- Relatie/designeaza:
  Situatie/context:
  Relatie/localizeaza:
  Loc: *sudul Statelor Unite*
  Termen [engleza]:
  Relatie/exprima:
  Sintagma: *sheep's bane*
- Relatie/designeaza:
  Termen [româna]:
  Relatie/exprima:
  Sintagma [s.m.sg.]: *laur de munte*
  Relatie/exprima:
  Sintagma [s.m.pl.]: *lauri de munte*

b) Concept: comandant de calarime

- Relatie/designeaza:
  Situatie/context:
  Relatie/localizeaza:
  Loc: Moldova
  Relatie/localizeaza:
  Loc: Tara Românească
  Relatie/dateaza:
  Perioada: sec. XVII-XVIII
  Termen [româna]:
  Relatie/exprima:
  Sintagma [s.m.sg.]: *serdar*
  Relatie/exprima:
  Sintagma [s.m.pl.]: *serdari*

c) Concept: boier de rang mijlociu

- Relatie/designeaza
  Situatie/context:
  Relatie/dateaza:
  Perioada: *sec. XVIII-XIX*
  Termen [româna]:
  Relatie/exprima:
  Sintagma: *serdar* [s.m.sg.]
  Relatie/exprima:
  Sintagma: *serdari* [s.m.pl.]

Tot ca o ilustrare, în fig. 4 se prezinta modul cum se consemneaza etimologia unui termen, cu ajutorul clasei 'situatie': o situatie de tip 'etimologie' se asociaza cu termenul de baza, iar termenii din care acesta provine sunt asociati cu situatia prin intermediul unor relatii de tip 'provineDin'.

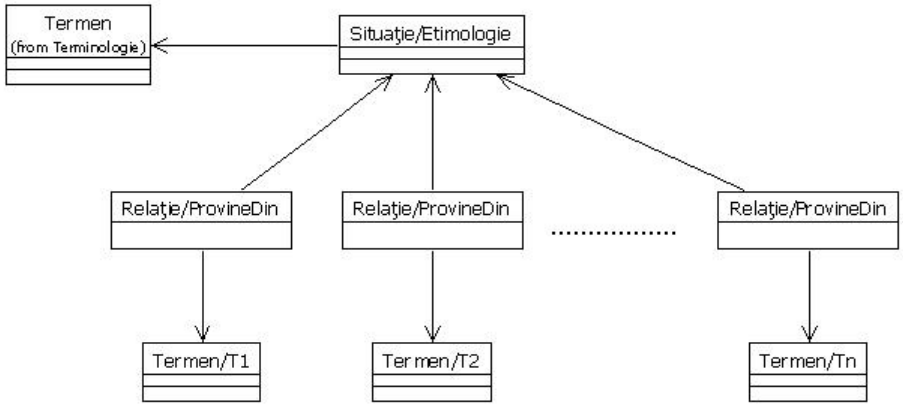


Figura 4 – Ilustrare a reprezentarii etimologiei

De pilda:
Concept: fixat la vârf
Relatie/designeaza:
Termen [româna]:
Relatie/exprima:
Sintagma: acrofix
Situatie/etimologie:

Relatie/provine din:
Termen [greaca]:
Relatie/exprima:
Sintagma: acro
Relatie/provinedin:
Termen [latina]:
Relatie/exprima:
Sintagma: fixus

B.1.2. Sectiunea bibliografica

Fig. 5 prezinta entitatile (i.e. subclasele clasei 'entitate') de natura bibliografica, cu alte cuvinte este o sectiune de metadate. Sectiunea pare simpla, deoarece o buna parte din multitudinea de date bibliografice sunt consemnate cu ajutorul relatiilor. Clasa esentiala este 'editie'; cea care consemneaza fisa bibliografica a unei editii citate.

Entitatea 'lucrare' consemneaza metadatele specifice unei creatii (mai ales textuale, în cazul nostru), i.e. „abstractizeaza" ceea ce au în comun toate editiile unei lucrari. Utilitatea ei imediata este colocarea tuturor manifestarilor unei lucrari, indiferent de limba sau editie. O subclasa importanta a clasei 'lucrare' este entitatea 'serial'. Aici se consemneaza si periodicele, adica entitatile ce grupeaza instantele clasei 'NumarPeriodic', cu alte cuvinte publicatiile-gazda ale articolelor. Discutia asupra acestor clase si a relatiilor între ele depaseste cadrul acestui articol.

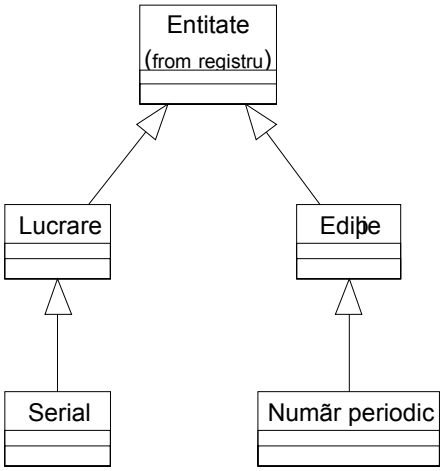


Figura 5 – Sectiunea bibliografica

B.2. Sectiunea administrativa

În fig. 6 se prezinta clasele de natura administrativa si relatiile esentiale între ele. Rolul acestor clase este de a consemna modificarile survenite în baza de date, în succesiunea lor. În acest fel se poate urmari geneza înregistrarilor si se pot identifica responsabilitatile. În plus, deoarece se prevede si stocarea datelor modificate, se creeaza premizele revenirii la stari anterioare ale bazei de date. În instantele clasei 'interventie' se consemneaza fiecare modificare operata asupra unei înregistrari. Fiecare asemenea instanta este asociata — prin intermediul instantelor clasei 'contributie' — cu agentul (i.e. operatorul) care a produs-o. În plus o interventie este asociata si cu sursele ei documentare. Se observa cum clasa 'referinta' poate avea ca instante atât referinte bibliografice (citând o editie), cât si referinte personale (citând o comunicare personala).

Clasa 'ÎnregistrareArhiva' este foarte importanta, instantele ei fiind chiar versiunile „desuete” (i.e. cele dinainte de modificari) ale atributelor înregistrarilor.

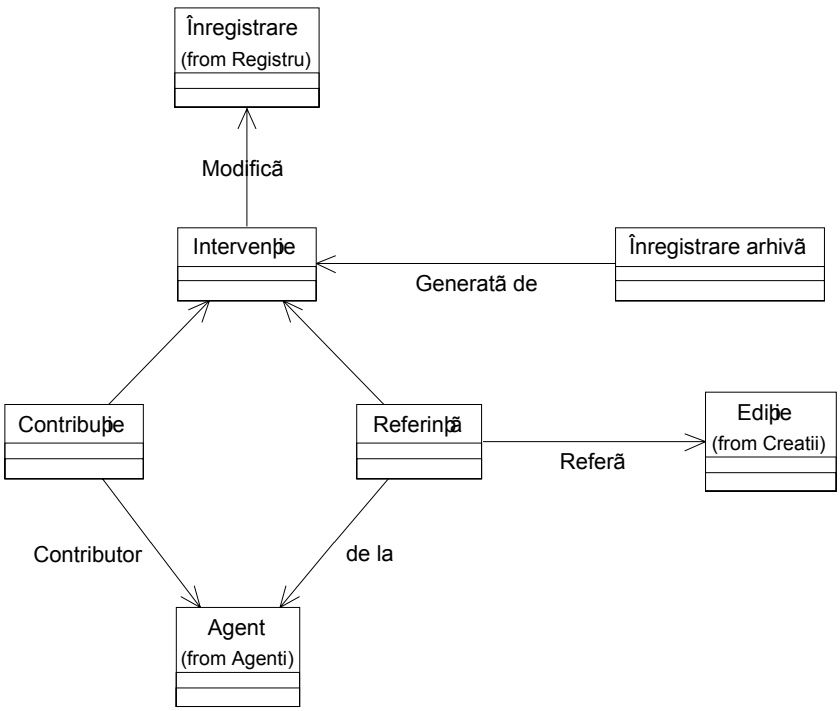


Figura 6 – Sectiunea administrativa

C. Remarci finale

Modelul prezentat pare suficient de flexibil pentru a satisface cerintele functionale atât ale unei baze de date terminologice, cât si a uneia lexicografice (mai ales datorita distinctiei între termeni si expresiile lor lingvistice). El este si suficient de abstract pentru ca schema unei baze de date ce l-ar folosi ca fundament sa fie relativ comoda la implementare.

TermRom are în curs un proiect de elaborare a unei astfel de baze de date terminologice. Dupa finalizarea acesteia, este de asteptat un proces traumatic de convertire a bazei de date curente. Sporul de functionalitate obtinut va compensa însa efortul.

D. Referinte

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

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

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

[ISO 16642] *ISO/CD 16642:1999, Computer applications in terminology – Metamodel for representing terminological data collections*

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

Probleme de reprezentare a datelor terminografice într-o baza de date relationala

Dr. Sorin GHETARU

Oriunde si oricând se creeaza, comunica, înregistreaza, prelucreaza, stocheaza, transforma sau refoloseste informatie sau cunostinte de specialitate este implicata într-un fel sau altul si terminologia. Comunicarea într-un anumit domeniu a devenit un discurs specializat cu texte de specialitate diferentiate în nenumarate forme. Atunci când se defineste terminologia ca o multime structurata de concepte si denumirile lor într-un anumit domeniu, ea poate fi socotita ca fiind infrastructura cunoasterii de specialitate. Scrierea textelor tehnice si documentarea tehnica devin astfel imposibile fara o utilizare corecta a unor resurse terminologice. Deoarece producerea textelor tehnice implica frecvent mai multe limbi, terminologiile multilingve de înalta calitate au devenit bunuri mult dorite greu de gasit pe înfloritoarea piata a industriilor limbajelor si cunoasterii.

Exista numeroase baze de date terminologice disponibile pentru interogare on-line sau pe CD-ROM (TERMIUM, EURODICAUTOM), pe dischete sub forma unor dictionare electronice sau ca baze de date personale realizate si întretinute de ingineri, specialisti în calculatoare, chimisti care lucreaza ca terminologi, traducatorii, autori de texte tehnice. Aceste baze de date sunt utilizate pentru:

- traducere asistata de calculator;
- scrierea de texte tehnice si stiintifice asistata de calculator;
- sisteme informatice (administrarea componentelor etc.);
- cercetari terminologice în lingvistica, filozofia stiintei, sociologia tehnologiei etc.

Pentru asemenea obiective au fost dezvoltate aplicatii specializate (programe de management al bazelor de date terminologice), unele disponibile pe piata terminologica internationala, altele ca prototipuri în cadrul unor proiecte de cercetare academice.

MARTIF este formatul standardizat pentru managementul informatiei terminologice. Posibilitatea organizarii terminologiei în baze de date având formate diferite face nerealista presupunerea ca s-ar putea cadea de acord asupra unui anumit format de baza de date relationala, asa cum este SQL, care sa fie folosit pentru schimburile terminologice. De aceea s-a mers pe linia producerii unui format la dispozitia publica fara obligatiii materiale si care sa fie independent de platforma de lucru. Rezultatul este MARTIF (**Ma**chine-**R**eadable **T**erminology **I**nterchange **F**ormat cunoscut si ca ISO 12200.

In ISO 12620 sunt descrise 150 de categorii de date, un numar imens care nu urmareste decât sa le arate pe cele posibile si modul în care acestea pot fi structurate. Categorille MARTIF sunt împartite în 10 sectiuni 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;
- tnformație 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);
- tnformație 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 ca MARTIF este scris folosind cod SGML este acela ca, desi se poate aprecia ca lectura codului nu este facila, ea este totusi posibila ca urmare a faptului ca nu face apel decât la caracterele ASCII. Un alt avantaj al sistemului MARTIF este acela ca el accepta referinte catre alte documente chiar din interiorul documentului. Initial MARTIF presupune ca înainte de implementarea produselor software pentru importul sau exportul datelor programatorii sunt obligati sa examineze sursele implicate. Pentru a asigura un acces asa numit "orb" care sa permita oricui sa transfere baze de date terminologice din orice sistem spre sau dinspre MARTIF este necesara o standardizare suplimentara a categoriilor de date, domeniilor specifice etc.

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

<termEntry>	Set complet unic de date terminologice pentru un concept exprimat într-o singura limba, si cuprinzând unul sau mai multi termeni si datele descriptive si administrative asociate lor, sau, în cazul unei abordari bilingve sau multilingve, doua sau mai multe concepte foarte apropiate, exprimate în fiecare limba, precum si datele descriptive si administrative asociate lor. Atributele includ: type, care clasifica setul de date terminologice conform categoriile de date specificate de ISO 12620.
<langSet>	Limba; în cadrul unui element <termEntry> va fi folosit pentru a grupa mai multe <tig> si <ntig> asociate unei singure limbi. Prezenta atributului lang este obligatorie, în afara cazului în care el este mostenit.

<tig>	Grup de informatii terminologice; în cadrul unui element <termEntry>, va contine elemente de informatii asociate cu un singur termen, fiecare dintre acestea functionând la acelasi nivel; cu alte cuvinte nu este permisa imbricarea între elementele subordonate unui <tig>. Prezenta atributului lang este obligatorie, în afara cazului în care el este mostenit.
<ntig>	Grup încuibat de informatii terminologice; va fi folosit în cadrul unui element <termEntry> daca anumite elemente informationale sunt asociate mai curând cu elemente interne, decât cu întregul <tig>. Urmatoarele elemente vor fi folosite în cadrul <ntig> pentru a gazdui alte date terminologice: <termGrp>, <termNoteGrp>, <descripGrp> si <adminGrp>. Prezenta atributului lang este obligatorie, în afara cazului în care el este mostenit.
<term>	Va contine un termen format dintr-un singur cuvânt sau din mai multe cuvinte, sau o desemnare simbolica privita ca un termen tehnic.
<termGrp>	Va contine un element <term> si, posibil, cel putin înca un element încuibat în plus fata de termen.
<termNote>	Va contine informatii legate de termen. Atributele includ: type, care clasifica <termNote> conform categoriilor de date specificate în ISO 12200.
<termNoteGrp>	Va contine un element <termNote> si posibil cel putin un element încuibat în plus fata de informatia legata de termen. Va fi folosit pentru a gazdui un nivel suplimentar de imbricare în cadrul elementului <termGrp>
<descrip>	Va contine informatii descriptive precum definitia, contextul sau explicatii descriind concepte si termeni. Atributele includ: type, care clasifica <descrip> potrivit categoriilor de date specificate în ISO 12200.
<descripGrp>	Va contine un element <descrip> si, posibil, cel putin un element imbricat în plus fata de informatia descriptiva.
<admin>	Va contine date administrative. Atributele includ: type, care clasifica <admin> în functie de categoriile de date specificate în ISO 12200.
<adminGrp>	Va contine un element <admin> si, posibil, cel putin un element imbricat în plus fata de informatiile administrative.
<date>	Va contine o singura data de formatul YYYY-MM-DD, cu optiunea notarii data-timp YYYY-MM-DD hh:mm:ss. Atributele includ: type, care clasifica <date> dupa categoriile specificate în ISO 12200.
<note>	Va contine o nota 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 informatiilor de tipul <note> folosite în cadrul <descripGrp> când continutul notei este legat de o lista de optiuni.
<adminNote>	Va fi folosit în cazul informatiilor de tipul <note> folosite în cadrul <adminGrp> când continutul notei este legat de o lista de optiuni.
<ptr>[70]	Va consta dintr-un indicator catre o alta locatie din documentul curent. Atributele includ: type, care clasifica <ptr> conform Anexei A, A.12 target, care precizeaza destinatia referirii, ca unul sau mai multi identificatori SGML.
<ref>	Va defini o referire catre o alta locatie din documentul curent, în termeni de unul sau mai multe elemente identificabile. <ref>GI este asociat cu text suplimentar drept continut al elementului, deci consta dintr-o eticheta-start cu o tinta integrata, urmata de textul asociat si închisa de o eticheta-sfârsit. Atributele includ: type, care clasifica <ref> conform Anexei A. target, care precizeaza destinatia referirii ca unul sau mai multi identificatori SGML.
<xref>[71]	Va defini o referinta la un grafic, ilustratie, figura, tabel sau alt document extern sau fisier folosind o notatie indicativa extinsa ca valoare a atributului tinta a <xref>, de ex. <xref target='documentIdentifier'>, unde valoarea 'documentIdentifier' este un cod de identificare pentru documentul tinta. Utilizatorul va documenta notatia indicativa extinsa care este folosita incluzând un comentariu adecvat în elementul <encodingDesc> ale header DTD. Atributele includ: type, care clasifica <xref> conform Anexei A. target, care precizeaza destinatia referirii ca unul sau mai multi identificatori SGML.
<hi>[72]	Va fi folosit pentru a marca un cuvânt sau o fraza ca evidentiat grafic în contrast cu textul înconjurator. Atributele includ: type, care clasifica <ref> conform Anexei A. target, care precizeaza destinatia referirii ca unul sau mai multi identificatori SGML.

[70] *Nota – <ptr> GI nu poate fi asociat cu text suplimentar drept continut al elementului, întrucât consta doar dintr-o eticheta-start cu o tinta integrata. Elementele <ptr>, <ref> si <xref> sunt toate considerate link-uri pentru ca ele conecteaza locatia lor curenta cu o alta locatie tintita în cadrul unui document sau cu o locatie externa documentului.*

[71] *Nota – Elementele externe tintite de <xref> trebuie sa fi accesibile sistemului-tinta pentru scopuri de importare.*

[72] *Nota – În managementul terminologiei o utilizare frecventa a <hi> se face pentru a sublinia termeni necesari, adica termeni folositi într-o definitie, nota sau alt material textual care sunt definiti altundeva în resursa terminologica. Vezi de asemenea Anexa A, A.2.2.2.*

<foreign>	Va identifica un cuvânt sau o fraza ca apartinând altei limbi decât cea a textului înconjurator. Atributele includ: lang, care identifica limba cuvântului sau frazei marcate.
<refObjectList>	Va fi folosit în back-matter si va contine unul sau mai multe obiecte back-matter, mai ales resurse comune ca: date bibliografice, date de responsabilitate, identificatori de namespace (URL-uri si FPI-uri), material textual la care se fac referiri dese, liste de locatii geografice, fisiere externe si altele asemenea. Atributele includ: type, care clasifica <refObjectList> dupa categoriile de date specificate în ISO 12620 Anexa A, A.11.4.1.
<refObject>[73]	Va contine o data constând în general dintr-o resursa comuna ca: date bibliografice, date de responsabilitate, identificatori de namespace (URL-uri si FPI-uri), material textual la care se fac referiri dese, liste de locatii geografice, fisiere externe si altele asemenea. Datele bibliografice ar trebui sa rezide în back matter sau într-un document extern (caz în care se va face referire la datele bibliografice din back matter folosind elementul <xref>). Atributele includ: type, care clasifica <refObject> dupa categoriile de date specificate în ISO 12620 Anexa A, A.11.4.2. Daca se specifica altfel, tipul <refObject> este mostenit de la <refObjectList> respectiv.
<itemSet>	Va fi folosit în back matter si va contine unul sau mai multe obiecte individuale care în mod traditional sunt grupate împreuna, de ex. obiectele numele autorului si prenumele autorului vor fi grupate împreuna într-un <itemSet> de tip=autor Atributele includ: type, care clasifica <itemSet> în principal conform categoriilor de date listate în ISO 12620 Anexa B. Totusi acest Standard International nu specifica întregul spectru al categoriilor de date care pot fi folosite cu <itemSet>
<item>	Va contine un exemplu individual de informatie back matter. Atributele includ: type, care clasifica <itemSet> în principal conform categoriilor de date listate în ISO 12620 Anexa B pentru informatii bibliografice. Totusi acest Standard International nu specifica întregul spectru al categoriilor de date care pot fi folosite cu <item>

[73] *Nota – Unele documente terminologice cuprind date bibliografice complete în format nediferentiat drept continut al categoriei de date sursa (vezi ISO 12620:1999, A.10.19). Aceasta practica încurajeaza redundanta si efortul marit pentru îngrijirea datelor. Aceste informatii ar trebui convertite în obiecte back matter (informatii bibliografice) daca este posibil.*

<itemGrp>	Va contine unul sau mai multe <item> împreuna cu <ptr>, <ref> sau <note>. Atributele includ: type, care clasifica <item> în principal conform categoriilor de date listate în ISO 12620 Anexa B pentru informatii bibliografice. Totusi acest Standard International nu specifica 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 informatiilor bibliografice. Instantierea elementelor enumerat mai sus se face prin intermediul „categoriilor de date" standardizate de ISO 12620. Numarul acestora este de aproximativ 200. În cea mai ampla resursa terminologica (EURODICAUTOM) sunt în prezent prezente mai putin de 20 astfel de categorii de date.

Uniunea Eropeana în activitatea sa este unul dintre utilizatorii majori ai procedurilor de translatare a textelor si terminologiei. Aceasta se datoreste partial faptului ca legislatia sa este direct aplicabila în statele membre si de aceea ea trebuie sa fie disponibila în toate limbile de lucru oficiale. Ca rezultat, traducatorii Comisiei Europene produc mai mult de 1 milion de pagini pe an si au de-a face cu cel putin 6-7 milioane de termeni (în medie sunt 8 sau 9 termeni care ridica probleme pe fiecare pagina).

Unitatea pentru Terminologie a Comisiei Europene este destinata asigurarii suportului lingvistic pentru toate limbile oficiale ale Uniunii Europene. Au fost elaborate glosare de specialitate, multe dintre le în noua limbi. Domeniile acoperite sunt tratatele importante cum ar fi cele de la Maastricht si Roma, cele economice si administrative (Taxa pe Valoarea Adaugata, buget) dar si unele legate de subiectele centrale sau puternic inovatoare ale stiintei si tehnologiei (fizica plasmei, biotehnologie, minerit). Deosebit de rolul lor de resurse terminologice si de surse terminologice pentru domeniile de inovare, aceste glosare documenteaza ceea ce se numeste "Eurolect", adica frazele si cuvintele care îsi au origina în cadrul Uniunii Europene si pentru care nu exista echivalente nationale.

Monitorizând toate modificarile aparute ca urmare a unei evolutii permanente a bazei de date EURODICAUTOM am constatat ca, recent, a avut loc schimbarea suportului hardware si odata cu aceasta pot fi observate urmatoarele:

- Indicarea mult mai frecventa a referintei la documentul sursa a termenului;
- Indicarea frecventa a referintei la documentul sursa al definitiei acestuia;
- Indicarea documentului sursa si pentru sinonime si abrevieri;
- Utilizarea mai frecventa a notelor pentru adaugarea unor informatii suplimentare asupra termenilor, acestea putând fi grupate astfel:
  - {NTE} explicatii si informatii generale asupra termenilor;
  - {TXT} contextul (de cele mai multe ori un exemplu de utilizare a termenului respectiv);
  - {GRM} informatii gramaticale (gen, numar);

- {USG} indicarea mediului în care este utilizat termenul: "*technical jargon*";
- {REG} nota asupra unor utilizari locale speciale sau asupra regionalismelor;
- {DOM} indicarea unui domeniu sau subdomeniu care complementeaza clasificarea obisnuita folosita anterior si care a ramas înca prezenta.

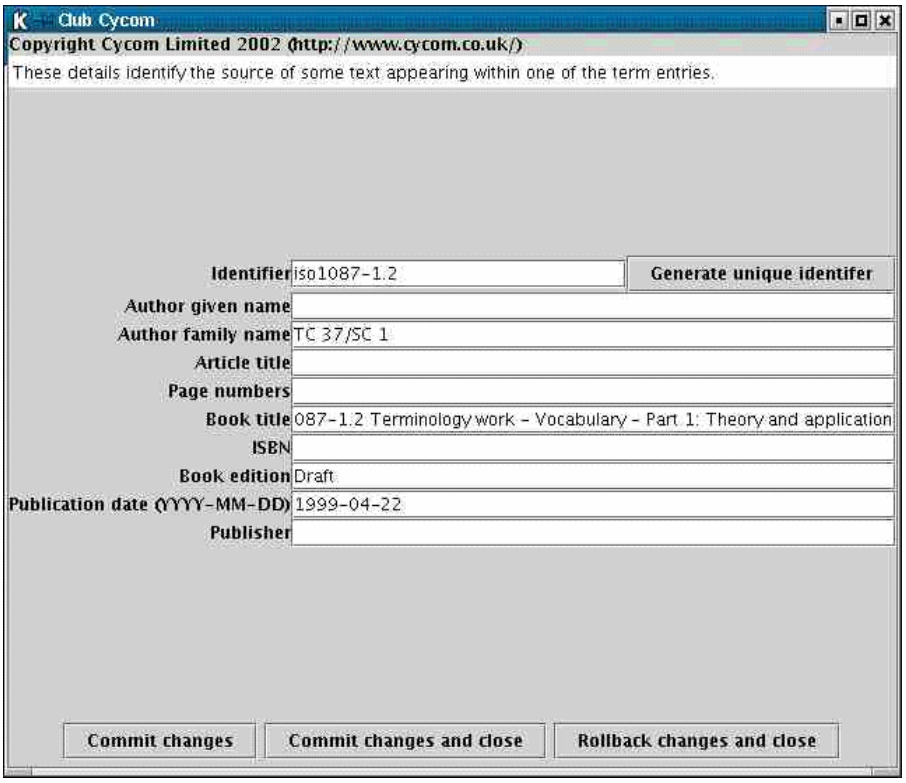
De asemenea se prevede ca în cel mai scurt timp sa fie implementate urmatoarele:

- afisarea tuturor caracterelor si diacriticelor (ca si a informatiei nelingvistice, daca se cere);
- îmbunatatirea sistemului de clasificare a domeniilor;
- introducerea link-urilor interne si externe.

Modelele de date terminologice orientate în exclusivitate catre terminologie au avantajul de a fi relativ intuitive pentru terminolog. Transcrierea directa a elementelor si reltiilor dintre acestea într-o baza de date este din ce în ce mai dificila si mai riscanta.

Exista încercari meritorii de realizare a unor interfete "cuprinzatoare" pentru consultarea resurselor terminologice. Exemplele urmatoare sunt edificatoare în acest sens.

Primul exemplu ar putea provoca comentarii legate de complexitatea reala a înregistrarii referintelor bibliografice cele mai obisnuite.



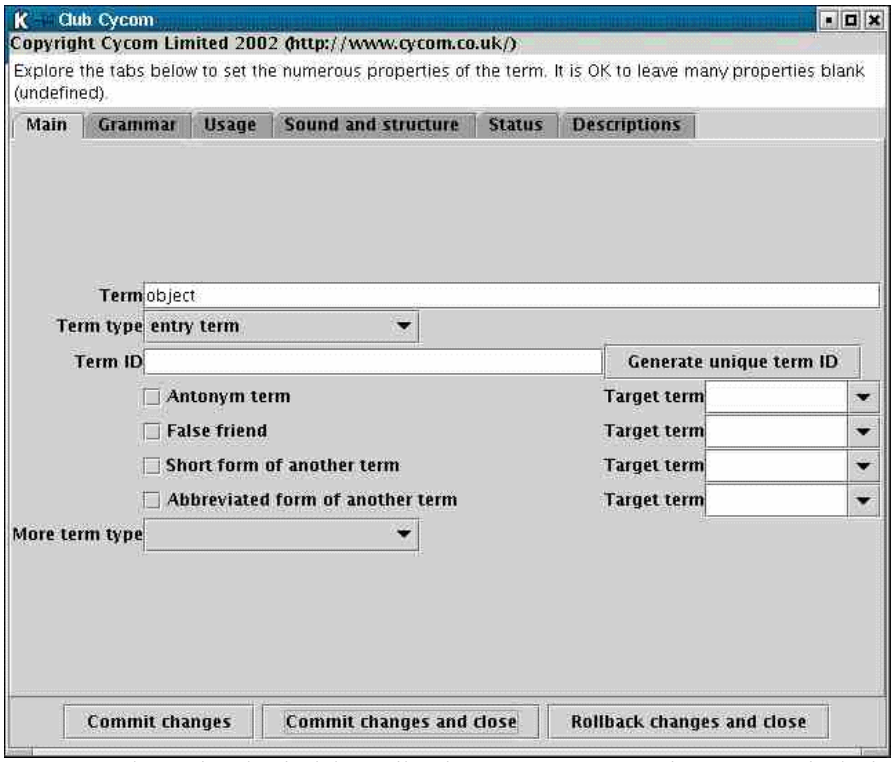
Al doilea, ne determina sa luam în considerare urmatoarele:

La nivelul Uniunii Europene numarul limbilor pentru care este necesar suport terminologic este atât de mare (si speram înca în crestere) încât nu mai este posibila multiplicarea tabelelor bazelor de date potrivit numarului de limbi de lucru. Din fericire, "balizarea" documentelor permite identificarea si prelucrarea corect dependenta de limba în care au fost concepute acestea. Se vine astfel în sprijinul "globalizarii" aplicatiilor informatice care sunt suport al resurselor terminologice multilingve dând posibilitatea acceptarii, prelucrarii si prezentarii numeroaselor scrisuri, formate de date si limbi existente. În acelasi timp trebuie adaptata si interfata utilizator potivit locului si culturii careia îi apartine acesta printr-un proces nu mai putin important de "localizare"

Multa vreme, prelucrarea automata a datelor a fost considerata satisfacator realizabila prin utilizarea setului ASCII de caractere. În prezent este însa absolut necesar ca:

- Utilizatorul calculatorului sa poata tasta caractere si simboluri (vest-europene, est-europene, grecesti si cirilice, cel putin) folosind o claviatura standard.
- Aplicatia sa prelucreze si sa afiseze sau imprime siruri de caractere formatate corect folosind seturi de caractere specifice fiecarei limbi.

Aceste cerinte pot fi realizate prin valorificarea calitatilor standardului Unicode de codificare prin utilizarea unor coduri de 16 biti pentru reprezentarea tuturor caracterelor pentu calculatoarele moderne care includ simbolurile tehnice si semnele speciale necesare imprimarii textelor.



Cu alte cuvinte la nivelul seturilor de semne necesare unei resurse terminologice multilingve se poate conta pe serviciile standardului Unicode si pe cele ale oricarei baze de date relationale care accepta Unicode.

Pentru indicarea formatelor de prezentare (fonte, punere în pagina, seturi de caractere) si a limbii utilizate se face apel la balizare astfel încât la nivelul câmpului vom gasi siruri de caractere Unicode balizate.

Înscrierea datelor terminologice este facilitata de înscrierea lor în „categorii de date” bine definite (vezi ISO 12620). Dar numarul mare al acestor categorii si mai ales

incidenta ridicata a aparitiilor neprevazute dinainte a unora noi face imposibila alocarea unui câmp de date fiecarei categorii de date. Aceeasi observatie poate fi facuta si asupra relatiilor dintre diferitele categorii de date care reflecta direct relatiile dintre elementele MARTIF. O solutie este o abstractizare suplimentara a datelor terminologice dupa încadrarea lor succesiva în siruri de caractere balizate, categorii de date, elemente MARTIF.

În centrul modelului de date se afla un set de 13 entitati (atomi):

Entitate	Descriere
data category	o anumita clasa de informatii terminologice (de exemplu: term, part of speech)
data category name	un nume agreat de utilizator (user-friendly), dependent de limba, al unei anumite categorii de date (de exemplu, în româna, "termen" pentru term)
data category index type	o strategie de indexare corespunzatoare unei anumite categorii de date (ISO 12620) (de exemplu: nu se indexeaza, se indexeaza ca valoare unica, se indexeaza cuvânt cu cuvânt)
lang	o anumita limba, care dispune de o schema de codare uniforma care utilizeaza un singur set de caractere (de exemplu: French, German, Italian)
charset	o combinatie unica de caractere care poate fi utilizata pentru reprezentarea unei singure sau mai multor limbi (de exemplu: ISO 8879-1. ISO 8859-2)
picklist	o multime de valori posibile ale unor date terminologice apartinând unei anumite categorii de date (ISO 12620) (de exemplu, pentru categoria "parte de vorbire": noun, verb, adjective)
element	o data terminologica unica
date value	o data (time stamp) care constituie valoarea unui element
number value	un numar care constituie valoarea unui element
picklist value	un membru al unei liste care reprezinta valoarea unui element
text value	sir de caractere care constituie valoarea unui element
index value	un sir de caractere care reprezinta forma normalizata indexata a unui element particular sau a unei parti a acesteia
link	legatura între doua elemente

Primele 6 "articole" sunt "meta-entitati"; ele sunt create si tabelele corespunzatoare sunt completate cu informatii înainte de încarcarea oricarei date terminologice în baza de date. Prin completarea acestor table se contureaza si se activeaza chiar modelul de date al bazei de date terminologice. Cu alte cuvinte, ansamblul "meta-tabelelor" defineste structura care impune conditii si unifica datele terminologice de nivel molecular. Ele pot fi considerate atomi catalizatori ai reactiilor necesare combinarii altor atomi în interactiuni moleculare.

Celelalte 7 entitati se încarca direct prin proceduri de introducere a datelor sau prin import si cuprind datele terminologice vizibile pentru utilizatorul bazei de date. Informatiile continute de aceste entitati pot fi validate la nivel molecular folosind interogari SQL

standard. Majoritatea interogarilor formulate de utilizatorii bazei de date se concentreaza aproape în întregime asupra informatiilor încarcate în aceste entitati.

Elementul central al aplicatiei pentru întretinerea unei astfel de baze de date este componenta de tip *parser* pentru crearea, validarea si prelucrarea documentelor MARTIF în particular (fara a ignora documentele SGML, HTML, XML). În mod obisnuit un *parser* este un modul software care examineaza un document SGML prin confruntarea acestuia cu DTD-ul corespunzator. Rezultatul acestei examinari este de cele mai multe ori simplu: 'da' în situatia în care documentul reprezinta o instantiere valida a DTD-ului si 'nu' în cazul contrar. De cele mai multe ori *parser*-ul este capabil sa 'normalizeze' documentul validat (aducându-l la o 'forma canonica') astfel încât faciliteaza formatarea, editarea si încarcarea documentului în baza de date.

Alaturi de *parser* si legat de acesta se afla un *editor structurat*. Pornind de la DTD acesta propune utilizatorului pas cu pas optiunile de compunere, sau modificare a unui document în conformitate cu definitia tipului corespunzator documentului. În cazul în care obiectivul este compunerea unui document SGML el poate asigura completarea *tag*-urilor necesare.

De cele mai multe ori sistemele de management al bazelor de date orientate spre text folosesc *fisiere inversate de indexare* a continutului acestora pentru regasirea informatiilor. Cautarea poate urmari aparitia unui anume cuvânt, sau a unui model oarecare într-un document sau în o parte a acestuia. Identificarea subdiviziunilor documentului se poate face folosind tocmai tag-urile cu acesta este marcat, respectiv modul în care acestea au fost transcrise în relatiile dintre tabelele bazei de date.

În fine, o componenta deosebit de importanta este aceea care realizeaza functiile de *import-export* ale datelor terminologice spre si dinspre baza de date.

Terminologia calitatii

Realizarea unor resurse terminologice multilingve este de mai multa vreme în centrul preocuparilor Asociatiei Române pentru Terminologie (TERMROM). Începând de anul trecut pe lista temelor având aceeasi orientare se înscrie proiectul "Terminologie armonizata cu prevederile EURODICAUTOM
în domeniul calitate si standardizare". Proiectul a fost initiat de Ministerul Educatiei si Cercetarii si este finantat în cadrul Programului CALIST.

Obiectivele principale ale acestui subprogram sunt:

- Asigurarea flexibilitatii necesare pentru a raspunde operativ la cerintele concrete de rezolvare a unor teme de cercetare care decurg din prioritatile stabilite prin strategiile guvernamentale adoptate pe domenii specifice, în procesul integrarii României în U.E.

- Asigurarea conditiilor de dezvoltare si armonizare a sistemului de standarde nationale în conformitate cu cerintele organismelor de standardizare europene si internationale;
- Asigurarea unei baze terminologice stiintifice pentru elaborarea standardelor de calitate românesti, precum si în ceea ce priveste conditiile de aplicabilitate a prevederilor standardelor internationale si europene adaptate ca standarde românesti;
- Clarificarea conditiilor pe care trebuie sa le îndeplineasca produsele românesti în vederea patrunderii lor pe piata unica a Uniunii Europene si produsele introduse în România.

Pentru realizarea obiectivelor proiectului au fost prevazute urmatoarele activitati:

- Întocmirea unui Proiect Terminologic pentru definirea si înregistrarea terminologiei domeniilor calitate si standardizare utilizate în documentele oficiale ale Uniunii Europene, conform prevederilor EURODICAUTOM si standardelor internationale;
- Extragerea, traducerea si structurarea terminologiei domeniilor calitate si standardizare;
- Proiectarea, programarea si implementarea unei Baze de date conform Proiectului Terminologic capabila sa gestioneze toate domeniile EURODICAUTOM;
- Inregistrarea în baza de date a terminologiei domeniilor calitate si standardizare;
- Elaborarea unei aplicatii informatice de administrare a bazei de date terminologice si de transfer de date terminologice conform formatului standard ISO pentru lucrul în retea;
- Realizarea unui site web pentru promovarea Bazei de date terminologice si punerea acesteia la dispozitia publicului.

A fost avizat Proiectul Terminologic, au fost stabilite cerintele pe care sa le satisfaca suportul informatic, s-a constituit un fond de termeni specifici extrasi din EURODICAUTOM si din Tezaurul rational al CEI si au fost demarate activitatile pentru realizarea unei baze de date relationale 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. **ISO 12200:1999**
    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. **ISO 12620:1999**
    Computer applications in terminology - Data categories
14. **ISO 15188:2001**
    Project management guidelines for terminology standardization

SECTIUNEA II

TEHNOLOGII ALE LIMBAJULUI SCRIS

Ro-balkanet - ontologie lexicalizata, în context multilingv, pentru limba româna

Dan TUFIS, Institutul de Inteligenta Artificiala, Academia Româna, Bucuresti, tufis@racai.ro

Dan CRISTEA, Facultatea de Informatica, Universitatea A.I.Cuza, Iasi dcristea@infoiasi.ro

Rezumat

Cerintele crearii unei ontologii multilingve de tipul EuroWordNet sunt frecvent contradictorii si daca problemele de compatibilitate nu sunt considerate în etapele timpurii ale constructiei o armonizare tardiva se poate dovedi dificila sau imposibila. Mai exact, exista doua probleme majore de compatibilitate care trebuie avute în vedere si anume: acoperirea conceptuala – în sensul ca fiecare lexicon monolingv ar trebui sa contina lexicalizari ale aceluiasi fond conceptual si coeziunea interpretativa – în sensul ca interpretarea relatiilor folosite în fiecare din ontologiile cuprinse în ontologia multilingva trebuie sa fie identica. În lucrare sunt discutate ambele aspecte si prezentate solutiile adoptate în vederea satisfacerii criteriilor de consistenta si coerenta multilinguala a wordnet-ului pentru limba româna.

1. Limba, resurse lingvistice si comunicare electronica

Cercetarea în domeniul tehnologiilor limbajului este un domeniu ce are deja istorie în stiinta calculatoarelor, dar, actualmente, motivatiile sale depasesc sfera interesului pur stiintific sau comercial. Pastrarea identitatii limbilor si culturilor nationale în cadrul globalizant al societatii informationale si a cunoasterii readuce în actualitate avertismentul lui Alain Danzin (1992): ***„În era electronica, este esential pentru supravietuirea unei limbi ca ea sa fie folosita în sistemele de informare electronica."*** Avansul stiintific si tehnologic obtinut în cei 10 ani scursi de la raportul prezentat de Danzin Comisiei Europene, a condus la maturizarea unor teorii, tehnologii, metode si la dezvoltarea altora noi, dar mai ales a permis definirea unor standarde pentru realizarea unitara a ceea ce generic se numeste *resurse lingvistice fundamentale* ale unei limbi. Caracterul multilingual al societatii cunoasterii, în care conceptul de „unitate prin diversitate" se refera în primul rând la prezervarea limbilor si culturilor actuale, a generat o deosebita efervescenta,

puternic stimulata de organismele internationale – în primul rând Comisia Europeana – asupra cercetarii în domeniul resurselor multilingve. Metodologic, tehnologia limbajului natural creeaza o distinctie neta între prelucrari si date, între „masinaria software de prelucrare a limbajului" numita si *lingware* si cunostintele lingvistice, numite cum aratam *resurse lingvistice*, necesare functionarii acestei masinarii. Dihotomia *lingware - resurse lingvistice*, sustinuta de standardele de reprezentare si codificare a cunostintelor lingvistice permite dezvoltarea independenta a celor doua componente ale unui sistem de prelucrare a limbajului. Lingware-ul este independent de limba si intra tot mai pregnant în zona ingineriei software. El poate fi dezvoltat de specialisti de oriunde fara ca acestia sa fie preocupati de limba pentru care va fi folosit. Resursele lingvistice însa sunt de competenta specialistilor vorbitori nativi ai limbii respective. În conditiile în care aceste resurse lingvistice sunt realizate în conformitate cu standardele sau practicile internationale, ele pot fi integrate în sistemele de comunicare electronica, nu doar pentru prelucrare monolingva ci mai ales pentru prelucrari multilingve. Beneficiile alinierii la standardele internationale în realizarea resurselor lingvistice sunt enorme, si putem considera un exemplu foarte simplu. Sa presupunem ca suntem interesati de un anumit subiect si, folosind imensul ocean informational ce este Internet-ul, apelam la un asa numit „motor de cautare", un program a carui functionalitate asigura identificarea documentelor electronice ce contin informatii potential relevante pentru subiectul nostru de interes. Acest gen de serviciu informational este asigurat de „motoare de cautare" precum Google, Altavista, Excite si multe altele. Documentele interesante din punctul nostru de vedere ar putea sa fie scrise în limba engleza, franceza, germana, româna sau orice alta limba. Dar pentru a le regasi pe toate, indiferent în ce limba am formulat cererea noastra de regasire, motorului general de cautare îi sunt necesare resursele lingvistice specifice limbilor în care documentele ar putea exista. Daca aceste resurse lingvistice exista pentru engleza, franceza, germana, italiana etc. si ele sunt reprezentate în acelasi format standardizat, rezultatul cercetarii noastre documentare va fi o colectie de documente tratând subiectul de interes în oricare dintre aceste limbi. Un astfel de serviciu, numit regasire documentara multilingva este o realitate pentru toate limbile „mari", o calificare ce nu are acoperire în substratul cultural ci doar în ceea ce se numeste „nivelul de informatizare al limbii". Procesul de informatizare a unei limbi naturale permite potentarea si diseminarea ei prin mijloacele tehnologice ale societatii informationale.

2. Lexicalizarea abordarilor în tehnologia limbajului si conceptul „wordnet"

Lexicul este fara îndoiala cea mai importanta resursa lingvistica a unei limbi. Marea majoritate a cercetarii actuale, atât în lingvistica formala cât mai ales în tehnologia limbajului, plaseaza componenta lexicala în centrul modelelor de limba, sub influenta a ceea ce a fost numita abordarea *lexicalizata* sau *lexicalista* a studiului limbii. Nu este de mirare, deci, enormul interes pentru dezvoltarea de resurse lexicale multilingve. Studiul computational al dictionarelor electronice, natura informatiei ce trebuie inclusa în ele si tipul de prelucrari pe care le poate facilita o anumita structurare a unui mare volum lexical

a fost, fara îndoiala, fundamental influentat de proiectul WordNet, lansat în urma cu mai mult de 25 de ani la Universitatea din Princeton sub conducerea reputatului psiholingvist George Miller. WordNet, resursa publica, este o uriasa retea semantica lexicala în care peste 100.000 de *întelesuri* lexicalizate în limba engleza prin mai mult de 130.000 de cuvinte sunt asociate între ele prin relatii semantice si/sau lexicale (Fellbaum, 1998). Fondul lexical este distribuit în 4 retele semantice corespunzând categoriilor gramaticale deschise: substantive, verbe, adjective si adverbe. Notiunea de *înteles* (*meaning*) este în WordNet echivalata cu cea de concept si este reprezentata printr-o serie sinonimica în care fiecare cuvânt al seriei are asociat un numar ce identifica sensul în care cuvântul respectiv are întelesul asociat conceptului. Seria sinonimica ce identifica un înteles se numeste *sinset*. Relatiile existente între sinseturi sunt de diferite tipuri, depinzând de categoria gramaticala a cuvintelor ce alcatuiesc un anumit sinset (antonimie/sinonimie, hiponimie/hiperonimie, holonimie/meronimie, troponimie etc.). Influenta proiectului WordNet a fost enorma în domeniul tehnologiei limbajului (exprimata poate si prin faptul ca acum, în limbajul tehnic cel putin, cuvintele „wordnet" si „synset" au devenit substantive comune, importate prin calchiere în mai toate limbile) iar beneficiile acestui concept sunt atât de evidente încât Comisia Europeana, între 1996 si 1998, a finantat un proiect similar de mare anvergura numit EuroWordNet (Bloksma et al., 1996). Acest proiect, extrem de ambitios si-a propus nu numai realizarea concertata de wordneturi monolingve pentru limbile europene de circulatie internationala (engleza, franceza, germana, italiana, olandeza, spaniola) dar a introdus o cerinta fundamental noua, anume corelarea multilinguala a celor 6 retele semantice lexicale, astfel încât dintr-un sinset al unei limbi sa se poata ajunge în echivalentul de traducere al oricaror celorlalte 5 limbi. Fata de relatiile originale din WordNet, EuroWordNet propune un inventar mult mai bogat (90) de relatii cum ar fi cele tematice de tip cazual (Agent, Patient, Instrument, Location, Direction) sau cele corelând sensurile derivatilor lexicali (XPOS-SYNONYMY: a adora - adoratie).

Solutia tehnica pentru corelarea multilinguala a retelelor semantice monolingve a fost definirea unui index interlingual (ILI), independent de limba, continând reprezentari conceptuale ale întelesurilor lexicalizabile în limbile proiectului. Fiecare înteles din oricare din limbile reprezentate în reteaua semantica multilingva este pus în corespondenta, în general, cu un singur concept al indexului interlingual. Aceste corespondente se realizeaza prin intermediul a 20 de tipuri distincte de relatii binare. Sinseturile (seriile sinonimice) din doua sau mai multe limbi care sunt puse în corespondenta cu acelasi concept din ILI sunt considerate echivalenti de traducere, natura echivalentei de traducere fiind definita de tipul relatiilor ce definesc corespondenta dintre sinseturile respective si conceptul comun.

Initial, indexul multilingual a fost constituit ca o multime nestructurata a tuturor întelesurilor lexicalizate în WordNet (cu alte cuvinte în engleza). Ulterior, prin dezvoltarea wordneturilor monolingve, ILI a fost îmbogatit si cu reprezentari conceptuale cu lexicalizare ce nu se regasesc în engleza.

O alta inovatie a proiectului EuroWordNet a fost adoptarea unei multimi de primitive semantice, independente de limbaj, în termenii carora asa-numitele *concepte de*

baza din ILI au fost asociate cu descrieri *ontologice*. Prin importul acestor descrieri la nivelul lexicalizarilor prin echivalenti de traducere (si, prin mostenire, la hiponimii acestora) în fiecare dintre wordneturile monolingve, în EuroWordNet se poate vorbi de o ontologie lexicala multilingva. O prezentare în detaliu a proiectului EuroWordNet se poate gasi în (Vossen, 1998).

Dupa 3 ani, proiectul EuroWordNet initial a fost extins pentru o perioada de înca doi ani (EuroWordNet II) si a încorporat înca 4 limbi: basca, catalana, ceha si estoniana. Proiectul EuroWordNet II s-a încheiat în anul 2000 cu realizarea unor nuclee a caror extensie a ramas în exercitiul financiar al autoritatilor nationale.

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

In septembrie 2001 a fost lansat proiectul european BALKANET (IST – 2000 – 29388), o continuare fireasca a proiectului EuroWordNet II care aduce alaturi de cele 10 limbi europene alte 5 limbi din zona balcanica: bulgara, greaca, româna, sârbo-croata, turca (Stamou et al., 2002). Ca si în EuroWordNet, ontologiile lexicale monolingve sunt corelate printr-o multime de concepte interlinguale, corespondentele fiind stabilite cu ajutorul unor relatii de echivalenta complexe (*eq-synonymy, eq-near-synonymy, eq-has-hyperonym, eq-has-hypernym, etc.).*

Reprezentantii din România în acest proiect, care va dura trei ani, sunt Institutul Academiei Române de Cercetari pentru Inteligenta Artificiala din Bucuresti (coordonator Dan Tufis) si Facultatea de Informatica a Universitatii A.I.Cuza din Iasi (coordonator Dan Cristea) si în realizarea obiectivelor proiectului sunt implicati numerosi specialisti, atât informaticieni cât si lingvisti. Desigur, participarea româneasca în acest proiect si angajarea fata de obiectivele proiectului nu s-a bazat numai pe entuziasm ci pe activitati si rezultate anterioare importante, pe *surse lingvistice* primare (Tufis, 2001) de referinta ale limbii române, implementate ca *resurse lingvistice* (ibid.) în format standardizat si pe o multitudine de programe de prelucrare dezvoltate de-a lungul a multi ani de cercetare, în cea mai mare parte prin finantare internationala.

3.1. Corpusuri

În cadrul proiectelor europene Multext-East si TELRI (Erjavec et al.,1997), (Dimitrova et al., 1998), (Tufis, Bruda, 1997), (Tufis et al., 1997, 1999) a fost creat un corpus paralel în 7 limbi, foarte detaliat adnotat, bazat pe romanul "1984" al lui Orwell si un alt corpus paralel în 25 de limbi, bazat pe "Republica" lui Platon. Adnotarea folosita initial a fost conforma cu standardul TEI (http://www.tei-c.org/),dar ulterior, odata cu cristalizarea standardului CES (Ide, 1998), corpusurile au fost re-adnotate (automat) în conformitate cu CES. Acestea sunt doua corpusuri relativ mici (câte aproximativ 110.000 cuvinte în fiecare limba) dar, datorita acuratetei proceselor de etichetare si de aliniere

(validate manual), au fost extrem de folositoare pentru diverse aplicatii, de la construirea modelelor lingvistice pentru etichetare morfo-sintactica (Tufis, 1999), clasificare a documentelor (Tufis et al., 2000), extragere de echivalenti de traducere (Tufis, 2002), pâna la discriminarea automata a sensurilor (Ide et al., 2002). Pe lânga corpusurile multilingve s-au construit alte doua corpusuri monolingve mult mai mari: un corpus literar bazat pe diverse romane (continând aproximativ 1.500.000 cuvinte) si un corpus jurnalistic (continând peste 100.000.000 cuvinte). Ambele corpusuri au fost segmentate, etichetate si lematizate automat[74].

3.2. Dictionare explicative: WEB-LEX si XML-LEX

Principalul dictionar pe care l-am folosit în analiza noastra este Dictionarul Explicativ al Limbii Române (DEX, 1996), referinta lexicografica pentru limba româna contemporana, dictionar realizat de Institutul de Lingvistica „Iorgu Iordan"[75] al Academiei Române. În urma analizelor statistice de frecventa în corpusurile mentionate, au fost selectate si introduse în format electronic cele mai frecvente 23.000 de cuvinte titlu din DEX. Acest nucleu DEX a fost convertit într-o baza de date lexicala în cadrul proiectului european CONCEDE (*CONortium for Central European Dictionary Encoding*) (Tufis et al., 1999) si al proiectului prioritar al Academiei WEB-LEX (Tufis, 2000). Ulterior, îmbogatit continuu prin culegere manuala din alte câteva dictionare explicative (DEX'84, DOOM, DLRM), la initiativa unor tineri entuziasti atât din tara cât si din diaspora (vezi de pilda: http://dex.francu.com), WEB-LEX a fost corectat sub aspect sintactic-structural si codificat într-un format standardizat, respectând conventiile lexicografice utilizate de DEX si, în masura posibilului, continutul sau textual. Uneori, din considerente legate de consistenta structurala, s-au operat o serie de modificari asupra continutului. De asemenea, o serie de erori evidente în sursa primara au fost corectate de specialisti avizati. Desi mai bogat (în prezent WEB-LEX contine aproape 70.000 de intrari, fata de cele circa 56.000 de intrari din DEX'96), influenta DEX a fost fundamentala în dezvoltarea WEB-LEX. Pe de alta parte, eventualele critici asupra continutului, acolo unde ne-am despartit de DEX, în nici un caz nu trebuie puse în seama Institutul de Lingvistica „Iorgu Iordan-Al. Rosetti" ci a noastra. Din acest motiv, preferam sa ne referim la WEB-LEX ca la un dictionar ***de tip*** DEX si nu ca varianta computationala a DEX-ului.

Codificarea continutului WEB-LEX, s-a realizat folosind limbajul de adnotare XML. Implementarea, ce expliciteaza toate conventiile tipografice precum si informatiile implicite, a condus la un volum textual de date de circa 8-10 ori mai mare fata de continutul textual echivalent al DEX-ului. Adnotarea XML a fost realizata automat, cu ajutorul compilatorului `DIC` (Tufis, 2000). Compilatorul a fost generat automat folosind JavaCC©, pe baza unei gramatici LL(7) ce descrie structura formala a intrarilor în DEX. `DIC` poate fi folosit pentru a genera documente XML (conform cu DTD-ul CONCEDE) pentru orice

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

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

dictionar ce foloseste conventiile tipografice adoptate în DEX. În (Vintila-Radulescu, 2002) sunt prezentate o multitudine de dictionare realizate sau aflate în curs de realizare la Institutul de Lingvistica „Iorgu Iordan-Al. Rosetti" si presupunând ca ele urmaresc conventiile tipografice si lexicografice adoptate în DEX, toate aceste surse lingvistice de referinta pentru limba româna ar putea fi transformate, cu efort minim, în resurse computationale fundamentale pentru prelucrarea automata.

Varianta codificata a dictionarului nostru este numita XML-LEX iar structura sa este descrisa 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
   xml:space  (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       (%basetags; | 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     >
<!--  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|acc|lang|gram|obj|
          subj|verb|hint|geo|dom|register|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)       "u"
    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

Aceasta structura de codificare a fost adoptata în implementarea unui numar mare de dictionare, reprezentând un standard „de facto" în lexicografia computationala actuala (Erjavec *et al*., 2000). Detalii suplimentare privind semantica entitatilor folosite în codificare si a atributelor acestora, pot fi gasite în documentatia tehnica a proiectului la adresa www.itri.bton.ac.uk/projects/ concede/.În tabelul de mai jos, sunt exemplificate reprezentarea tipografica (de tip DEX) si reprezentarea codificata în XML.

DEX	XML-LEX
ZA[2], zale, s.f. 1. Fiecare dintre ochiurile unui lanț; p. gener. (la pl.) lanț. ♦ Lănțișor de metal întrebuințat uneori ca podoabă. ♦ Cusătură în formă de lănțișor, 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 oștenii în antichitate şi în evul mediu, spre a se apăra de loviturile dușmanilor. [Var.: (reg.) zálă,zea s.f.] - Cf. ngr. záva.	<entry type="homonym" id="ZA.2"> <hw>ZA</hw> <alt> <brack> <gram>nominativ_feminin_singular_indefinit</gram> <orth>ZA</orth> </brack> <brack> <gram>nominativ_feminin_plural_indefinit</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>Lănțișor de metal întrebuințat uneori ca podoabă.</def> </struc> <struc type="Sec"> <def>Cusătură în formă de lănțișor, 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 oștenii în antichitate şi în evul mediu, spre a se apăra de loviturile dușmanilor.</def> </brack> </alt>

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

Figura 2: Continut primar si codificarea echivalenta
în XML (cf. CONCEDE.dtd)

În tabelul din Figura 2, sunt exemplificate reprezentarea tipografica (de tip DEX) si reprezentarea codificata în XML. Mentionam ca reprezentarea tipografica din coloana stânga a Figurii 2 s-a obtinut automat, folosind un convertor XML de format, proiectat astfel încât rezultatul generarii (interpretarea marcajului XML) sa fie cât mai apropiat de aspectul dictionarului tiparit. Structura de dictionar, definita mai jos, este suficient de generala pentru a permite implementarea diferitelor tipuri de dictionare. În fapt, DTD-ul CONCEDE a fost utilizat pentru codificarea a doua dictionare bilingve: un dictionar Sloven-Englez si un dictionar Român-Francez.

Adnotarea XML fiind independenta atât de conventiile tipografice cât si de limba dictionarului, este posibila cautarea multi-criteriala a informatiei în unul, doua sau mai multe dictionare explicative ale unor limbi diferite. De pilda, o cautare multi-criteriala ar putea fi parafrazata astfel:

Gaseste si afiseaza toate intrarile ce corespund ***substantivelor feminine, de origine neo-greaca*** *si al caror cuvinte titlu* ***încep cu secventa de litere ZA***.O astfel de cautare va avea ca rezultat tiparirea cel putin a intrarii corespunzatoare cuvântului titlu ZA^2**:**

ZA^2, zale, s.f. **1.** Fiecare dintre ochiurile unui lant; *p. gener.* (la pl.) lant. ♦Lantisor de metal întrebuintat uneori ca podoaba. ♦ Cusatura în forma de lantisor, executata de obicei la broderii. **2.** (La pl.) Împletitura executata din inele mici de fier legate unul de altul; *p. ext.* armura facuta din aceasta împletitura, cu care se îmbracau ostenii în

antichitate si în evul mediu, spre a se apara de loviturile dusmanilor. [Var.: (reg.) **zála**,**zea** s.f.] - Cf. ngr. záva.

3.3. Alte dictionare, lexicoane; indexul interlingual

Unul dintre rezultatele proiectului Multext-East îl constituie un lexicon de forme ocurenta (LFO), cu peste 450.000 de intrari, care contine triplete de tipul <cuvânt, lema, cod_morfo-sintactic>. Acest lexicon va fi completat cu formele flexionare (generate automat) a lemelor din XML-LEX nereprezentate în LFO. Codificarea folosita este compatibila cu recomandarile Eagles (http://www.ilc.pi. cnr.it/EAGLES/home.html) pentru adnotarea morfo-sintactica si este documentata pe larg în (Tufis *et al.*, 1997).

O alta resursa lexicala esentiala a fost Dictionarul de Sinonime al Limbii Române – DSLR (Seche, Seche, 1997), care a fost transpus în forma electronica la Facultatea de Informatica a Universitatii "A.I.Cuza" din Iasi. Forma electronica a DSLR a fost convertita în format XML astfel încât aceeasi interfata ce a fost dezvoltata pentru XML-LEX functioneaza si cu XML-DSLR.

Din corpusurile paralele mentionate mai sus si folosind programul ce implementeaza metodologia noastra de extragere a echivalentilor de traducere (Tufis, Barbu, 2001a, 2001b, 2002) s-a construit un dictionar bilingv Român – Englez (de asemenea transpus în format XML). Acest lexicon bilingv a fost validat manual si îmbogatit cu noi intrari din diverse surse publice.

În sfârşit, o resursă extrem de valoroasă a fost şi Indexul Interlingual al EuroWordNet, exportat în format XML cu editorul VisDic produs la Universitatea Masaryk din Brno (Pavelek, Pala, 2002).

3.4. Alegerea nucleului lexical

Vom da câteva definitii ale unor notiuni pe care le vom folosi în cele ce urmeaza.

Când ne plasam într-un context monolingv, vorbim despre *sensuri, întelesuri* si *sinseturi*. Un cuvânt are unul sau mai multe *sensuri*. Un sens refera un *înteles*. În EuroWordNet sensurile unui cuvânt sunt numerotate în functie de frecventa lor, iar sensul unei leme este denotat adaugând numarul sensului la forma ortografica a acesteia. O multime de sensuri astfel specificate (ex. *action2, activity1, activiteness1*) care refera acelasi înteles este numit *sinset* si constituie el însusi denotatia întelesului sensurilor din sinset. Cu alte cuvinte, un sinset reprezinta *lexicalizarea unui înteles* în contextul monolingv curent.

Daca abstractizam notiunea de *înteles*, definita ca mai sus, astfel încât sa nu mai facem referirea la un anumit context monolingv, vom vorbi despre *concepte* care sunt referite de *întelesurile* lexicalizate în diferitele limbi. Asadar, putem vorbi despre concepte care au sau nu realizare lingvistica într-o limba sau alta. Un concept este un construct cognitiv, independent de limba, care în EuroWordNet este totdeauna lexicalizat cel putin într-una dintre limbi. Un concept este mai departe rafinat în termeni de distinctii semantice

elementare (trasaturi semantice), deci putem vorbi despre gruparea conceptelor în functie de trasaturile lor semantice.

În EuroWordNet si deci si în BALKANET, ILI este definit ca o colectie nestructurata de intrari de forma: <ILI–index><descriere ontologica><glosa> {domeniu}. Indexul interlingual initial a fost construit plecând de la versiunea 1.5 a Wordnet-ului si deci glosele pentru fiecare concept au fost importate direct din sinsetul englezesc care se refera la întelesul conceptualizat în ILI.

Pentru a facilita o cât mai buna intercorelare a wordneturilor monolingve din cadrul proiectului si pentru a înlesni extensia lor ulterioara, consortiul proiectului a decis ca procesul implementarilor paralele sa fie centrat pe concepte (independente de limba) selectate de comun acord, la momente succesive de timp.

O prima selectie a constituit-o multimea asa-numitelor „concepte de baza" definite în EuroWordNet ca fiind acele concepte din ILI lexicalizate în limba engleza (în WORDNET) prin sinseturi plasate pe un nivel ierarhic cât mai sus si, în plus, care au un numar mare de hiponimi directi (tot în WORDNET). Ratiunea acestei decizii a constat în faptul ca, aceste concepte fiind foarte generale si totodata productive în definirea unor concepte mai particulare, este foarte probabil ca ele sa fie lexicalizate în majoritatea limbilor de interes. Acest lucru a fost probat atât în EuroWordNet cât si în BALKANET. Multimea c*onceptelor de baza* (o motivatie mai detaliata a selectiei lor este prezentata in (Vossen, 1998) în raport cu obiectivele EuroWodNet) contine 1.310 concepte, fiecaruia dintre ele fiindu-i atasata o glosa explicativa si o *descriere ontologica* (vezi Rodriguez et al., 1998).

Dupa implementarea, în toate cele 5 limbi ale proiectului, a nucleelor de ontologii lexicale corespunzând conceptelor de baza, s-a facut o noua selectie, de data aceasta, continând 4.000 de noi concepte interlinguale.

Selectia a avut în vedere, pe de o parte maximizarea compatibilitatii cu EuroWordNet, iar pe de alta parte relevanta stocului lexical pentru fiecare limba din perspectiva monolingva. Primul criteriu a fost operationalizat alegându-se acele concepte lexicalizate în cele mai multe limbi din EuroWordNet. Limita inferioara a numarului de limbi a fost fixata la 5, astfel încât dupa implementarea acestor concepte în BALKANET ele sa fie lexicalizate în cel putin 10 limbi.

Criteriul relevantei monolingve a condus la propunerea mai multor multimi candidate de concepte. Pentru fiecare limba a proiectului au fost efectuate analize cantitative în context strict monolingv. Metodele de analiza au diferit de la partener la partener, în raport cu datele si instrumentele disponibile pentru limbile în cauza. Dupa analiza acestor multimi, au fost incluse în multimea finala acele concepte ce au aparut în cel putin doua propuneri. Multimea finala a conceptelor a fost ordonata dupa numarul de limbi din EuroWordNet ce le lexicalizeaza si dupa numarul de limbi din BALKANET care le-au propus. Primele 4000 de noi concepte în aceasta lista au fost de comun acord alese ca tinta comuna pentru cea de a doua etapa a proiectului.

În continuare prezentam metodologia folosita pentru limba româna privind selectia fondului lexical în cadrul BALKANET. Analiza cantitativa s-a efectuat asupra unui corpus foarte mare, format din mai multe romane si dintr-o colectie de texte jurnalistice culese de pe web. Corpusul (continând mai mult de 100 de milioane de cuvinte) a fost supus unor prelucrari statistice, fiind etichetat si lematizat automat, iar cuvintele care prezentau interes (substantive comune, verbe, adjective si adverbe) au fost sortate în functie de frecventa lor în texte. Am extras în acest fel o lista de mai mult de 30.000 de leme. În functie de frecventa acestora în textele din corpus, aceasta lista a fost împartita în trei parti, corespunzând celor mai frecvente 10.000 de leme (I), urmatoarele cele mai frecvente 10.000 (II) si restul (III). Frecventa dintr-un corpus este considerata de multi lexicografi un criteriu subiectiv. Printre cele mai puternice argumente se numara volumul si reprezentativitatea textelor incluse în corpusul folosit la analiza cantitativa. Luând în calcul faptul ca din ce în ce mai multe texte sunt disponibile pe web, marimea corpusului nu mai reprezinta o problema semnificativa, însa reprezentativitatea ramâne în continuare un punct slab. Definirea exacta a naturii textelor care trebuie incluse într-o analiza cantitativa face obiectul unei îndelungi polemici si nu vom insista asupra ei. Având în vedere ca datele noastre constau aproape în întregime din texte jurnalistice, problema reprezentativitatii poate fi cu îndreptatire ridicata. Dictionarul de Frecvente al Cuvintelor Românesti FDRW (Juilland *et al.*, 1965), publicat cu mult timp în urma, bazat pe un *corpus balansat* de 500.000 de cuvinte (teatru, nuvele si scurte povestiri, eseuri memorii si corespondente, texte jurnalistice, literatura tehnica) contine cele mai frecvente 5.000 de leme. Chiar daca este foarte controversat, FDLW este înca folosit de multi lingvisti români ca o referinta. Comparatia pe care am facut-o a aratat ca mai toate cele 5.000 inventariate de FDRW se gasesc si în lista obtinuta de noi, chiar daca nu cu aceleasi scoruri de frecventa. Pe lânga frecventa în corpus am apelat si la alte doua criterii mai putin controversate si care au putut fi operationalizate în raport cu resursele lingvistice disponibile si instrumentele noastre de analiza a corpusurilor. Primul este numarul de sensuri pe care un cuvânt (împreuna cu sintagmele si expresiile în care participa) îl are într-un dictionar. Al doilea este numarul de definitii de dictionar în care apare un anumit cuvânt. Al treilea criteriu, ne-inclus înca în analiza, ar putea fi numarul de derivate lexicale ale unui cuvânt. Pentru o pertinenta analiza din acest punct de vedere, o excelenta lucrare este (Dinu, 1996).

În aceasta faza a proiectului BALKANET, ne-am concentrat atentia asupra substantivelor din limba româna, iar datele experimentale raportate mai jos se refera doar la acestea. Având însa în vedere ca procedurile tehnice nu depind de categoria gramaticala, metodologia si procedura vor fi aceleasi si pentru verbe, adjective si adverbe. Luând în calcul numai primele doua clase de frecventa descrise mai sus (primele 20.000 cele mai frecvente din corpusul jurnalistic) am extras din XML-LEX mai mult de 8.000 de intrari de substantive si substantive compuse (care însumeaza aproximativ 35.000 de sensuri) astfel încât productivitatea definitionala PD (numarul de definitii în care participa un substantiv) sa fie cel putin 3. Lista a fost sortata în functie de productivitatea definitionala si numarul de sensuri ale fiecarui cuvânt titlu.

Substantiv	Productivitate definiționala	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 candidatilor

Pentru toate aceste substantive am extras traduceri englezesti din dictionarul de echivalenti de traducere. Procedurile pentru extragerea automata a echivalentilor de traducere din corpusuri paralele ca si procedura de discriminare a sensurilor sunt descrise pe larg în (Tufis, Barbu, 2001a,b), (Erjavec et al. 2001), (Tufis, 2002), (Ide *et al.*, 2002). Fiecare substantiv din limba româna a fost pus în corespondenta cu lista tuturor conceptelor din ILI corespunzatoare traducerilor sale în engleza. Conceptele astfel identificate, au fost sortate dupa rangul corelat al substantivelor românesti de la care s-a pornit.

Interesant de remarcat ca dintre cele 4000 de concepte selectate în final prin armonizarea propunerilor tuturor partenerilor, circa 2600 s-au regasit si în primele 4000 de concepte ale ierarhiei noastre. Toate cele 4000 de concepte selectate de consortiu se regasesc printre primele 6000 de concepte ale ierarhiei noastre.

Toate substantivele reprezentând potentiale lexicalizari ale celor 4000 de concepte din cea de a doua selectie au fost automat puse în corespondenta cu toate definitiile lor din XML-LEX. De asemenea, ele au fost corelate cu lexicalizarile din limba engleza ale celor 4.000 de concepte. Prin intermediul dictionarului de echivalenti de traducere englez-român, fiecare concept a fost asociat cu lexicalizarea din limba engleza (extrasa din WORDNET) si cu potentialele lexicalizari în limba româna.

Dictionarul de Sinonime al Limbii Române (DSLR), digitizat si codificat în XML, a fost folosit pentru a extrage seriile sinonimice pentru cuvintele românesti selectate. În XML-DSLR unii membri ai seriilor sinonimice sunt arhaisme sau regionalisme. Discutiile preliminare au condus catre ideea de a elimina toate cuvintele care fac parte din aceste clase (ne-am bazat pe cerinta de a construi un nucleu lexical de uz general în limba româna contemporana). Totusi, pentru eventualitatea în care aceste cuvinte filtrate (împreuna cu informatiile despre uz) vor fi necesare mai târziu, s-a asigurat recuperabilitatea lor. Seriile sinonimice românesti au fost considerate ca posibile sinseturi si adaugate la asociatiile descrise mai sus.

4. Instrumente software dezvoltate pentru proiectul BALKANET

Materialul lingvistic de baza descris în sectiunea anterioara, a fost asamblat prin intermediul unor programe unitare, astfel încât toata aceasta informatie este disponibila într-o interfata „prietenoasa", prin care lexicograful alege echivalentele corecte de sens dintre cele potentiale. Aceasta interfata este generata si „personalizata" automat în functie de multimea conceptelor interlinguale furnizata ca parametru de intrare unui generator de interfete. Printr-un astfel de model arhitectural, a fost posibil ca sarcina construirii wordnet-ului pentru limba româna sa fie distribuita între membrii celor doua colective românesti participante la proiect si judicios controlata. Pentru fiecare dintre acestia s-a generat o interfata personalizata pentru o submultime distincta de concepte dintre cele agreate de consortiul proiectului. Utilizatorul acestei interfete, pe care generic îl numim în continuare *lexicograf*, va lucra în mod independent de ceilalti, construind, ca urmare a interactiunii, fragmente ale wordnetului pentru limba româna. La un moment dat, lexicograful alege un concept din multimea ce i-a fost repartizata caruia doreste sa-i ataseze un sinset românesc. El are la dispozitie simultan, sinsetul ce lexicalizeaza în limba engleza conceptul respectiv si, pentru fiecare cuvânt englezesc din acest sinset, toate potentialele lui traduceri în limba româna, aceste traduceri având atasate toate definitiile continute în XML-LEX. În plus, fiecare cuvânt românesc are atasate toate seriile sinonimice din XML-DSLR în care el este prezent. Ceea ce trebuie sa decida lexicograful este (vezi figura 4):

a. care este cuvântul românesc a carui definitie este cea mai apropiata de definitia conceptului lexicalizat în limba engleza;
b. care este cea mai buna serie sinonimica a acestui cuvânt;
c. care dintre definitiile atasate cuvintelor dintr-o serie sinonimica este cea mai adecvata pentru a fi aplicabila tuturor cuvintelor din seria respectiva.

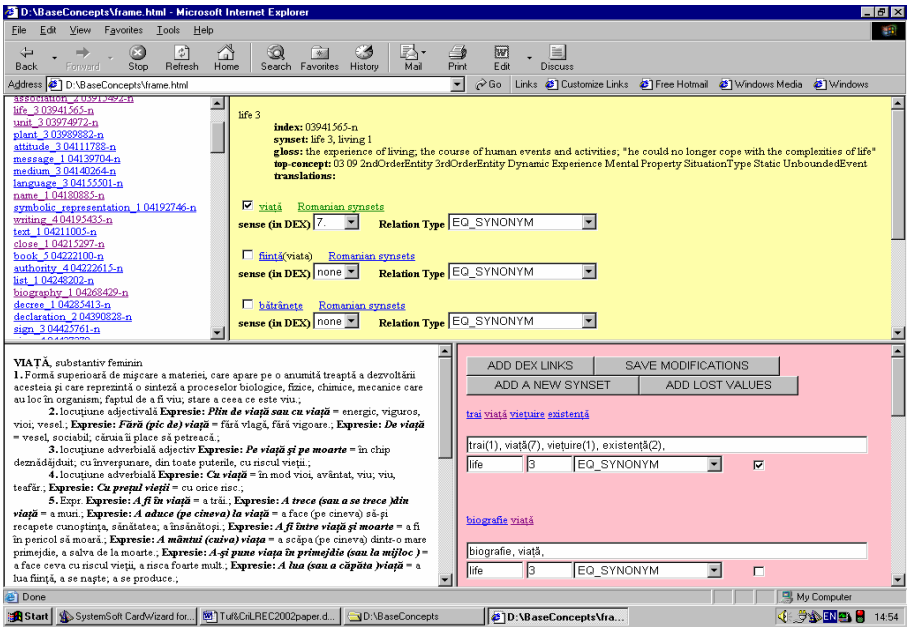


Figura 4: Editorul pentru construirea sinseturilor

În majoritatea cazurilor, definițiile extrase din XML-LEX corespunzând sinonimelor dintr-un sinset nu sunt identice, lexicografii alegând pe cea mai apropiată de definiția conceptului corespunzător (vezi figura 5).

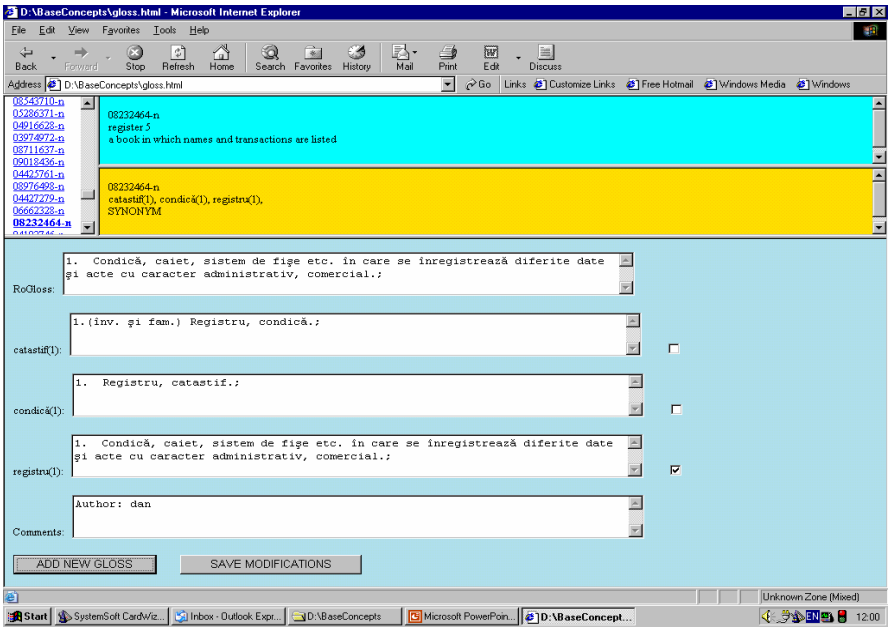


Figura 5: Editorul pentru asignarea gloselor

Merita mentionat ca în faza asocierii gloselor a devenit evidenta incorectitudinea alcatuirii unor sinseturi, ele fiind modificate. În alte cazuri Dictionarul Explicativ al Limbii Române include în aceeasi definitie doua sensuri care sunt demarcate în ILI ca doua concepte diferite. În astfel de situatii strategia generala a fost sa se desparta definitia românească si sa se ataseze ca glosa partea relevanta.

Fragmente create de fiecare lexicograf sunt agregate în mod incremental în structuri din ce în ce mai complexe si mai acoperitoare din punct de vedere lexical. Acest proces de agregare se realizeaza în mod centralizat, astfel încât corectitudinea structurilor rezultate sa poata fi controlata si, în cazul conflictelor, sa se poata identifica si corecta sursele de conflict (de exemplu: acelasi sens pus în corespondenta cu concepte diferite, sensuri diferite ale aceluiasi cuvânt puse în corespondenta cu acelasi concept, literali fara identificatori de sens etc.). Corectarea unor conflicte între doua portiuni ale structurii agregate poate sa genereze conflicte între alte parti ale sale. Pentru evitarea acestui pericol au fost proiectate mecanisme de control centralizat al unificarii subseturilor de wordnet ce gestioneaza efectul global al oricaror modificari locale.

4.1. Importul relatiilor taxonomice; vizualizare sincronizata a mai multor wordneturi

Constructia sinseturilor si punerea lor în corespondenta cu conceptele interlinguale reprezinta doar una din cele doua dimensiuni fundamentale ale procesului de construire a unei retele semantice lexicale pusa în corespondenta cu indexul interlingual, respectiv cea de implementare a nodurilor si echivalarea acestora cu conceptele interlinguale. Cea de a doua dimensiune a procesului constructiei retelei o constituie definirea relatiilor (intralinguale) între nodurile create si echivalate în prima faza. Deosebit de importante sunt relatiile taxonomice care stabilesc o ierarhie de generic-specific între sinseturile unui wordnet.

Stabilirea relatiilor taxonomice între sinseturile wordnetului pentru limba româna s-a facut automat (urmata de validarea umana) în baza principiului „echivalentei ierarhice interlinguale" (Tufis, Cristea, 2002). În esenta, acest principiu afirma ca:

1. daca sinsetul S_{1LA} din limba LA si sinsetul S_{1LB} din limba LB sunt echivalate cu acelasi concept C_1 din ILI si
2. daca sinsetul S_{2LA} din limba LA si sinsetul S_{2LB} din limba LB sunt echivalate cu acelasi concept C_2 din ILI si
3. daca în limba A sinseturile S_{1LA} si S_{2LA} sunt într-o relatie ierarhica H^+ (H^+ denota compunerea de un numar de ori cel putin egal cu 1 a relatiei H, în cazul nostru: has-as-hypernym),atunci:

în limba B sinseturile S_{1LB} si S_{2LB} sunt într-o relatie ierarhica similara H^+ (desi lanturile de relatii H pot fi de lungimi diferite în cele doua limbi).

Principiul expliciteaza necesitatea ca interpretarea relatiilor folosite în ontologia multilingva sa fie similara, asadar defineste **coeziunea interpretativa** a relatiilor ontologice în toate limbile participante la proiect. Acest principiu este reprezentat schematic în figura 6:

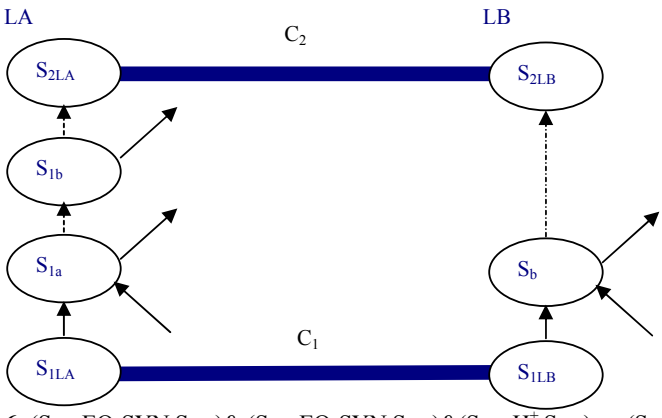


Figura 6: $(S_{1LA}\ \text{EQ-SYN}\ S_{1LB})\&\ (S_{2LA}\ \text{EQ-SYN}\ S_{2LB})\&(S_{1LA}\ H^+\ S_{1LB}) \Rightarrow (S_{2LA}\ H^+\ S_{2LB})$

În sectiunea urmatoare vom arata pe un caz concret cum poate fi exploatat acest principiu pentru a importa (si eventual valida/corecta manual) relatiile dintr-un wordnet în care structurile ierarhice au fost stabilite, într-un wordnet pentru care au fost stabilite doar relatiile de echivalenta translationala cu indexul interlingual (ILI).

Ultima etapa a construirii unui grup de sinseturi este transformarea rezultatelor interactiunii lexicografului cu interfata descrisa anterior într-un format independent de limba (codificare XML) si specific editorului multilingual de ontologii lexicale numit VisDic (Pavelek si Pala, 2002). Odata generat acest format, el poate fi încarcat în VisDic, iar wordnetul pentru limba româna poate fi vizualizat în mod sincron cu toate celelalte wordneturi încarcate. In figura de mai jos este ilustata afisarea în mod sincron a sinsetului românesc *(fiinta_1, forma de viata_1, vietuitoare_1, vietate_1)* si a celui englezesc *(being_1 life form_1, living thing_1, organism_1)* si a arborilor lor de hiponimi. Cele doua sinseturi sunt aliniate via ILI, ambele fiind echivalate independent cu conceptul interlingual cu identificatorul *00002728-n*.

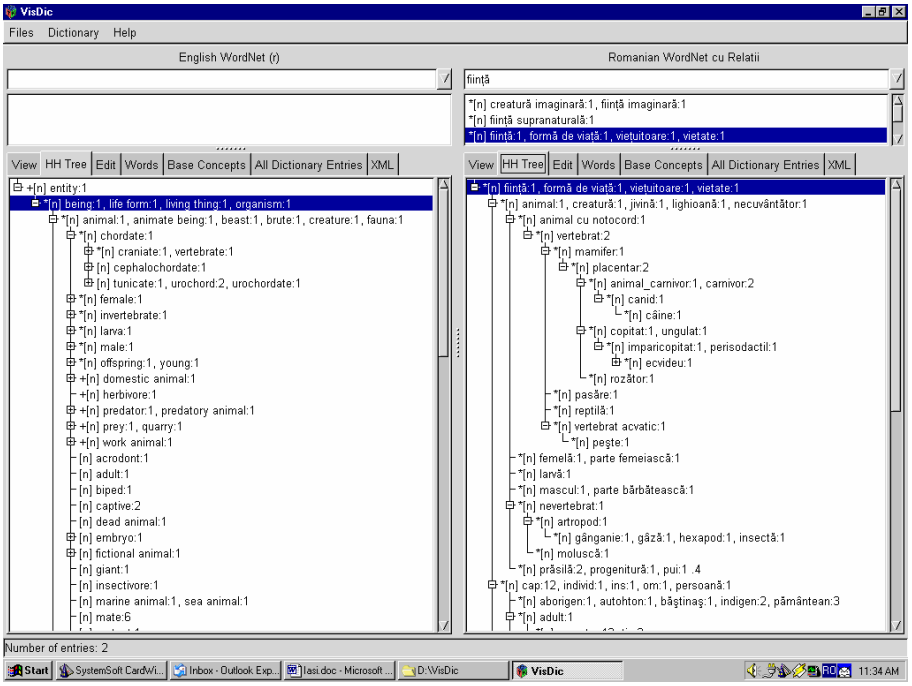


Figura 7: Vizualizarea sincronizata a doua ontologii lexicale cu ajutorul VisDic

Editorul de ontologii multilingve, VisDic, a fost dezvoltat în cadrul proiectului BALKANET pentru a substitei functionalitatea asigurata în cadrul EuroWordNet de

editorul Polaris, dezvoltat de firma Lernout & Hauspie. Implementat initial pentru ca rezultatele proiectului BALKANET sa poata fi utilizate în regim liber de restrictii comerciale (Polaris poate fi utilizat doar contra cost), VisDic este constant îmbunatatit cu facilitati noi a caror necesitare apare pe masura evolutiei proiectului BALKANET, fiind deja unul dintre cele mai puternice instrumente existente pentru gestiunea ontologiilor multilinguale.

5. Principiul conservarii trans-linguale a ierarhiei lexicale. Studiu de caz: Condimente, mirodenii, sosuri si alte ingrediente

Vom considera fragmentele din RO-WordNet si WordNet 1.5 aratate în figura 8. Sagetile reprezinta relatiile de hiponimie (de la hiponime spre hipernime) în cele doua wordneturi. Liniile groase reprezinta relatiile de echivalenta de traducere (EQ-SYN) dintre sinseturile celor doua limbi, aceasta însemnând ca sinseturile respective sunt puse în corespondenta cu acelasi concept din ILI. Linia groasa întrerupta reprezinta o relatie EQ-SYN identificata ca nerespectând principiul conservarii trans-linguale a ierarhiilor lexicale din cele doua wordneturi. Inconsistenta este semnalata deoarece în româna relatiile ierarhice (de hiponimie) dintre *mirodenie*(RO) si *condiment*(RO) ca si dintre *ketchup*(RO) si *sos*(RO) nu sunt verificate de echivalentii lor în limba engleza: *spice*(EN) este frate cu *condiment*(EN) si respectiv *ketchup*(EN) este frate cu *sauce*(EN). Daca structura variantei 1.5 a WordNet este considerata cea corecta, acest exemplu arata ca principiul pastrarii ierarhiei nu este irefutabil. Pe de alta parte, daca ar fi rezonabil sa consideram ca WN 1.5 este amendabil (de exemplu facând *mustard*(EN) si *ketchup*(EN) hiponimii directi ai lui *sauce*(en)) ca în figura 9, atunci principiul pastrarii ierarhiei ar putea fi o puternica proba a consistentei[76].

În urma restructurarilor ierarhice si de echivalare translationala, necesare pentru respectarea principiului conservarii trans-linguale a ierarhiei lexicale (aratate în figura 9), interesant este faptul ca a disparut relatia de echivalenta între cuvântul românesc *condiment* si cuvântul englezesc *condiment*.

[76] *Consultată recent asupra acestei probleme, Christiane Felbaum a confirmat esistența unei erori în ierarhia WN1.5, probată, de altfel, şi de glosa lui* ketchup *(*thick spicy **sauce** made from tomatoes*)*.

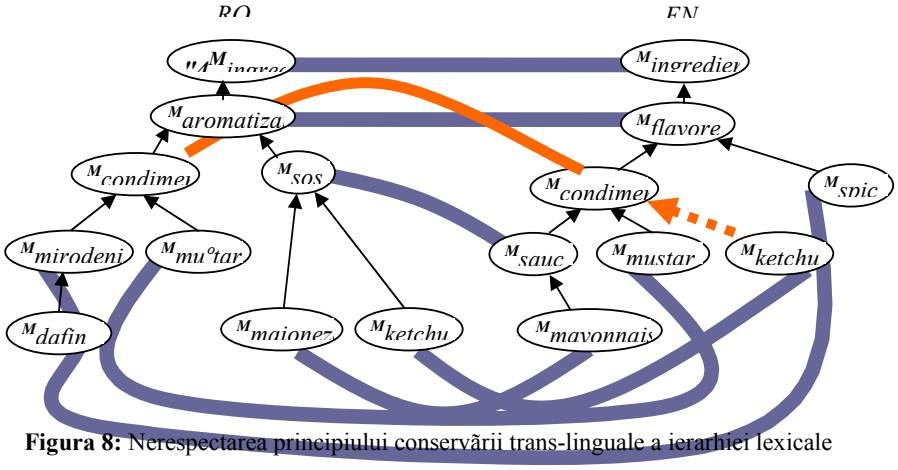


Figura 8: Nerespectarea principiului conservării trans-linguale a ierarhiei lexicale

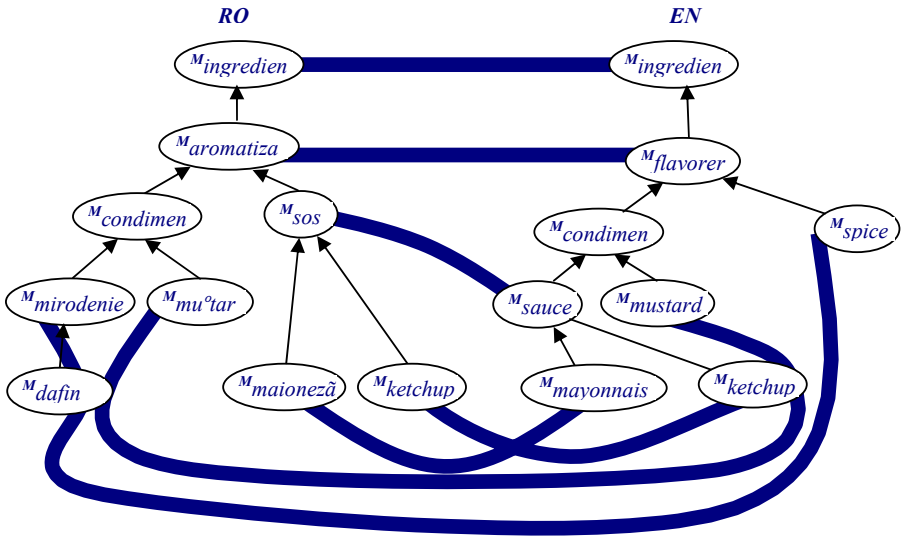


Figura 9: Nerespectarea principiului conservării
trans-linguale a ierarhiei lexicale

Pentru ca aceasta echivalenta sa fie posibila, în conditiile principiului conservarii trans-linguale a ierarhiei lexicale, ar trebui ori ca în limba engleza *spice* **sa fie** un hiponim al lui *condiment* iar *sauce* **sa nu fie** un hiponim al lui *condiment* ci frate, ori în limba româna *sos* **sa fie** un hiponim al lui *condiment* iar *mirodenie* **sa nu fie** un hiponim al lui *condiment* ci frate. Ambele variante au fost respinse de expertii consultati, lexicografi si vorbitori nativi ai limbii engleze si respectiv române. Singura concluzie posibila este ca în româna si engleza cuvântul *condiment* nu reprezinta exact acelasi lucru.

5. Concluzii

Realizarea ontologiei lexicale pentru limba româna, în contextul multilingual definit de proiecte de tipul EuroWordNet, Balkanet si GlobalWordnet (www.globalwordnet.org), este esentiala pentru procesul de informatizare a limbii române. Experienta internationala arata ca un astfel de proiect nu este niciodata închis, reclamând actualizare si întretinere continua, aparând mereu noi idei de îmbunatatire a performantelor si noi cerinte de exploatare. Specialistii de la Princeton au anuntat deja versiunea 1.7.1 a Wordnet, mult îmbunatatita. În variantele ce vor urma, pe lânga extensia în continuare a fondului lexical, toate cuvintele nefunctionale aparând în definitii vor contine referinte spre sinsetul corespunzator contextului de utilizare. Cu alte cuvinte, Wordnet va deveni simultan si un dictionar si un corpus adnotat la nivelul sensului. O alta dezvoltare semnificativa o va reprezenta traducerea definitiilor din Wordnet în formule logice, adecvate prelucrarilor inferentiale. Acest proiect, coordonat de Dan Moldovan si Sanda Harabagiu se afla în derulare la Universitatea Texas din Dallas (Moldovan, 2001), (Harabagiu *et.al.*, 1999).

Astfel de extensii vor trebui considerate în viitor si în wordnetul pentru limba româna aflat deocamdata în faza incipienta. Obiectivul final prevazut 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 putin de un an de la începerea proiectului, wordnetul românesc se afla cu mult înaintea graficului prevazut, având deja create peste 6.000 de sinseturi. Se poate estima ca, în conditii normale, în cei peste doi ani care au mai ramas wordnetul românesc va ajunge la peste 20.000 de sinseturi, acoperind peste 40.000 de literali. Atingerea unui volum lexical similar cu al altor wordneturi necesita însa continuarea proiectului si dupa anul 2004, atragerea unor noi colective de specialisti în aceasta întreprindere si desigur gasirea surselor de finantare, în principal interne, care sa permita dezvoltarea si întretinerea wordnetului românesc. Operationalizarea acestui obiectiv poate fi facilitata de contextul organizatoric creat de curând prin înfiintarea la Academia Româna a Comisiei de Informatizare pentru Limba Româna (CILR) precum si a Consortiului de Informatizare pentru Limba Româna (ConsILR: http://www.consilr.info.uaic.ro/), for executiv al CILR.

A fost construita o platforma software de dezvoltare incrementala a retelei semantice ce permite implementarea independenta de regiuni ale retelei si integrarea ulterioara a acestora. Viabilitatea acestui concept arhitectural si a demersului de dezvoltare distribuita a wordnetului au fost validate prin implicarea în procesul de construire a 10 specialisti, carora li s-au adaugat înca 12 studenti masteranzi de la Facultatea de Litere a Universitatii Bucuresti si Facultatea de Informatica a Universitatii "A.I. Cuza" (cele doua facultati ce au programe de Master în domeniul prelucrarii limbajului natural si al lingvisticii computationale). Rezultatele produse în mod independent au fost agregate fara nici o dificultate. Mediul lingware de dezvoltare contine un modul special de verificare a corectitudinii deciziilor lingvistice la crearea sinseturilor românesti sau la punerea lor în corespondenta cu conceptele indexului interlingual. Dupa cum era de asteptat, procesul de integrare a rezultatelor partiale furnizate de fiecare membru al celor doua echipe de realizare a evidentiat o serie de inconsistente cu explicatii diverse:

- neatentie în asignarea sensurilor, generata de oboseala expertului decident uman;
- granularitate semantica diferita între sensurile explicitate în XML-LEX si sensurile conceptelor din ILI;
- absenta lexicalizarii în limba româna a unor concepte existente in ILI si introducerea unor forme perifrastice cu definitii ad-hoc;
- erori sau incompletitudini existente în sursele lingvistice primare folosite în implementare.

Inconsistentele depistate, atât de natura structurala, dar mai ales cele de natura semantica au fost înregistrate, analizate si unele dintre ele corectate. Altele, necesita o analiza mai profunda si rezolvarea lor a fost amânata pentru o etapa ulterioara a proiectului. Aceasta cu atât mai mult cu cât, prin analiza similara pe care am efectuat-o asupra wordneturilor pentru celelalte limbi din proiect, am constatat ca exista multe similaritati ale acestor genuri de inconsistente. Sunt puse astfel în evidenta o serie de concepte din ILI pentru care diferenta semantica dintre ele este prea mica pentru a fi sesizata usor chiar si de catre un vorbitor nativ al limbii respective. Distinctii atât de rafinate au, din perspectiva prelucrarii automate si mai ales al traducerii automate, o utilitate limitata iar în context multilingv pot fi chiar surse de eroare. Pericolul micsorarii distantei semantice (am putea numi acest fenomen pulverizarea conceptuala) între conceptele din ILI este amplificat de adaugarea unor concepte ce au lexicalizari într-o singura limba sau într-un numar mic de limbi. O solutie pentru evitarea idiosincrasiilor lexicale într-un context multilingv si a disparitatilor de traducere este gruparea conceptelor foarte apropiate semantic în ceea ce s-ar putea numi *concepte agregat.* Lexicalizarile întelesurilor din doua sau mai multe limbi, puse în corespondenta cu aceleasi concepte din ILI sau cu concepte membre ale unui agregat, vor putea fi folosite ca echivalenti de traducere în pofida unor diferentieri semantice specifice unei limbi sau alteia (*ciorba*, *sarmale*, *pepper pot*, *porcupine ball* etc.; vezi si exemplele din sectiunea precedenta). Analiza inconsistentelor interumane în echivalarea întelesurilor dintr-o limba cu conceptele interlinguale din ILI, precum si

identificarea conceptelor distincte puse în corespondenta cu echivalenti de traducere (extrasi automat din corpusuri paralele sau gasiti într-un dictionar bilingv clasic) pot furniza informatii calitative mult mai interesante (cel putin din perspectiva psiho-lingvisticii) si mai demne de încredere decât o analiza statistica. Aceasta este o promitatoare directie de cercetare ce se dezvolta în paralel cu activitatea principala de constructie a wordnetului pentru limba româna.

Referinte bibliografice

Bloksma, L., Diez-Orzas and Vossen, P. (1996) The User Requirements and Functional Specification of the EuroWordNet-project *EWN-deliverable D.001*, LE-4003

Danzin, A. (1992) „Towards a European Language Infrastructure" raport al Comisiei Europene

Dinu, M. (1996). Personalitatea limbii române, Editura ALL, 368 p.

DEX (1996). Coteanu, I., Seche, L., Seche, M. (coord.). Dicţionarul Explicativ al Limbii Române, Ediția a II-a, *Univers Enciclopedic*, Bucureşti

Erjavec, T., Ide, N., Tufiş, D.(1997) 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)

Erjavec, T., Evans, R., Ide, N., Kilgarriff, A. (2000). The CONCEDE Model for Lexical Databases. *Proceedings of the Second Language Resources and Evaluation Conference* (LREC), Athens, Greece, 355-362.

Erjavec, T., Ide, N., Tufiş, D.(2001) *Automatic Sense Tagging Using Parallel Corpora*, in Proceedings of the 6th Natural Language Processing Pacific Rim Symposium, Tokyo, Japan, 27-29 November, pp. 212-219, 2001

Fellbaum, Ch. (ed.) *WordNet: An Electronic Lexical Database*, MIT Press, 423 p.

Harabagiu, S., Miller, G., Moldovan, D. (1999). „WordNet 2 - A Morphologically and Semantically Enhanced Resource", in *Proceedings of SIGLEX-99*, Univ. of Maryland, pp 1-8.

Ide, N. (1998) *Corpus Encoding Standard: SGML Guidelines for Encoding Linguistic Corpora* First International Language Resources and Evaluation Conference, Granada, Spain. See also http://www.cs.vassar.edu/CES/.

Ide, N., Erjavec, T., Tufiş, D. (2002): „Sense Discrimination with Parallel Corpora" in Proceedings of the SIGLEX Workshop on Word Sense Disambiguation: Recent Successes and Future Directions. ACL2002, Philadelphia, pp. 54-60.

Juilland, A., Edwards, P.M.G, Juilland, I. (1965). The Frequency Dictionary of Rumanian Words. *Mouton & CO.,* London-The Hague-Paris

Miller, G.A., Beckwidth, R., Fellbaum, C., Gross, D., Miller, K.J. (1990) "Introduction to WordNet: An On-Line Lexical Database" 1990 In International Journal of Lexicography, Vol. 3, No. 4 (winter), pp. 235-244

Moldovan, D. (2001). "Question Answering Systems in Knowledge Management", IEEE Intelligent Systems, vol 16, nr. 6, pp 90 – 92.

Pavelek, T., Pala, K. (2002) *VisDic: A new Tool for WordNet Editing* in Proceedings of the 1st International Wordnet Conference, Mysore

Rodriguez, H., Climent, S., Vossen, P., Bloksma, L., Peters, W., Alonge, A., Bertagna, F., Roventini, A.(1998) 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, Vol. 32, Nos. 2-3

Seche, L., Seche, M.(1997) *Dicţionarul de sinonime al limbii române.* Univers Enciclopedic, Bucureşti

Stamou, S., Oflazer, K., Pala, K., Christoudoulakis, D., Cristea, D., Tufiş, D., Koeva S., Totkov, G., Dutoit, D., Grigoriadou, M. (1997) BALKANET A Multilingual Semantic Network for the Balkan Languages, in *Proceedings of the International Wordnet Conference*, Mysore, India

Tufiş, D., Bruda, Şt. (1997). 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

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

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

Tufis, D., Rotariu, G., Barbu, A.M. (1999) TEI-Encoding of a Core Explanatory Dictionary of Romanian. In Kiefer, F. and Pajzs J. (eds.) *Papers in Computational Lexicography*, Hungarian Academy of Sciences pp. 219-228

Tufiş, D., Popescu, C., Roşu, R (2000).: Automatic classification of documents by random sampling in *Proceeding of the Romanian Academy*, Series A, vol 1, no. 2, pp. 18-28

Tufis, D. (2001), "*Promovarea Limbii Române în SI-SC*", in "Societatea Informationala - Societatea cunoasterii: concepte, solutii si strategii pentru România", Florin Gh. Filip (coord.), Ed. Expert, Bucuresti, 2001, ISBN 973-8177-42-1, pp. 131-142

Tufiş, D., Barbu, A.M.(2001a) *Computational Bilingual Lexicography: Automatic Extraction of Translation Dictionaries*, in International Journal on Science and Technology of Information, Romanian Academy, ISSN 1453-8245, Vol.4, No.3-4, 2001, pp.325-352

Tufiş, D., Barbu, A.M.(2001b) *Extracting multilingual lexicons from parallel corpora*, in Proceedings of the ACH-ALLC conference, New York, 12-17 June, 4p.

Tufiş, D., Barbu, A.M(2002). "Lexical token alignment: experiments, results and applications" In Proceedings of LREC2002, Las Palmas, Spain, pp.458-465

Tufiş, D. (2000). Blurring the distinction between machine readable dictionaries and lexical databases. R*esearch Report, RACAI-RR56, p. 56*

Tufiş, D. (2002). "A cheap and fast way to build useful translation lexicons" in *Proceedings of the 19th International Conference on Computational Linguistics, COLING2002,* Taipei, 25-30 August, 2002, 7 p

Tufiş, D., Cristea, D. (2002). „Methodological issues in building the Romanian Wordnet and consistency checks in Balkanet", In Proceedings of LREC2002, Las Palmas, Spain, May, pp. 35-41

Vintilă-Radulescu, I. (2002). "Resurse lingvistice pentru limba română elaborate la Institutul de Lingvistică «Iorgu Iordan»", în acest volum.

Vossen, P. (ed.) (1998). "A Multilingual Database with Lexical Semantic Networks", Kluwer Academic Publishers, Dordrecht

Algoritmi de segmentare a textului în unitati de tip clauzal

D. GÂLEA,
Institutul de Informatica Teoretica, Academia Româna, Filiala Iasi,
dgalea@iit.tuiasi.ro

N. CURTEANU,
Institutul de Informatica Teoretica, Academia Româna, Filiala Iasi,
curteanu@iit.tuiasi.ro

C. LINTES,
Institutul de Informatica Teoretica, Academia Româna, Filiala Iasi

1. Introducere

Scopul lucrarii este dublu: (*a*) Sa prezinte si sa compare doi *algoritmi de segmentare* a frazei (românesti) în unitati de tip clauzal. (*b*) Sa întregeasca si sa sustina doua componente de baza ale *strategiei lingvistice* SCD (Segmentare-Coeziune-Dependenta) [1], [2] de analiza a limbajului natural (LN): procesul de *segmentare* a textului de LN, si *teoria* FX-bar [3]. Segmentarea textului poate continua sau interfera cu stabilirea arborilor de dependenta între unitatile clauzale si subclauzale (unitati sintagmatice) ale textului. Unitatile de tip-clauzal corespund, în general, *relatiilor retorice* dintre unitatile minimale ale discursului, astfel încât algoritmii de segmentare pot fi (si chiar sunt) utilizati în aplicatii ce tin de *teoria* si *procesarea discursului*. Primul algoritm este o aplicare la limba româna a segmentarii frazei în unitati de tip-clauzal, algoritm dezvoltat de Daniel Marcu în [4], [5] (si prescurtat în cele se urmeaza "*algoritm-Marcu*", sau "*algoritm-M*"). Al doilea algoritm reprezinta o rafinare a segmentarii în clauze si grupuri sintagmatice din cadrul strategiei lingvistice SCD [1], [2], [3], [6] (prescurtat în cele se urmeaza prin "*algoritm SCD*"). Acesti algoritmi sunt implementati într-un mediu specializat de procesare (dezvoltat sub C++), si este realizata o comparatie computationala a executiei segmentarii de tip-clauzal pe un set consistent de fraze românesti [7].

Segmentarea textului LN a devenit în ultimii ani un subiect intensiv cercetat si cu multiple aplicatii. O atentie speciala a primit segmentarea textului de LN în unitati de discurs, în particular, segmentarea frazei în *unitati minimale* de *discurs*, de multe ori si pe buna dreptate asociate cu unitati de tip clauzal în numeroase teorii sintactice, semantice, si de discurs. Unitatile textuale de tip-clauzal obtinute (sau proiectate) prin mijloace orientate

sintactic, pe analiza de suprafata [Eng: *shallow*], s-au dovedit a fi esentiale în numeroase tipuri de procesare a LN: parsare, traducere automata, generare de LN, interpretare de discurs, extragere de date lingvistice, regasirea informatiei, rezumare automata, rezolutia anaforei, etc. Un caz special de segmentare a textului este ceea ce numim '*chunking*', un proces dedicat obtinerii unor tipuri de "*segmente*" [Eng: *chunks*] dominate de anumite categorii (verb, substantiv, adjectiv-adverb, clauza). În continuare, vom folosi doar termenul de *segmentare* a textului de LN, considerând *chunking*-ul drept un caz particular al procesului de segmentare a LN.

În analiza si implementarea celor doi algoritmi de segmentare, *algoritmul-M* si *algoritmul*-SCD, cel putin doua aspecte le consideram a fi importante: **(a)** Se demonstreaza ca *algoritmul-M* de segmentare este scufundat în *algoritmul*-SCD, ceea ce înseamna ca primul dintre cei doi algoritmi poate fi obtinut ca un caz particular al claselor de marcheri, ierarhiei acestor clase, si a segmentarii (dependentelor) obtinute 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 cascada, fiecare sistem component fiind format din *seturi si subseturi specifice de etichete*, (b2) *o ierarhie* stabilita între câteva dintre cele mai importante *clase* ale acestor *etichete*, si (b3) *o gramatica formala* (sau un automat finit) pentru recunoasterea (sub)secventelor si arborilor de etichete (în concordanta cu ierarhia claselor de etichete). În abordarea prezenta, aceste aspecte sunt exemplificate de catre o implementare C++ a celor doi algoritmi într-un mediu specializat, o baza de date a marcherilor (de discurs) românesti, si o ierarhie specifica a claselor de marcheri lingvistici. Cei doi algoritmi de segmentare considerati sunt executati si comparati pentru un set consistent de fraze românesti. Posibile dezvoltari si aplicatii sunt mentionate [21], [22].

Importanta segmentarii de tip clauzal a frazei în procesul de parsare a textului a fost scoasa în evidenta înca de la începutul anilor '80, iar studii teoretice dateaza mult mai devreme. În România, primele lucrari stiintifice si contracte de lingvistica computationala au continut, printre alte realizari meritorii, si primele încercari de realizare a segmentarii automate a frazei în clauze finite (si infinite) [8], [9], [10], [11]. În pofida unor modeste mijloace formale (gramatici formale) si de programare (retele ATN) disponibile la acel timp, ideile principale pe care se bazau abordarile mentionate, nu numai ca au reprezentat premiere pentru acele timpuri, dar multe din ideile de atunci îsi pastreaza înca o surprinzatoare actualitate, aceste fenomene de *come-back* ciclic fiind frecvente (si perfect explicabile de evolutia tehnologica) în momentul de fata. Trebuie mentionate aici folosirea intensiva a *marcherilor de discurs* (*cue phrases*, *connectives*), întâlnita si în [4], [5], [12], [13], a *predicativitatii* (aparitia categoriilor '*deverbale*') [13], [14], a utilizarii automatelor finite în analiza LN, etc.

De fapt, o versiune a gramaticii formale preluata din [8] este folosita în *Pasul 6* al *algoritmului SCD*-2002 de segmentare, în concatenarea marcherilor de nivel M3 si M2 (vezi Sectiunea 4), în timp ce rudimente ale unor reguli similare din aceeasi gramatica se

regasesc în algoritmul-*M* de segmentare, la compunerea actiunilor care lucreaza cu aparitia multipla a marcherilor [4], [5] (vezi Sectiunea 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 unitati de tip-clauzal [4], [5] functioneaza ca un automat finit, sau ca o retea de tranzitie, bazat pe un set de *stari* si *actiuni*. În [4] se face o analiza de corpus a potentialilor *marcheri de discurs*, numiti si "*sintagme indicatoare*" [Eng: *cue phrases*] si "*conective*", cu scopul de a evalua contributia potentiala a diferitilor marcheri la determinarea (delimitarea) unitatilor textuale elementare pe care sunt definite *relatiile retorice*, în cadrul unitatii textuale standard care este fraza. În încercarea de a stabili principalele tipuri de functii ale marcherilor, si anume de tip clauzal, frazal, de discurs, sau pragmatic, algoritmul de segmentare *M*-1997 considera mai întâi urmatoarele *trei clase* de marcheri:

(Mar1) 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 (Mar1) 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*, numiți "*marcherii de frază/clauză*", joacă în discurs, pentru majoritatea fragmentelor de text în care apar, rolul de *adiacenți* la alți 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* conține marcheri care s-au dovedit că joacă un rol de *delimitare a clauzelor* în majoritatea fragmentelor de text investigate în [4]; ei vor fi referiți, simplu, ca "*marcheri clauzali*" (sau "*de clauză*"). (Mar3) include, de asemenea, acei marcheri pentru care analiza de corpus nu a putut distinge între funcția 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 stocați şi procesați într-o *bază de date* ale cărei înregistrări conțin următoarele câmpuri:

a. Câmpul denumit *Exemplu* contine un *fragment de text* din care a fost extras marcherul.

b. Câmpul *Marker* codifica *marcherul* însusi, împreuna cu marcherii de punctuatie contextuali si, atunci când este necesar, ceilalti *marcheri adiacenti*.

c. Câmpul *Usage* furnizeaza unul sau mai multe dintre *rolurile functionale* ale marcherului:

(c1) *Frazal*/*clauzal* (S), atunci când marcherul nu îndeplineste *nici o functie* în structurarea discursului;
(c2) *De discurs* (D), când marcherul evidentiaza o *relatie de discurs* între doua unitati textuale;
(c3) *Pragmatic* (P), daca exista o *relatie* între o constructie lingvistica (sau non-lingvistica) care contine marcherul, si convingerile, planurile, intentiile si/sau scopurile de comunicare ale vorbitorului.

d. Câmpul *Break_action* (actiune de oprire) contine un nume de *actiune* din multimea actiunilor ce vor fi executate în cadrul procesului de segmentare. Acest proces este controlat de catre un set de semnalizatori (*flaguri*). Executia unei actiuni din multimea {NOTHING, NORMAL, COMMA, NORMAL_THEN_COMMA, END, MATCH_PAREN, COMMA_PAREN, MATCH_DASH, SET_AND, SET_OR, DUAL} are unul dintre urmatoarele efecte:

(d1) creează o *margine* pentru unitatea textuală elementară în *string*-ul de intrare;
(d2) setează un semnalizator (*flag*).

e. Câmpul *Position* specifica pozitia marcherului de discurs în cadrul unitatii textuale careia îi apartine. Valorile acestui câmp sunt *B*, *M* si *E*, dupa cum marcherul este situat *la început* (*B*), *în mijlocul* (*M*) sau, respectiv, *la sfârsitul* (*E*) unitatii textuale.

3. Algoritmului de segmentare *M*-1997

Algoritmul *M*-1997 primeste în intrare o fraza S si masivul *markers*[*n*] al marcherilor potentiali de discurs si clauzali din fraza S. Masivul *markers*[*n*] contine marcherii recunoscuti în S. Fiecare element al acestui masiv este caracterizat de catre urmatoarea *structura de trasaturi*:

- *Actiunea* asociata acelui marcher;
- *Pozitia* marcherului în cadrul unitatii textuale elementare (*B*, *M* sau *E*);
- *Semnalizatorul has_discourse_function* care initial este setat la valoarea "*no*".

Câteva dintre variabilele importante cu care lucreaza algoritmul *M*-1997 sunt: "*status*", "*parenthetical*" si "*clauses*".

Algoritmul M-1997 pentru identificarea unitatilor de tip-clauzal din cadrul unei fraze are *doua* parti principale:

(1) Când variabila *"status"* este NIL, algoritmul *M*-1997 executa actiuni care pot introduce margini ale unitatii textuale sau pot modifica variabila, influentând procesarea marcherilor ulteriori. Pentru partea (1) a algoritmului *M*-1997, atunci când variabila *"status"* ia valoarea NIL, sunt considerate urmatoarele situatii:

(1a) Daca tipul de marcher este DUAL, determinarea marginilor unitatii textuale depinde de marcherul adiacent care precede marcherul curent analizat. În aceasta situatie, algoritmul *M*-1997 seteaza variabila "*status*" la aceeasi valoare ca si în cazul unui marcher de tip COMMA.

(1b) Daca marcherul analizat curent nu este adiacent cu marcherul imediat precedent, atunci este identificata o margine a unitatii textuale.

(1c) Cel mai frecvent tip de marcher (si de actiune) este NORMAL, marcher care identifica o noua unitate de tip clauzal a carei margine-dreapta este data 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 identifica o noua unitate de tip-clauzal ca si în cazul marcherului de tip NORMAL.

În oricare dintre cazurile (1c), (1d), (1e), variabila "*status*" este actualizata astfel încât o margine a unitatii textuale sa fie identificata la prima aparitie a unei virgule (COMMA).

(1f) Pentru marcherul de tip NOTHING, singura actiune consta în atribuirea marcherului o utilizare specifica discursului.

(1g) Marcherii care introduc posibile aparitii de unitati textuale parantetice (texte între paranteze) au doar efectul de a actualiza variabila "*status*", ca si în cazul aparitiei marcherilor "*si*" si "*sau*".

(2) Atunci când variabila "*status*" nu este NIL, algoritmul *M*-1997 executa actiuni specifice pentru a realiza:

(2a) Tratarea informatiei din paranteze. O data identificata o paranteza deschisa, o linie-de-despartire [Eng: *dash*] (între doua asemenea liniute se introduce de obicei o apozitie sau un text explicativ), sau un marcher de discurs a carui actiune asociata este COMMA_PAREN, algoritmul *M*-1997 cauta prima paranteza închisa, linie-de-despartire, sau virgula, ignorând toti ceilalti marcheri întâlniti pe parcurs. Acest tratament atrage dupa sine faptul ca informatiei parantetizate *nu* îi este atribuita nici o stare pentru unitatile textuale elementare. Totusi, algoritmul *M*-1997 evita stabilirea de margini parantetizate în cazurile în care prima virgula care urmeaza dupa un marcher COMMA_PAREN este imediat urmata de un marcher "*si*" ori "*sau*". De mentionat este, de asemenea, ca tratamentul aplicat informatiei dintre paranteze în algoritmul *M*-1997 poate conduce la rezultate eronate, ca în exemplul "*I-am dat lui Ion o racheta de tenis, care i-a placut si o minge de plastic, care nu i-a placut*". Acest tip de erori poate fi evitat printr-o tratare mult mai adecvata în cadrul algoritmului de segmentare *SCD*.

(2b) Daca variabila "*status*" contine actiunea COMMA, aparitia primei virgule care nu este adiacenta unui marcher "*si*" ori "*sau*" determina identificarea unei noi

unitati elementare de discurs. Algoritmul *M*-1997 nu este, capabil, în general, sa distinga suficient de precis între rolurile de discurs si frazale/clauzale ale marcherilor "*si*" si "*sau*". Anumite situatii sunt totusi recunoscute ca introducând functii de discurs, ca de exemplu aparitia unui marcher de discurs imediat dupa un "*si*" ori "*or*", caz în care valoarea semnalizatorului *has_discourse_function* este stabilita la "*yes*".

Forma originala a algoritmului *M*-1997 [4], [5] este extinsa si îmbunatatita în implementarea noastra pentru limba româna (subsectiunea 5.3) cu o analiza mai detaliata la nivelul ei superior, pentru aparitii multiple si corelate ale marcherilor de discurs/clauza.

4. Algoritmul de segmentare *SCD*-2002

Aceasta sectiune prezinta partea de segmentare si dependenta, în principal la nivel de clauza, desprinsa din strategia lingvistica *SCD* (*Segmentare-Coeziune-Dependenta*) [1], [2], [3], [6]. Forma actuala a algoritmului, referita în restul articolului prin prescurtarea *SCD*-1994, este foarte apropiata de versiunea publicata în [1], [2]. Noutatea principala a algoritmului *SCD*-2002 fata de *SCD*-1994 consta într-o rafinare a claselor de marcheri, o noua ierarhie a acestora, si în noul algoritm de stabilire a segmentarii si dependentei (structurarii) clauzelor si grupurilor sintagmatice. Vom pune în evidenta relatia dintre algoritmul *M*-1997 si algoritmii *SCD*-1994 si *SCD*-2002, aratând ca primul este scufundat în ceilalti doi.

Rezultatele obtinute prin executia algoritmilor de segmentare *M*-1997 si *SCD*-2002 pe aceleasi fraze conduc la aceeasi concluzie: *SCD*-2002 are o granularitate (mult) mai fina a claselor de marcheri în comparatie cu cea a claselor algoritmului *M*-1997, iar rafinarea actiunilor implicate în *SCD*-2002 conduce la delimitarea de unitati textuale de tip-clauzal mai precise (de fapt mai corecte si mai adecvate) decât cele obtinute de catre algoritmul *M*-1997, pretul computational ce trebuie platit pentru acest fapt ramânând sa fie analizat.

Este de mentionat ca segmentarea clauzala practicata de *SCD*-2002 este doar un aspect particular al segmentarii textului, deoarece se obtin si alte "*bucati*" mai mici de text dominate de nuclee semantice de tip N (Substantiv), V (Verb), A (Adjectiv-Adverb). Segmentarea rezultata din clasele de marcheri *SCD*-2002 se afla într-o strânsa relatie cu noua teorie X-*bar functionala* (FX-bar) [3], o alta componenta importanta a strategiei lingvistice generale *SCD*.

Din schema generala FX-bar propusa în [3] se detaseaza urmatoarele nivele de proiectie la nivel lexical si gramatical:

Tabelul 4.1.

Nivele de proiectie ale schemei FX-bar (vezi [3])

Marcheri	Nivelul de Proiecție	Structura gramaticală	Exemple

trăsătura PRED sau EXIST (OBJECT)	nivel de lexicon; prin convenţie, (BAR = –1)	forma de dicţionar a cuvântului; X = N, V, A, Pron, ...	*a ploua* *conducere* (trăsătura PRED) *clădire* (trăsătura EXIST înţelesul obiectual) *clădire* (PRED, pentru înţelesul acţional) *creion* (EXIST)
M0-marcher reprezintă aplicarea inflexiunii M0(X)=X0	X0 (BAR = 0)	forma lexicală (de text) a cuvântului; X=N, V, A, …	*plouă* ….
M1-marcher se aplică nucleului X0 M1(X0)=X1	X1 = CL0; (BAR=1) poate fi identificat şi cu nivelul 0 de proiecţie a clauzei, BAR-CL = 0	sintagme XG (X=N, V, A), i.e. grupuri nominale, verbale, adjectivale-adverbiale	*orice steag alb* *ploua* *aleargă repede* *nu aleargă deloc* *foarte bine studiat*
M2-marcher se aplică proiecţiei X1 M2(X1)=X2=CL1 M2 se aplică unei singure clauze	proiecţia X2 = CL1 BAR = 2 şi BAR-CL = 1	clauza finită sau infinită	*Maria i-a dat un măr fiicei sale.* *O femeie dăruind un măr unui bărbat conţine o clauză infinită.*
M3(CL1, CL1)=CL2 marcheri de discurs; M3 se aplică la două sau mai multe clauze	nivelul de proiecţie X3 = CL2; BAR = 3 şi BAR-CL = 2	relaţii de discurs între clauze finite	*Dacă plouă atunci plec mai devreme şi îmi i-au şi umbrela.*

4.1. Clasele de marcheri pentru algoritmul *SCD*-2002

Pentru algoritmul de segmentare *SCD*-2002 propunem o anumita rafinare a claselor de marcheri si a ierarhiilor acestor clase din [1], [2], schimbari ce constau în urmatorul set de marcheri, în concordanta cu *Tabelul* 4.1. de mai sus:

M3 = **{** *marcheri (de discurs) inter-clauzali* **}**.

Clasa de marcheri M3 este formata din *functii* sau *relatii* (atunci când marcherii sunt corelati), având ca argumente doua sau mai multe clauze finite (unele dintre ele pot fi infinite). Acesti marcheri sunt ceea ce [4], [5], precum si alte abordari numesc "*marcheri de discurs*", si se aplica proiectiilor sintactice de nucleu X2 = CL1 (si de nivel X3), de tip clauzal în teoria FX-bar (vezi Tabelul 4.1.).

M3 poate fi partitionata în urmatoarele subclase (în ordinea *descrescatoare* a *prioritatii* de definire a relatiilor de dependenta – vezi Fig. 4.1.1.):

M33 = { marcheri (de discurs) inter-clauzali care introduc o dependenta (neambigua) de *supra-ordonare stricta* }. *Supra-ordonarea stricta* înseamna *ridicarea* efectiva a *(cel putin) unui nivel* de dependenta clauzala, si este reprezentata de marcheri precum "*atunci*", "*altfel*", etc.

M32 = { marcheri (de discurs) inter-clauzali care introduc dependenta de *supra-ordonare*, incluzând *semnele de punctuatie* precum doua puncte, punct-si-virgula, paranteza închisa, linie-de-despartire, etc. }. *Supra-ordonarea* presupune *ridicarea* unuia sau mai multor nivele de dependenta clauzala, sau ramânerea pe acelasi nivel de dependenta în cadrul unei dependente de tip-*coordonare*. Exemple tipice de marcheri din clasa M32 sunt "*dar*", "*asadar*", "*chiar*", "*la_fel_(de)*", "*în_comparatie_(cu)*", etc.

M31 = { marcheri (de discurs) inter-clauzali care introduc unul sau mai multe nivele de dependenta de *sub-ordonare*, incluzând *semne de punctuatie* ca paranteza deschisa, linia-de-despartire, etc. } Aceasta este o clasa larga de marcheri de discurs formata din numeroase tipuri de relatii între clauze: logice, sintactice, semantice, pragmatice, etc.

Asa cum a fost mentionat mai sus, fiecare dintre clasele M33, M32 si M31 poate, la rândul ei, sa fie partitionata în subclase care contin marcheri de tip relational (exprimati prin corelatie), ce stabilesc relatii între clauze, sau ca functii de clauze (cu cel putin doua argumente).

M2 = **{** *marcheri care introduc o clauza (finita sau infinita), sau un grup sintagmatic al carui nucleu semantic este una din categoriile sintactice 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 divizata în urmatoarele subclase (în ordinea descrescatoare a prioritatii de introducere a relatiilor de dependenta):

M25 = { marcheri care introduc *clauza relativă* }.

Explicatia consta în faptul ca o clauza relativa reprezinta cea mai complexa unitate sintagmatica ce joaca rol de modificator, si care se aplica nucleului NG al clauzei relative:

M24 = { aparitia unui *grup verbal finit* (FVG) sau, echivalent, aparitia valorii FINITE pentru trasatura TENS atribuita unui verb, introducând deci o *clauza finita* }.

Întregul grup verbal poate mosteni valoarea trasaturii FINITE daca nucleul sau V sau alta componenta importanta din VG poarta aceasta valoare a trasaturii TENS (de exemplu, auxiliarul din VG).

M23 = { aparitia unei sintagme *predicationale* XG (sau X1), X=V, N, A, al carei nucleu semantic este o *categorie predicationala*, purtând valoarea PRED = ACT (posibil înca la nivel de lexicon), si introducând astfel o *clauza infinita* }.

Clasele de marcheri M24 si M23 introduc structuri de nivel-X2, si anume clauze finite sau infinite, formate dintr-o sintagma X1 (sau grup XG, X = N, V, A) care reprezinta *nucleul semantic*, *finit* (TENS = FINITE) sau *predicational* (PRED = ACT), al structurii de nivel-X2, urmata de sateliti (argumente si/sau adjuncti) corespunzatori de tip NG (inclusiv NG-uri prefixate de o prepozitie, deci clasica sintagma PP). Unele dintre argumente, cum este cazul clasic al *subiectului gramatical*, pot preceda nucleul semantic de tip X1 al clauzei careia îi apartin [3]. Sa mai precizam ca exista o *ordine sistemica* (*canonica*) [18], [19], a satelitilor, sau "*actantilor*" (argumente si adjuncti) dintr-o clauza (finita sau

infinita): ACT(or), PAT(ient), ADDR(essee), ORIG(ine), LOC(ation), etc. Ordinea canonica este specifica fiecarui LN, si se poate obtine în urma unei cercetari statistice si lingvistice foarte atenta [18].

Putem gasi recent un *principii de predicativitate* similare cu cel folosit în *strategia lingvistica* SCD, si aplicat la sintagmele nominale din limba italiana [14], sau la adjectivele "*deverbale*" [14], [16]. În timp ce predicativitatea verbelor este frecventa si naturala, trasatura de *nepredicativitate* [17, p. 22] (de fapt, nepredicationalitate) a verbelor de tip *existential* este si ea la fel de frecventa (formele lui "*a fi*"), valoarea lor FINITE, dublata sau nu de valoarea trasaturii PRED = ACT, anuntând totusi aparitia unei clauze finite.

M22 = { marcheri care introduc relatii de tip-JOIN, adica *conjunctii* de tipul "*si*", "*sau*", "*la_fel_ca_(si)*", "*împreuna_(cu)*" }.

M21 = { COMMA (sau VIRGULA) }.

Clasele M22 si M21 cuprind marcheri cu un grad important de ambiguitate deoarece pot introduce orice structura de tip X1 (grupuri XG, X = N, V, A) sau X2 (clauze finite sau infinite).

M1 = **{** *marcheri care delimiteaza (introduc) structuri XG* **}**.

Conform strategiei SCD si teoriei FX-bar [3], clasa de marcheri M1 consta în *marcheri de nivel*-X1, X = N, V, A, adica marcheri care se aplica constructiilor sintactice de *nivel*-X1 (denotat si XG, si numit X-grup). Aceste sintagme constau, de fapt, dintr-un *nucleu semantic* înconjurat de *modificatori* (adjective sau adverbe) si/sau *specificatori* (sau *cuantificatori*, unii generalizati, printre cuantificatori incluzându-se determinatorii, negatia, etc.).

Asa cum exista o ordine sistemica a satelitilor unui nucleu semantic într-o clauza (sintagma de nivel-X2), în mod similar exista o "*ordine structurala*", data de "*distanta*" modificatorilor, cuantificatorilor, prepozitiilor, etc. fata de nucleul X0, pentru constituentii unei sintagme de nivel-X1. Astfel, în limba româna (franceza, engleza), cel mai "apropiat" fata de nucleul X0 trebuie sa fie *modificatorul* (adjectivul sau adverbul), urmeaza apoi *cuantificatorul* (care ocupa locul modificatorului daca acesta lipseste), apoi prepozitia (ad-pozitia, în general), etc. De exemplu, nu este sintactic corecta sintagma "*frumos orice copil*", sau "*orice frumos pe copil*". Nucleul X0 înconjurat de modificatori si/sau specificatori (cuantificatori) poate fi marcat functional prin *pre-pozitii* (în cazul grupului nominal NG din româna, engleza, franceza), dar si prin *post-pozitii* (în cazul NG sau VG din engleza sau germana). Marcarea clitic-functionala (prin particule *pre-* sau *post-pozitionale*) poate exprima *cazul* (pentru NG), sau *timpul*, *semantica* (pentru VG), etc. Principalele elemente componente ale unei structuri XG corespund si subclaselor de marcheri ai clasei M1.

M1 poate fi divizata în subclase de marcheri, subclase utile în delimitarea substructurilor XG (X1), X = N, V, A, în conformitate cu un criteriu cum este *distanta* dintre nucleul semantic X0 si elementele functionale care îl "*înconjoara*"; un asemenea nucleu este, în ultima instanta, un substantiv comun obiectual (numit si *autosemantic* în

[19]), un nume propriu, sau un substantiv personalizat (dar fara nume propriu, denominalizat).

M14 = { aparitia unui substantiv comun *obiectual* (nepredicational, autosemantic), a unui nume propriu, sau a unui substantiv personalizat denominalizat }

M13 = { aparitia unui *modificator* (adjectiv, adverb, adjectiv pronominal) }

M12 = { aparitia unui *cuantificator* (generalizat) }

M11 = { pre-pozitii sau *post-pozitii* exprimând *cazul* (pentru N), timpul sau *semantismul* (pentru V), etc. }

Ultima clasa de marcheri, notata **M0** (sau M00 pentru uniformitate), si ai carei marcheri se aplica *formei de dictionar* a cuvântului, este reprezentata de *rolul functional* al *flexionarii*.

Recapitulând, clasele de marcheri considerate de strategia lingvistica SCD, în particular de *algoritmul de segmentare* SCD-2002, pot fi reprezentate grafic de urmatoarea ierarhie:

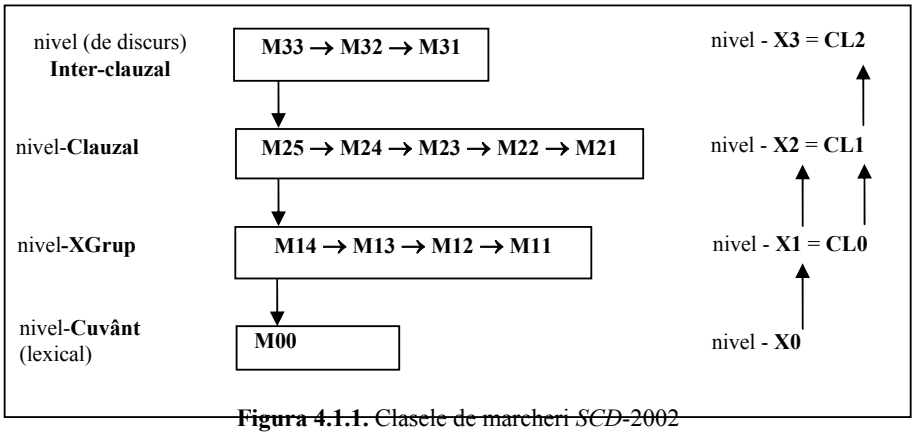


Figura 4.1.1. Clasele de marcheri *SCD*-2002 si ierarhia lor

Orientarea arcelor din Fig. 4.1.1., stabilite între clasele si subclasele de marcheri, provine dintr-o *ordine* de *prioritatea descrescatoare* între marcherii considerati, si este reprezentata mai jos prin relatia "≪" dintre clasele si subclasele de marcheri. Aceasta ierarhie este o *ipoteza de baza* impusa în *strategia lingvistica* SCD si, prin consecinta, si în *algoritmul de segmentare SCD*.

(4.1.2) $\forall(j = 1\div4)$ $M(k+1)(j+1) \ll M(k+1)j$ $(k = 0\div2)$;

(4.1.3) $\forall(k = 0\div2)$ $M(k+1)i \ll Mkj$ $(i = 1\div5), (j = 0\div5)$.

Aceste inegalitati ne spun ca marcherii din subclasa M(k+1)(j+1) sunt de *prioritate mai mare* în comparatie cu marcherii din subclasa M(k+1)j , (k = 0÷2), (j = 1÷4), în cadrul aceleiasi clase M(k+1) de marcheri aflata pe *acelasi nivel* de proiectie lingvistica, iar marcherii din aceeasi clasa M(k+1) au o *prioritate mai mare* fata de marcherii din clasa Mk de pe nivelul *inferior* de proiectie lingvistica.

Aceasta ierarhie a marcherilor si claselor de marcheri este considerata de noi ca fiind valida pentru *limba româna*. Probabil ca anumite modificari vor fi necesare când se trece de la un LN la altul. Daca ne situam în domeniul mai restrâns al limbajelor indo-europene (cum sunt franceza, engleza, germana, italiana, spaniola, posibil rusa), se poate aprecia ca structurile si clasele de marcheri propuse în Tabelul 4.1. si Fig. 4.1.1. ramân aceleasi sau foarte asemanatoare, cu anumite modificari parametrizate în functie de limbaj.

4.2. Algoritmul *SCD*-2002 de segmentare si stabilire a dependentelor

Urmând algoritmii de segmentare si dependenta (numiti si *meta-algoritmi* SCD) propusi în [1] si [2] (denotati în continuare *SCD*-1994), rafinati cu clasele de marcheri considerate în subsectia precedenta, se obtine forma prezenta a *algoritmului de segmentare SCD* (denotata *SCD*-2002). Dezvoltam aici forma *secvential-liniara* a acestui algoritm, însa în [1] sunt expuse si o forma *secvential-recursiva*, ca si o versiune *paralela* a algoritmului. O forma "*inversata*" (pentru care în intrare avem un arbore de derivare sau o formula logica, iar în iesire – ca si în intrarea în algoritmul standard – avem o fraza) poate fi folosita pentru a ghida procesul de generare a unei fraze de LN [2], schimbând operatia de recunoastere a marcherilor cu cea de generare a lor, si analiza (parsarea) compusilor sintactici cu generarea lor.

În *descrierea algoritmului* de segmentare *SCD*-2002 sunt folosite câteva *operații* al căror înțeles este bine să fie precizat de la început.

(4.2.a) *Recunoasterea marcherilor* înseamna inserarea în text a unor etichete adecvate, ce corespund marcherilor care realizeaza delimitarea unitatilor textuale sintactice, semantice, si de discurs.

(4.2.b) *Verificarea marcherului* înseamna preluarea, din baza de date a marcherilor, a celor mai importante valori din structura de trasaturi a acelui marcher.

(4.2.c) *Segmentarea* implica a analiza liniara (parsare) a secventei de etichete de marcheri, si *recunoasterea* unei subsecvente (eventual discontinua) care face parte din secventa originala de etichete de marcheri.

(4.2.d) *Recunoasterea structurii sintactice* înseamna *segmentarea* si *recunoasterea* structurilor sintactice elementare cum sunt NG, VG, AG, clauza infinita, si clauza finita.

(4.2.e) *Compunerea structurilor* (*de dependenta*) consta în stabilirea dependentelor (sub-ordonare, co-ordonare, supra-ordonare) succesive

dintre structurile sintactice recunoscute, pe baza rolului functional specific al marcherilor care delimiteaza aceste structuri, si utilizând ierarhia corespunzatoare dintre clasele carora le apartin acesti narcheri (vezi Fig. 4.1.1. si relatiile (4.1.2.-4.1.3.).

Algoritmul de segmentare SCD-2002

Step01. Recunoasterea pe text a marcherilor din clasa M3;

Step01. Recunoasterea pe text a marcherilor din clasa M2;

Step03. Verificarea contextuala si recunoasterea aparitiei corelate a marcherilor de tip M3 si M2[(1)];

Step04. Segmentarea frazei în clauze finite;

Step05. Segmentarea (chunking), daca este necesar, a clauzelor finite în clauze infinite;

[Stop: Daca scopul procesarii este de a obtine o structura liniara a clauzelor finite si/sau infinite din fraza**].**

Step06. Verificarea marcherilor M3 si stabilirea relatiilor de dependenta inter-clauzala[(2)];

[Stop: Daca scopul procesarii este doar de a obtine arborele de dependenta a clauzelor finite (si infinite) din fraza**].**

Step07. Recunoasterea pe text a marcherilor din clasa M1;

Step08. Verificarea contextuala si recunoasterea (eventualei aparitii corelate) a marcherilor M1[(3)];

Step09. Recunoasterea structurilor XG (X = N, V, A)[(4)];

Step10. Verificarea marcherilor M24 si M23, si stabilirea relatiilor de dependenta dintre structurile infinite, intra-clauzale de tip XG[(5)];

[Stop].

Indicii superiori (**n**) care apar în algoritmul de mai sus corespund urmatoarelor *remarci*:

(**1**) Marcherii corelati pot fi reprezentati ca upluri ordonate (liste) de marcheri.

(**2**) Relatiile de dependenta clauzala pot fi stabilite (ca în [8, Anexa 9, p. 108], de exemplu) prin utilizarea unei *gramatici formale* (ambigue) definita pe secvente de marcheri din (sub)clasele M3, M25, M22, si M21.

(**3**) Marcherii complecsi pot fi sintagme sau expresii de tipul gradelor de comparatie a adjectivelor, diferiti cuantificatori generalizati, etc.

(**4**) În executia acestui pas se realizeaza parsarea sintagmelor XG dintr-o clauza finita si infinita.

(**5**) Dependentele dintre structurile de tip XG sunt stabilite în principal prin utilizarea trasaturilor si valorilor de trasaturi TENS = FINITE sau INFINITE, si PRED = ACT sau EXIST, pe care le poseda nucleele semantice ale sintagmelor XG, X = N, V, A (a se vedea [3]). Aceste valori pot fi *mostenite* din reprezentarea de lexicon a cuvintelelor care poarta aceste trasaturi si care formeaza XG, sau pot fi *dobândite* de catre nucleul semantic al XG în procesul de recunoastere (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:

(5.1.1) **1**-marcheri = M3 ∪ M25 ∪ M22;

2-marcheri = M24;

3-marcheri = M23;

4-marcheri = M21 ∪ M1

Prezentam în continuare algoritmul de segmentare *SCD*-1994 (în forma secvential-recursiva), asa cum a fost expusa în [1, p.68-69], având ca scop parsarea LN. Algoritmul *SCD*-1994 (în forma secvential-liniara) si destinat sarcinii de *generare* a LN este prezentat în [2, p.172-173].

Algoritmul de segmentare SCD-1994 în forma secvential-recursiva (SR)

Step01. Recunoasterea marcherilor de clauza**.**

Step02. Recunoasterea sintagmelor VG (grupuri verbale) finite si infinite**.**

Step03. Verificarea contextuala a marcherilor**.**

Step04. Segmentarea clauzala**.**

Step05. Segmentarea sub-clauzala**.**

Step06. Recunoasterea **1**-marcherului**;**

Recunoasterea **1**-structurii:

Wait-until **1**-structura este completa**.**

Step07. Recunoasterea **2**-marcherului**;**

Recunoasterea **2**-structurii:

Wait-until structura de nivel-X2 este completa***.**

Step08. Recunoasterea **3-**marcherului**;**

Recunoasterea **3-**structurii**.**

Step09. Recunoasterea **4-**marcherului**;**

Procesarea **4-**structurii**.**

Step10. 3-structura completa**?**

Nu: *Go-to* Step08.

Da: Compune **3**-structuri**;** *Go-to* Step11.

Step11. 2-structura completa **?**

Nu: *Go-to* Step07.

Da: Compune **2**-structuri**;** *Go-to* Step12.

Step12. 1-structura completa **?**

Nu: *Go-to* Step06.

Da: Compune **1**-structuri**;** *Go-to* Stop.

Stop.

***** *Structuri* AX-bar (în original, în [1]), înțelegând structuri sintactice derivate din *schemele* X-bar *augmentate*, definite în [20] si extinse în [3]. Scopul acestui pas al algoritmului este de a completa clauza finita introdusa printr-un grup verbal finit.

Principala problema cu algoritmul de segmentare si dependenta *SCD*-1994 (forma SR) este ca sunt necesare "*multiple nivele de recursie pentru a completa si compune structurile*" [1, p.69].

5.2. Algoritmii de segmentare *M*-1997 şi *SCD*-2002

În aceasta subsectiune vom arata ca algoritmul de segmentare *M*-1997 este *scufundat* în algoritmul *SCD*-2002 (de fapt, si în *SCD*-1994).

M-1997 este un algoritm de "*suprafata*" destinat segmentarii discursului în unitati textuale de tip-clauzal. În timp ce, pentru acest scop, *M*-1997 foloseste numai *marcheri de discurs* ("*cue phrases*" sau *conective*), algoritmul *SCD*-2002 utilizeaza un set de clase de marcheri mai larg si în acelasi timp mai rafinat, set care include clasele de marcheri din *M*-1997 ca un caz particular. Mai precis, relatiile dintre clasele de marcheri Mar1, Mar2, si Mar3 (vezi Sectiunea 2) utilizate pentru *M*-1997, si clasele de marcheri Mkj ale algoritmului *SCD*-2002 sunt urmatoarele:

(5.2.1) $\text{Mar1} \cup \text{Mar2} \cup \text{Mar3} \subseteq \text{M3} \cup \text{M25} \cup \text{M22} \cup \text{M21}$

sau, posibil, mai precis:

(5.2.2) $\text{Mar1} \cup \text{Mar3} \subseteq \text{M3} \cup \text{M25}$ *şi* $\text{Mar2} \subseteq \text{M22} \cup \text{M21}$

Diferenta dintre algoritmii *M*-1997 si *SCD*-2002 nu consta doar în faptul ca al doilea algoritm are un numar mai mare de clase, care sunt mai fine (mai precise), ci, mai important este faptul ca aceste clase formeaza un *sistem ierarhic* (expus în Fig. 4.1.1.) ce este utilizat în procesele de *segmentare* si de *stabilire a dependentelor*. *SCD*-2002 furnizeaza noi clase de marcheri, cum sunt M23 si M24 (aparitia categoriilor predicationale si/sau având un timp finit), precum si clasa M1, cu subclasele sale (aparitia unor componente ale sintagmei XG, X = N, V, A). Acesta este un *prim argument* din care rezulta ca *M*-1997 este *scufundat* în *SCD*-2002. "*Scufundarea*" este un termen care reflecta, de fapt, un proces de rafinare si de crestere a preciziei în calculul marginilor (limitelor) unitatilor textuale si a dependentelor dintre ele, pentru *SCD*-2002 în comparatie cu *M*-1997.

Al doilea argument important care sustine validitatea relatiei afirmate între cei doi algoritmi este urmatorul: fiecare *actiune* din *M*-1997 are un corespondent într-o *operatie* (sau o multime de *operatii*) din algoritmul *SCD*-2002 (subsectiunea 4.2).

Pentru segmentare, algoritmul *M*-1997 asociaza fiecarui marcher, în baza de date a marcherilor, o anumita *actiune* ce este statistic determinata de catre analiza de corpus efectuata în [4]. Corespondenta dintre operatiile algoritmul-*SCD*, si o *actiune* din algoritmul-*M*, se face în felul urmator:

(5.2.a) *Actiunea* (si marcherii) NORMAL din algoritmul-*M* are acelasi efect cu operatiile de procesare a marcherilor de discurs din clasa M3 a algoritmului-*SCD*. Când este întâlnit un asemenea marcher, aceasta înseamna ca o clauza (în *SCD*-2002) sau o unitate de tip-clauzal (în *M*-1997) este pe cale de a se încheia si o alta clauza, respectiv unitate de tip-clauzal, este probabil ca va începe.

(5.2.b) *Actiunile* COMMA, SET_AND, si SET_OR din algoritmul-*M* sunt folosite pentru a dezambiguiza rolul unor marcheri din M3 pentru care nu se poate aplica întotdeauna regula generala (*actiunea* NORMAL). Acesti marcheri sunt urmatorii pentru limba româna: "**,**" [Eng: *comma*], "*si*", si "*sau*". Rolul acestor marcheri este ambiguu deoarece comportamentul lor nu este uniform în cadrul delimitarii unitatilor textuale. *SCD*-2002 rezolva aceste cazuri cu ajutorul utilizarii unei gramatici formale de marcheri care descrie principalele reguli de delimitare si dependenta a clauzelor (în limba româna). Aceasta gramatica (vezi indicele superior (**2**) din *SCD*-2002 si remarca corespunzatoare) are ca scop sa recunoasca secventele cele mai frecvente de marcheri din clasele M3 si 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 clauza finita în sensul gramatical al notiunii, asa cum este folosit în algoritmul-*SCD*. O asemenea unitate de tip-clauzal, în sens *M*-1997, poate fi o întreaga fraza, formata din mai multe clauze finite. *M*-1997 foloseste, de fapt, pentru segmentarea liniara a frazei în unitati de tip-clauzal numai *trei reguli* din

cele folosite de *SCD*-2002, iar aceste reguli sunt sintetizate de catre *actiunile* COMMA, SET_AND, SET_OR.

(5.2.d) *Actiunile* MATCH_PAREN, MATCH_DASH, COMMA_PAREN sunt utilizate de catre *M*-1997 pentru a delimita acele întinderi de text cate pot fi omise atunci când fraza este segmentata în unitati de tip-clauzal. Aceste parti "explicative" din text, considerate a nu fi importante, sunt, în text, puse între *paranteze*, (perechi de) *liniute-de-despartire*, sau (perechi de) *virgule*. Algoritmul *M*-1997 nu trateaza aceste întinderi "parantetizate" de text ca fiind unitati de tip-clauzal propriu-zise, ci le considera ca doar ca fiind scufundate în unitatea de tip-clauzal de care apartin. Pentru *SCD*-2002, aceste *actiuni M*-1997 nu au un corespondent specific deoarece *paranteza* (închisa si deschisa), *virgula*, si *liniuta-de-despartire* sunt tratate ca marcheri de discurs (M3), si fac parte din *gramatica de marcheri compusi* (concatenati) care este asociata cu algoritmul *SCD*-2002 de segmentare si dependenta a clauzelor dintr-o fraza.

(5.2.e) Din acelasi motiv ca cel mentionat mai sus, în (5.2.d), *actiunile* DUAL, NORMAL_THEN_COMMA din *M*-1997 nu au, nici ele, un corespondent în *SCD*-2002; aceste doua actiuni sunt de asemenea înglobate în *gramatica formala* de *secvente de marcheri de discurs*, care se dovedeste a fi, în mod clar, mai generala, usor de extins (sau de restrâns), este dependenta de LN specific analizat, si modeleaza comportamentul marcherilor simpli si compusi (concatenati) de tip M3 si M2.

Relatiile (5.2.1-2) si observatiile (5.2.a-e) demonstreaza ca algoritmul de segmentare *M*-1997 este (chiar strict) scufundat în algoritmii-*SCD* (atât *SCD*-2002 cât si *SCD*-1994). Acest fapt, stabilit teoretic aici, este confirmat de catre rezultatele empirice ale implementarilor, prezentate în subsectiunea care urmeaza.

5.3. Executia segmentarii pentru algoritmii *M*-1997 si *SCD*-2002

Actuala etapa de implementare a algoritmilor de segmentare este prezentata în exemplele care urmeaza. *Step06* din *SCD*-2002, si *Step12* din *SCD*-1994 stabilesc relatiile de *dependenta inter-clauzala*, folosind o gramatica formala pentru marcherii de discurs, simpli si compusi (concatenati), din clasele M3 si M2. Aceasta faza a algoritmului nu este înca implementata, în prezent. Sa mentionam ca stabilirea *dependentelor intra-clauzale* este (partial) implementata prin utilizarea, pentru moment, (numai) a subclaselor M2 si M1 de marcheri. Marginile inter-clauzale din text sunt reprezentate prin *paranteze patrate*, în timp ce pentru marginile si dependentele intra-clauzale sunt folosite *parantezele rotunde* (obisnuite). Indicii inferiori ai parantezelor patrate arata numarul curent al unitatilor textuale de tip-clauzal din algoritmul *M*-1997, respectiv numarul curent al clauzei obtinute din algoritmul *SCD*-2002.

Examplul 5.3.1.

Ex.5.3.1.Tag. (Etichetarea morfologica realizata cu mediul *TexTag* – vezi Fig. 5.4.1. si Fig. 5.4.2.)

<NSRY,23,0>Câmpul</NSRY,23,0> <V3,24,0>era verde</V3,24,0>
<CR,25,0>si</CR,25,0> <NSRY,26,0>vita</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</V3,30,0>
<S,31,0>cu</S,31,0> <NPN,32,0>lastari</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>soselei</NSOY,38,0> <V3,39,0>înfrunzisera</V3,39,0>
<CR,40,0>si</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 segmentarii (fara dependente), obtinut prin aplicarea algoritmului *M*-1997 în cadrul mediului *ClauSEGM* – vezi Fig. 5.4.3.)

[Câmpul era verde si vita de vie se acoperise cu lastari verzi, copacii de pe marginea soselei înfrunzisera si briza sufla dinspre mare.]$_1$

Ex.5.3.1.SCD. (Rezultatul segmentarii (fara dependente), obtinut prin aplicarea algoritmului *SCD*-2002 în cadrul mediului *ClauSEGM* – vezi Fig. 5.4.4.)

[(Câmpul) era verde]$_1$ si[(vita) (de (vie)) se acoperise (cu (lastari) (verzi))]$_2$,[(copacii) (de pe (marginea (soselei))) înfrunzisera]$_3$ si[(briza) sufla (dinspre (mare)).]$_4$

Examplul 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>ca</C,4,0> <V3,5,0>fulgera </V3,5,0> <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>noptile</NPRY,10,0> <V3,11,0>erau reci</V3,11,0> <CR,12,0>si</CR,12,0> <QZ,13,0>nu</QZ,13,0> <PPSD, 14,0>ti</PPSD,14,0> <PXA,15,0>se</PXA,15,0> <V3,16,0> parea</V3,16,0> <R,17,0>deloc</R,17,0> <C,18,0>ca</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]$_1$ [ca fulgera ca vara,]$_2$ [dar noptile erau reci si nu ti se parea deloc]$_3$ [ca se apropie furtuna.]$_4$

Ex.5.3.2.SCD.

[(În (întuneric)) ai fi zis]$_1$ [ca fulgera (ca (vara))]$_2$, [dar (noptile) erau reci]$_3$ si [nu (ti) se parea (deloc)]$_4$ [ca se apropie (furtuna).]$_5$

Example 5.3.3.

Ex.5.3.3.Tag.

<NSRY,46,0>Poarta</NSRY,46,0> <V3,47,0>era deschisa </V3,47,0> <COMMA,48,0>,</COMMA,48,0> <TSR,49,0>un</TSR,49,0> <NSN,50,0>soldat </NSN,50,0> <V3,51,0>sedea</V3,51,0> <S,52,0>la</S,52,0> <NSN, 53,0>soare< /NSN,53,0> <S,54,0>pe</S,54,0> <TSR,55,0> o</TSR,55,0> <NSRN,56,0> banca</NSRN,56,0><COMMA, 57,0>, </COMMA,57,0> <TSR,58,0>o</TSR,58,0> <NSRN, 59,0>ambulanta</NSRN,59,0> <V3,60,0>astepta</V3,60,0> <S,61,0>la </S,61,0> <NSRY, 62,0>usa</NSRY,62,0> <S, 63,0>de</S,63,0> <NSN,64,0> serviciu</NSN,64,0> <CR, 65,0>si</CR,65,0> <VG,66,0>intrând</VG,66,0> <V1,67,0> am simtit</V1,67,0> <NSRY,68,0>mirosul</NSRY,68,0> <NSOY,69,0> pardoselii</NSOY,69,0> <S,70,0>de</S,70,0> <NSRN,71,0>marmura</NSRN,71,0> <S,72,0>si</S,72,0> <S,73,0>de</S,73,0> <NSN,74,0>spital</NSN,74,0><POINT, 75,0>.</POINT,75,0>

Ex.5.3.3.Mar. (<u>întindere de text între paranteze acolade</u> {...})

[Poarta era deschisa, {un soldat sedea la soare pe o banca,} o ambulanta astepta la usa de serviciu si intrând am simtit mirosul pardoselii de marmura si de spital.]$_1$

Ex.5.3.3.SCD.

[(Poarta) era deschisa]$_1$,[(un (soldat)) sedea (la (soare) (pe (o (banca))))]$_2$,[(o (ambulanta)) astepta (la (usa) (de (serviciu)))]$_3$ si[intrând am simtit (mirosul (pardoselii)) (de (marmura) (si (de (spital)))).]$_4$

Example 5.3.4.

Ex.5.3.4.Tag.

<NPRY,1,0>Trupele</NPRY,1,0> <V3,2,0>treceau</V3,2,0> <S,3,0>pe lânga</S,3,0><NSRN,4,0>casa</NSRN,4,0><COMMA,5,0>,</COMMA,5,0> <S, 6,0>pe</S,6,0> <NSRN,7,0>sosea</NSRN,7,0><COMMA,8,0>,</COMMA,8,0> <CR,9,0>si</CR,9,0> <NSRY,10,0>praful</NSRY,10,0> <RELO,11,0>pe care</ RELO,11,0><Z,12,0>-</Z,12,0><PPSA,13,0>l</PPSA,13,0> <V3,14,0>ridicau</ V3,14,0> <PXA,15,0>se</PXA,15,0> <V3,16,0>asternea</V3,16,0> <S,17,0>pe </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 lânga casa, pe sosea, si praful]$_1$ [pe care-l ridicau se asternea pe frunzele copacilor.]$_2$

Ex.5.3.4.SCD. (clauza relativa – atributiva)

[(Trupele) treceau (pe lânga (casa)) , (pe (sosea))]$_1$,[si (praful) [pe care-(l) ridicau se asternea (pe (frunzele (copacilor)))]$_2$.]$_3$

Example 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 vazut</V3,79,0> <CR,80,0>si</CR,80,0> <QZ,81,0>n</QZ,81,0><Z, 82,0>-</Z,82,0><V1,83,0>am stiut</V1,83,0> <C,84,0>daca</C,84,0> <V3,85,0> e</V3,85,0> <NSRY,86,0>cazul</NSRY,86,0> <C,87,0>sa</C,87,0> <PPSA, 88,0>ma</PPSA,88,0> <V1,89,0>duc</V1,89,0> <S,90,0>la</S,90,0> <PPS,91,0> el</PPS,91,0> <C,92,0>sa</C,92,0><Z,93,0>-</Z,93,0><PPSA,94,0>i</PPSA, 94,0> <V1,95,0>raportez</V1,95,0> <C,96,0>ca</C,96,0> <V1,97,0>am sosit</ V1,97,0> <C,98,0>sau daca</C,98,0> <QZ,99,0>nu</QZ,99,0> <V3,100,0>e mai bine</V3,100,0> <C,101,0>sa</C,101,0> <PPSA,102,0>ma</PPSA,102,0> <V1, 103,0>duc</V1,103,0> <C,104,0>sa</C,104,0> <PPSA,105,0>ma</PPSA,105,0> <V1,106,0>aranjez</V1,106,0> <R,107,0>putin</R,107,0><POINT,108,0>.</POINT, 108,0>

Ex.5.3.5.Marc.

[Nu m-a vazut si n-am stiut $]_1$ [daca e cazul $]_2$ [sa ma duc la el $]_3$ [sa-i raportez $]_4$ [ca am sosit sau $]_5$ [daca nu e mai bine $]_6$ [sa ma duc $]_7$ [sa ma aranjez putin. $]_8$

Ex.5.3.5.SCD.

[Nu (m)-a vazut $]_1$ si[n-am stiut $]_2$ [daca e (cazul) $]_3$ [sa (ma) duc (la (el)) $]_4$ [sa-(i) raportez $]_5$ [ca am sosit $]_6$ [sau daca nu e mai bine $]_7$ [sa (ma) duc $]_8$ [sa (ma) aranjez (putin). $]_9$

Example 5.3.6.

Ex.5.3.6.Tag.

<NSRY,109,0>Fereastra</NSRY,109,0> <V3,110,0>era deschisa</V3, 110,0><COMMA,111,0>,</COMMA,111,0> <NSRY,112,0>patul</NSRY,112,0> <PSS,113,0>meu</PSS,113,0> <V3,114,0>era acoperit</V3,114,0> <S,115,0> cu</S,115,0> <NSRY,116,0>patura</NSRY,116,0><COMMA,117,0>,</COMMA, 117,0> <NSRY,118,0>masca</NSRY,118,0> <S,119,0>de</S,119,0> <NPN, 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>lunguiata</ASN,124,0> <S,125, 0>de</S,125,0> <NSRN,126,0>tinichea</NSRN,126,0> <CR,127,0>si</CR,127, 0> <NSRY,128,0>casca</NSRY,128,0> <S,129,0>de</S,129,0> <NSN,130,0> otel</NSN,130,0> <V3,131,0>erau agatate</V3,131,0> <S,132,0>pe</S,132,0> <DMSR,133,0>acelasi</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 deschisa, {patul meu era acoperit cu patura}, masca de gaze cu cutia ei lunguiata de tinichea si casca de otel erau agatate pe acelasi cuier. $]_1$

Ex.5.3.6.SCD.

[(Fereastra) era deschisa]$_1$,[(patul) (meu) era acoperit (cu (patura))]$_2$,[(masca) (de (gaze) (cu (cutia) (ei (lunguiata)) (de (tinichea)))) si (casca) (de (otel)) erau agatate (pe (acelasi (cuier))).]$_3$

Example 5.3.7.

Ex.5.3.7.Tag.

<V1,1,0>As vrea</V1,1,0> <C,2,0>sa</C,2,0><Z,3,0>-</Z,3,0><PPSD, 4,0>ti</PPSD,4,0> <V1,5,0>spun</V1,5,0> <C,6,0>ca</C,6,0> <CR,7,0>si</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>si</CR,11,0> <S,12,0>într</S,12,0><Z,13,0>-</Z,13,0><ASN,14,0>alta </ASN,14,0> <NSRN,15,0>parte</NSRN,15,0><COMMA,16,0>,</COMMA,16,0> <V1,17,0>am vazut</V1,17,0> <C,18,0>ca</C,18,0> <NPRY,19,0>lucrurile</ NPRY,19,0> <PXA,20,0>se</PXA,20,0> <V3,21,0>întâmpla</V3,21,0> <R,22,0> tot asa</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,27, 0> <NPN,28,0>precizari</NPN,28,0> <CR,29,0>si</CR,29,0> <V1,30,0>simt</ V1,30,0> <C,31,0>ca</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</V3,35,0> <NSRY,36,0>respiratia< /NSRY,36,0><COMMA,37,0>,</COMMA,37,0> <C,38,0>ca</C,38,0> <V1,39,0>as ocoli</V1,39,0> <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 {...})

[As vrea]$_1$ [sa-ti spun]$_2$ [ca si mai târziu, {si într-alta parte,} am vazut]$_3$ [ca lucrurile se întâmpla tot asa,]$_4$ [dar ar fi nevoie de oarecari precizari si simt]$_5$ [ca nu mi-ar ajunge respiratia,]$_6$ [ca as ocoli prea mult.]$_7$

Ex.5.3.7.SCD.

[As vrea]$_1$ [sa-(ti) spun]$_2$ [ca si (mai (târziu)) , si (într-(alta (parte))) , am vazut]$_3$ [ca (lucrurile) se întâmpla (tot asa)]$_4$,[dar ar fi nevoie (de (oarecari (precizari)))]$_5$ si[simt]$_6$ [ca nu (mi)-ar ajunge (respiratia)]$_7$,[ca as ocoli (prea (mult)).]$_8$

5.4. Programele *TexTag* şi *ClauSEGM*

În cele ce urmeaza sunt prezentate câteva imagini de executie în cadrul programelor *TexTag* si *ClauSEGM*, scrise în Visual C++ 5.0, si utilizate pentru a eticheta si segmenta texte de LN (limba româna). Figurile 5.4.1. si 5.4.2. se refera la *TexTag*, Figura 5.4.3. contine executia algoritmului de *segmentare M*-1997 în cadrul *ClauSEGM*, iar Figura 5.4.4. contine o executie a algoritmului de *segmentare SCD*-2002 sub mediul *ClauSEGM*. Stabilirea relatiilor de *dependenta inter*- si *intra-clauzale*, pentru aceleasi doua tipuri de algoritmi, urmeaza sa fie implementata în cadrul aceluiasi mediu *ClauSEGM*.

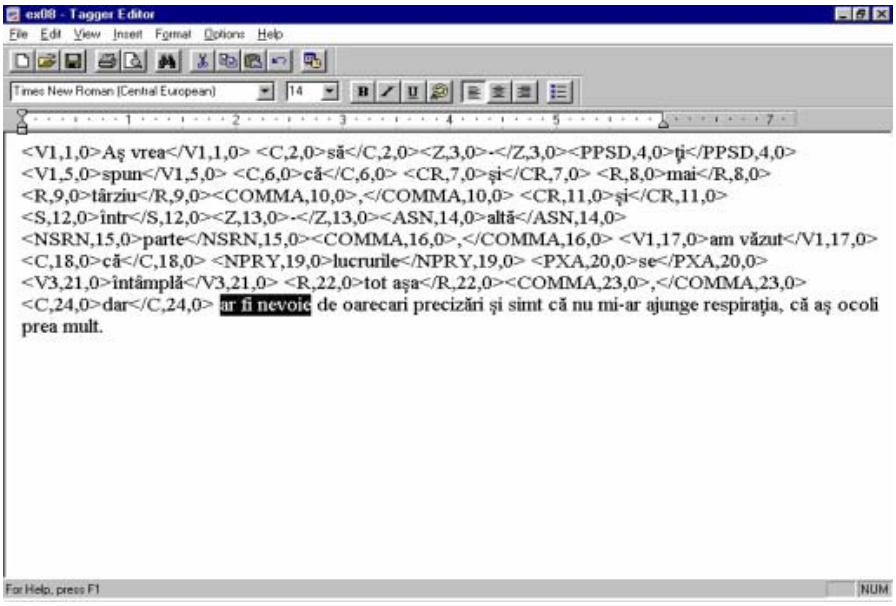


Figura 5.4.1. Rezultatul etichetarii morfologice sub *TexTag*

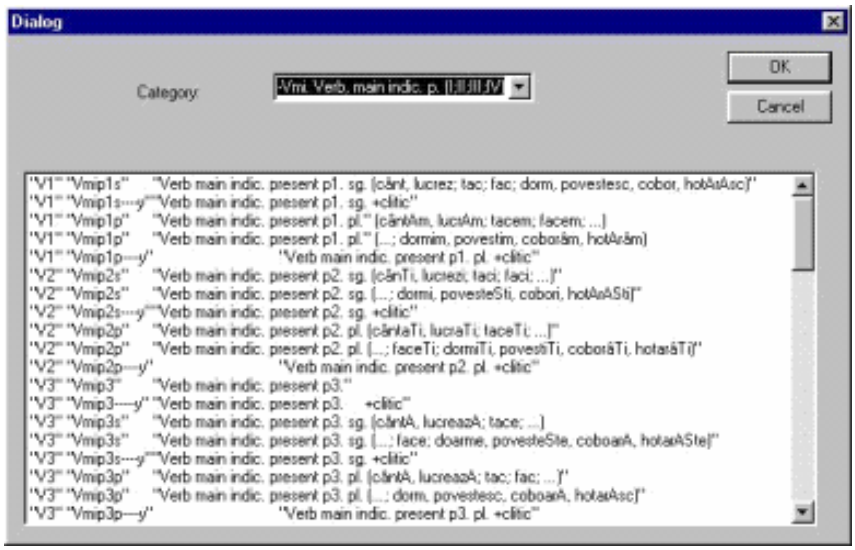


Figura 5.4.2. Lista de etichete selectata cu un meniu din *TexTag*

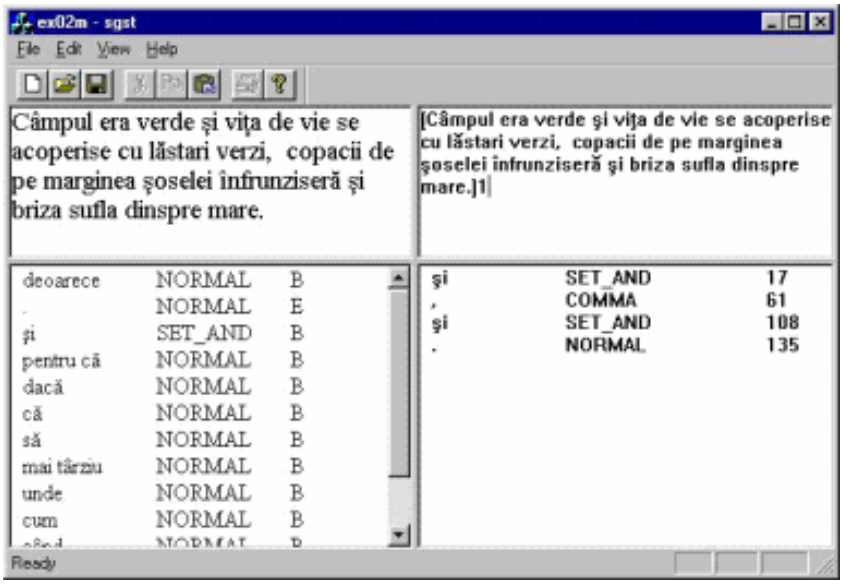


Figura 5.4.3. O executie a algoritmului de segmentare *M*-1997 cu programul *ClauSEGM*

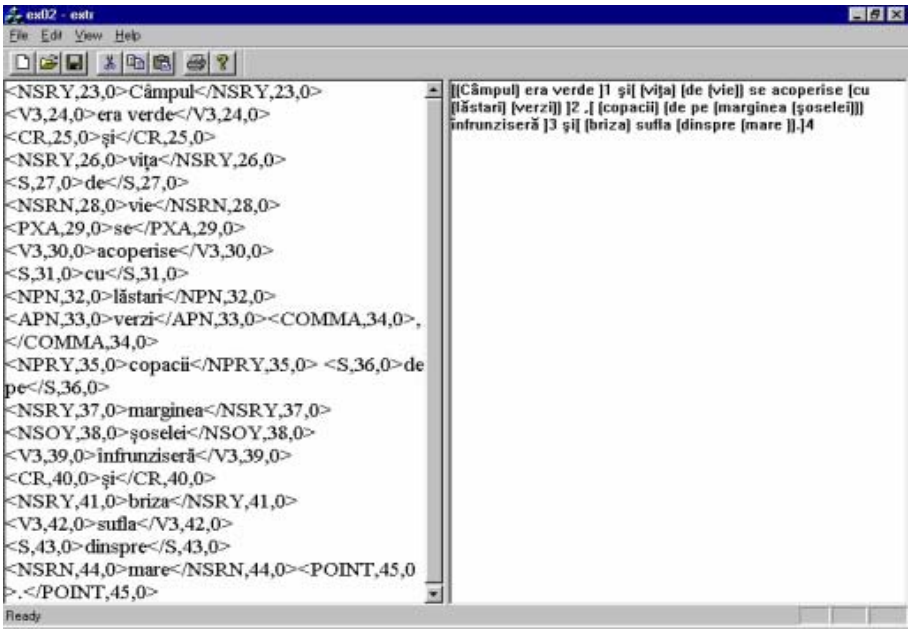


Fig. 5.4.4. O execuţie a algoritmului de segmentare *SCD*-2002
sub mediul *ClauSEGM*

6. Concluzii

Rezultatele obtinute în aceasta lucrare nu se refera strict la compararea si implementarea celor doi algoritmi de segmentare. Avem, de fapt, doua tipuri de algoritmi de segmentare (si dependenta), si fiecare din cele doua tipuri reprezinta linii specifice de cercetare, cu importante consecinte asupra domeniilor de procesare a LN carora se adreseaza: algoritmul *M*-1997 este destinat (teoriei si) aplicatiilor de procesare a discursului, generare automata a LN, si rezumarii automate, în timp ce algoritmul *SCD*-2002 se încadreaza mai curând în teorii sintactice ale LN, cum sunt teoria FX-bar [3], parsarea bazata pe teorii (principii) sintactico-semantice ale LN, dar si punerea în evidenta a structurilor (segmentelor) si relatiilor de discurs [6].

Demonstrarea relatiei (de scufundare) dintre cele doua tipuri de algoritmi de segmentare, schitarea (în sectiunea 1) a unui *cadru formal general* pentru *algoritmii de segmentare* a LN, în particular a segmentarii de tip *chunking*, propunerea (în cadrul algoritmilor-*SCD*) unei metode generale de segmentare în unitati textuale a LN si de stabilire a dependentelor între ele, toate acestea constituie posibile noi perspective pentru

abordarile teoretice si aplicative curente în procesarea automata a LN, inclusiv, si mai ales, pentru limba româna.

Revenind la aspectele concrete expuse în acest articol, extinderea algoritmilor catre analiza complexa a structurilor semantico-discursive antrenate de clasele de marcheri, si perfectionarea actualelor implementari ramân principalele directii de continuare a prezentei abordari.

Referinte bibliografice

[1] Neculai Curteanu (1994). *From Morphology to Discourse Through Marker Structures in the SCD Parsing Strategy*, Language and Cybernetics, Akademia Libroservo, Prague, p. 61-73.

[2] N. Curteanu, G. Holban (1996). *Strategia lingvistica SCD aplicata la analiza si generarea limbii române*, Limbaj si Tehnologie (D. Tufis, Ed.), Editura Academiei Române, p. 169-176.

[3] Neculai Curteanu (2000). "*Towards a Functional X-bar Theory*", Technical Report, Institute of Theoretical Informatics, Romanian Academy, Iasi Branch, 32p.

[4] Daniel Marcu (1997). *The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts*, Ph.D. Thesis, Univ. of Toronto, Canada, 331 p.

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

[6] O. Popârda, N. Curteanu (2002). *L'évolution du discours juridique français analysé par la stratégie linguistique SCD*, LINCOM Studies in Theoretical Linguistics, Lincom Europa (va apare).

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

[8] N. Curteanu, D. Cristea, P. Mihaescu (1982). *Cercetari în domeniul comunicarii om-calculator prin intermediul limbajului natural*. Contract de cercetare nr. 4774/1982, Universitatea Iasi - ICI Bucuresti.

[9] Neculai Curteanu (1983). *Algoritmi de analiza sintactica a frazei si propozitiei românesti*. Lucrarile Conferintei INFO-IASI'83, p. 553-548.

[10] N. D. Cristea, N. Curteanu, P. Mihaescu (1983). *Implementarea analizorului morfologic si definitivarea proiectului de analiza sintactica*. Contract de cercetare nr. 1906/1983, Universitatea Iasi - ICI Bucuresti.

[11] N. Curteanu (1984). *Aspecte ale analizei logice a limbajului natural*. Contract de cercetare nr. 4709/1984, Universitatea Iasi - ICI Bucuresti.

[12] Rebecca Passonneau, Diane Litman (1997). *Intention-based segmentation*: *human*

reliability and correlation with linguistic cues, in Proc. 31th 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 sémantique des facteurs hétérogcnes qui régissent la distribution des adjectifs*, Conferinta Internationala "Representations du Sens Linguistique", Bucuresti.

[17] Dumitru Irimia (1997). *Morfo-sintaxa verbului românesc*. Editura Universitatii "Al. I. Cuza", Iasi.

[18] Eva Hajicova, H. Skoumalova, P. Sgall (1995). *An Automatic Procedure for Topic-Focus Identification*. Computational 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 Tufis (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 Tufis, A.M. Barbu (2001). *Computational bilingual lexicography: automatic extraction of translation dictionaries*, In Romanian Journal on Information Science and Technology, vol. 4, no. 3.

O metoda automata pentru inserarea diacriticelor în texte în limba româna

Rada F. MIHALCEA
University of Texas at Dallas, Richardson, Texas, U.S.A.
rada@utdallas.edu

Vivi A. NASTASE
University of Ottawa, Ottawa, Canada
vnastase@site.uottawa.ca

1. Introducere

Problema restaurarii diacriticelor consta în inserarea diacriticelor într-un text în care lipsesc. Cresterea continua a numarului de texte disponibile prin Internet face ca metodele automate de inserare a diacriticelor sa devina o componenta esentiala în multe aplicatii importante, cum ar fi extragerea de informatii, traducerea automata, colectionarea de texte, construirea dictionarelor electronice si multe altele. Corectarea erorilor ortografice poate sa aiba un impact major asupra calitatii rezultatelor obtinute în aceste aplicatii. De exemplu, în absenta unei metode de restaurare a diacriticelor, unele cuvinte devin ambigue, cum este cazul cuvintelor din limba româna *peste, peste* sau *paturi, paturi*. O cautare bazata pe astfel cuvinte poate returna multe texte irelevante (de exemplu, o cautare pentru *peste* ar returna si documente continând *peste*). De asemenea, traducerea unor astfel de cuvinte într-o limba straina poate fi eronata (de exemplu, traducerea corecta a cuvântului *paturi* în limba engleza este *blankets*, dar în absenta diacriticei este tradus gresit ca si *beds*).

Metodele dezvoltate pâna în prezent pentru rezolvarea acestei probleme se bazeaza în general pe dictionare si pe diverse procesoare lexicale si/sau sintactice. Multe dintre limbile care se confrunta cu problema restaurarii diacriticelor nu beneficiaza însa de astfel de resurse, si ca urmare aplicabilitatea acestor metode este limitata la limbi bine studiate care dispun de suficiente resurse. Lucrarea de fata prezinta o metoda automata de reinserare a diacriticelor în text care necesita doar o colectie de texte de dimensiuni modeste. Spre deosebire de alte metode dezvoltate anterior, metoda introdusa în aceasta lucrare nu necesita nici un fel de dictionare sau procesoare morpfologice si/sau sintactice, si prin urmare poate fi folosita pentru prelucrarea de texte în orice limba care dispune de un numar minim de texte cu diacritice. Datorita lipsei de restrictii, metoda propusa este foarte generala si usor aplicabila pentru orice limba. Pentru a demonstra aceasta afirmatie, dupa

ce vom prezenta experimentele pentru texte în limba româna, vom arata câteva rezultate obtinute pentru limbile ceha, poloneza si maghiara.

2. Experimente anterioare

Restaurarea diacriticelor nu este în sine o problema dificila. Experimente efectuate pâna în prezent au demonstrat ca folosirea de dictionare electronice poate duce la o acuratete de peste 90% în restaurarea accentelor pentru limbile franceza si spaniola [9],[11],[5]. Metoda descrisa de Michael Simard în [9] este o îmbunatatire adusa unei metode propusa anterior de El-Bčze [4]. Aceasta metoda se bazeaza pe Hidden Markov Models si învata folosind cuvintele învecinate. Precizia raportata este de 99%. Tufis si Chitu [10] propun o metoda similara pentru inserarea diacriticelor în texte în limba româna. Yarowsky prezinta în [11] un set de metode folosite pentru restaurarea accentelor in limbile franceza si spaniola. Majoritatea algoritmilor pe care îi prezinta se bazeaza pe dictionare si cuvinte învecinate pentru a decide asupra ortografiei potrivite pentru fiecare cuvânt ambiguu. Yarowsky compara N-gram taggers, clasificatoare Bayesiene si liste de decizii cu metoda de baza care consta în folosirea unui dictionar. Pentru cele doua limbi considerate în experimentele raportate, listele de decizii duc la performantele cele mai ridicate. Toate aceste tehnici se bazeaza însa pe context, dictionare, si în unele cazuri pe informatii aditionale de natura morfologica si sintactica. Nagy et al. prezinta în [7] o abordare diferita a problemei, în care siruri de litere sunt extrase din fiecare cuvânt si folosite pentru a obtine statistici. Folosind metoda propusa, s-a observat o precizie foarte buna obtinuta pe texte în limba franceza. Experimentele prezentate in [7] sunt asemanatoare cu cele raportate în [1], unde masuri de similaritate între trigrame sunt folosite pentru a automatiza corectarea greselilor de ortografie.

Majoritatea studiilor efectuate pâna în acest moment pe aceasta tema, s-au ocupat de limbi bine cunoscute si raspândite, cum ar fi franceza si spaniola. Foarte putine studii s-au concentrat pe limbi mai putin mediatizate cum ar fi ceha, slovena, turca sau alte limbi care folosesc diacritice. Tabelul 1[77] prezinta diacriticele folosite în limbile europene cu alfabet latin. Dupa cum rezulta din aceasta lista, numeroase limbi se confrunta cu problema restaurarii diacriticelor. Din setul de 36 de limbi cuprinse în tabel, engleza pare sa fie singura limba pentru care diacriticele nu constituie o problema. Cuvintele din engleza care contin diacritice au fost împrumutate din alte limbi, si varianta acestora fara diacritice nu are un corespondent care sa duca la ambiguitate. Diacriticele par însa sa aiba un rol

[77] *Tabelul cuprinde numai litere mici. Fiecarei litere mici îi corespunde o litera mare. Informatia din acest tabel a fost agregata din liste de diacritici în limbi europene, disponibile la adresa www.tiro.com/di_intro.html*

important în diferentierea cuvintelor. Engleza, care dupa cum spuneam nu are diacritice per se, are în schimb o ambiguitatea semantica mai ridicata[78]

Tabel 1

Diacritice din limbile europene cu alfabet latin

Limba	Diacritice	Limba	Diacritice
Albaneză	ç ë	Malteză	c g h ż
Bască	ń ü	Norvegiană	á ć ř
Bretonă	â ę ń ů ü	Olandeză	á ŕ â ä é č ę ë i í ě î ď ó ň ô ö ú ů ű ü
Catalană	ŕ ç č é í ď ? ň ó ú ü	Poloneză	ą ć ę ł ń ó ś ź ż
Cehă	á č ď é e í ň ó ř š ť ú ů ý ž	Portugheză	â ă ç ę ó ô ő ö
Daneză	Í ć ř	Română	â ă î ş ţ
Engleză	None	Sami (Laponă)	á ď č đ ń ņ š t ž
Estoniană	ä č õ ö š ü ž	Serbo-croată	ć č đ š ž
Faroeză	á ć đ í ó ř ú ý	Slovacă	á ä č đ é í ľ ń ó ô ŕ š ť ú ý ž
Finlandeză	ä Í ö š ž	Slovena	č š ž
Franceză	ŕ â ć ç č é ę ë î ď ô ś ů ű	Spaniolă	á é í ó ú ü ń
Galiţiană	á é í ó ú	Suedeză	ä Í ö
Germană	ä ö ü ß	Turcă	ç ğ ö ş ü
Islandeză	á ć đ é í ó ö ú ý ţ	Sorbiană (1)	ć č e ł ń ŕ ś š ź ž
Italiană	ŕ é č í ě ď ó ň ú ů	Sorbiană (2)	ć č e ł ń ó ř š ž
Maghiară	á é í ó ö ő ú ü ű	Welsh	â ę î ô ű w y

Aplicabilitatea metodelor mentionate anterior este limitata în urmatoarele cazuri:

1. Dictionarele electronice nu sunt disponibile, sau doar dictionare de dimensiuni relativ mici sunt facute publice. Mai mult decât atât, în cazul în care dictionarul însusi nu are diacritice, metodele care se bazeaza pe aceasta resursa pentru restaurarea diacriticelor devin inaplicabile.

[78] *Studii efectuate pe corpusuri bilingve paralele, ar arătat că vocabularul construit dintr-un text în limba engleză este aproximativ jumătate din vocabularul construit pe baza aceluiaşi text într-o altă limbă. Competiţia SENSEVAL [6] raportează de asemenea precizii mult mai mici pentru engleză comparativ cu alte limbi în rezolvarea ambiguităţii semantice. Lipsa diacriticelor în limba englezã ar putea constitui o explicaţie a acestui fenomen.*

2. Procesoarele folosite pentru analiza morfologica si/sau sintactica, considerate folositoare pentru problema restaurarii diacriticelor, nu exista sau nu sunt public disponibile.
3. Numarul de texte disponibile continând diacritice este relativ mic. Marimea corpusurilor publice sau disponibile prin Internet influenteaza marimea vocabularului care poate fi construit ad-hoc pe baza acestor texte. În plus, majoritatea siturilor care publica texte pe Internet prefera în multe cazuri sa evite diacriticele din motive de simplitate, uniformitate, sau pur si simplu lipsa de mijloace necesare pentru codificarea diacriticelor.

Lucrarea de fata prezinta o metoda de restaurare a diacriticelor bazata pe învatarea la nivel de litera, si nu la nivel de cuvânt. Avantajul principal al acestei metode este faptul ca ofera posibilitatea de generalizare dincolo de cuvinte. Metoda este folositoare mai ales pentru limbile pentru care resursele disponibile sunt limitate, în speta limbi care nu au dictionare electronice mari cu diacritice. Limbi cunoscute si bine studiate, precum franceza si spaniola, pot de asemenea beneficia de aceasta metoda pentru procesarea cuvintelor necunoscute.

Experimentele prezentate în aceasta lucrare adreseaza în principal problema restaurarii diacriticelor în texte în limba româna. Precizia observata pe limba româna este de 99%, masurata la nivel de litera. Experimente similare au fost efectuate pe alte trei limbi, si anume poloneza, maghiara si ceha, de asemenea cu rezultate foarte bune. Avantajul principal al metodei este faptul ca nu necesita nici o etapa de preprocesare, ci numai un corpus relativ mic format din texte cu diacritice. Datorita simplitatii algoritmului, viteza de procesare este foarte mare, de aproximativ 20 pagini de text pe secunda, masurata pe un calculator cu un procesor Pentium III cu frecventa de 500MHz si 250MB memorie.

Practic, metoda propusa încearca sa învete reguli aplicabile la nivel de litera. În loc de a învata reguli care se aplica la nivel de cuvânt, cum ar fi *„anuncio se scrie anunció atunci când are functia de verb”*, dorim sa învatam reguli aplicabile la nivel de litera, cum ar fi *„s urmat de i si spatiu si precedat de spatiu se scrie s”*. Astfel de reguli, învatate la nivel de litera, sunt mai generale si au aplicabilitate mai mare, în special în cazurile în care dictionarele disponibile sunt de dimensiune redusa, când se întâlnesc multe cuvinte necunoscute în textul dat, sau când procesoare pentru analiza morfologica sau sintactica nu sunt la îdemâna.

Este evident ca în analiza limbajului literele constituie nivelul cu granularitatea cea mai scazuta, si de aceea au si cel mai mare potential de generalizare. În loc de aproximativ 150.000 de unitati candidate potentiale pentru algoritm (marimea aproximativa a vocabularului de uz general a unei limbi), vom avea mai mult sau mai putin 26 caractere pe baza carora se vor constitui datele de intrare pentru algoritmul de dezambiguare[79].

[79] *Numãrul de litere depinde de limba care se analizeazã. S-a arãtat de exmplu cã aproximativ 85% dintre cuvintele în limba francezã nu au o formã ortograficã cu diacritice, ºi deci numai 20.000 de cuvinte sunt potenþial ambigue. Pe de altã parte, numai 7 litere sunt ambigue în limba francezã.*

3. Experimente

Scopul experimentelor descrise în aceasta lucrare este de a arata ca învatarea la nivel de litera este posibila si poate rezolva, cu precizie mare, problema restaurarii diacriticelor. Pe lânga faptul ca metoda propusa constituie o problema de cercetare, scopul învatarii la un nivel de granularitate atât de scazut este de a oferi o metoda viabila pentru limbile pentru care resursele lexicale si semantice disponibile sunt limitate, si pentru care restaurarea diacriticelor prin învatare la nivel de cuvânt este greu de realizat.

3.1. Date

Prezentam în primul rând experimentele efectuate pe texte în limba româna. Limba româna nu este o limba foarte raspândita, si în consecinta nu are foarte multe resurse public disponibile pentru pre-procesare, iar dictionarele electronice sunt de dimensiuni relativ mici. În al doilea rând, am avut de rezolvat o problema specifica de restaurare a diacriticelor într-un dictionar electronic român-englez care contine aproximativ 75.000 de cuvinte, dar are dezavantajul ca diacriticele lipsesc. Am considerat ca este avantajos sa studiem problema restaurarii diacriticelor si sa folosim acest dictionar, în loc sa ne bazam pe alte dictionare cu diacritice de dimensiuni reduse. În plus, pentru procesoarele pe care am dori sa le dezvoltam pentru limba româna avem nevoie de numeroase texte electronice în limba româna. De obicei aceste texte nu au diacritice, si deci reinserarea diacriticelor este din nou necesara. Avem de asemenea posibilitatea de a compara eficacitatea acestei metode cu rezultate obtinute în experimente efectuate pe aceiasi limba constând în metode în care învatarea se face la nivel de cuvânt [10].

Pentru a aplica metoda descrisa în lucrarea de fata, avem deci nevoie de o colectie de texte românesti cu diacritice. În acest scop, am colectat articole din „România Literara”[80], un ziar românesc publicat saptamânal, cu articole legate în general de literatura. Ziarul are o versiune care contine diacritice începând din anul 2000. Colectia disponibila on-line la data colectarii datelor (august 2001) cuprindea 2780 articole. În pasul urmator, textul a fost extras din fisierele HTML. Atentie deosebita a fost acordata doar caracterelor românesti. Alte caractere cu diacritice întâlnite ocazional, cum ar fi c, é, etc. au fost transformate în forma lor echivalenta, fara diacritice, având în vedere ca suntem interesati doar de caracterele românesti, si nu de caractere franceze sau din alte limbi. Dupa toate aceste faze premergatoare, am obtinut un corpus continând aproximativ 3 milioane de cuvinte.

Literele mari au fost transformate în litere mici. Cazul literelor *â* si *î* este special în limba româna: desi pronuntia lor este identica, folosirea lor este guvernata de reguli bazate pe pozitia lor în cuvânt. La începutul cuvântului se foloseste întotdeauna *î*, iar *â* se foloseste în interiorul cuvântului. Este bine cunoscut faptul ca folosirea acestor litere a fost controversata de-a lungul timpului. O lege din anii `60 a schimbat ortografierea de la *â* la *î* , singura exceptie fiind cuvintele derivate din radacina *român*. La începutul anilor `90 ortografia veche a fost reintrodusa, si astfel s-a ajuns la cazuri de texte inconsistente, în care se întâlnesc scrieri

[80] *Accesibil prin http://www.romlit.ro*

diferite ale aceluiasi cuvânt. De exemplu, *cîntec* si *cântec* sunt forme ale aceluiasi cuvânt care pot fi întâlnite în acelasi text. Ziarul „*România Literara"* pastreaza înca ortografia cu *î*, cu mici exceptii (de exemplu, articole scrise de scriitori invitati care prefera sa scrie folosind *â* în loc de *î*).

3.2. Algoritmi de învatare

Pentru a rezolva problema restaurarii diacriticelor, am ales sa folosim un algoritm bazat pe învatarea de instante (IBL). Exista doua motive importante care au stat la baza luarii acestei decizii. În primul rând, este faptul demonstrat ca exceptiile au un rol important în procesarea limbajelor naturale. Algoritmii de tip IBL sunt recunoscuti pentru faptul ca iau în considerare fiecare exemplu de antrenament în luarea unei decizii de clasificare [2], si deci folosirea acestui tip de algoritmi prezinta un avantaj deosebit în probleme de limbaj natural. În al doilea rând, acest gen de algoritmi sunt foarte eficienti relativ la timpul de antrenament si testare.

Învatarea pe baza de instante se desfasoara în felul urmator: în pasul de antrenament, toate exemplele de intrare sunt memorate. În faza de testare, fiecare exemplu din set este comparat cu exemplele memorate, si va primi clasificarea data de exemplul memorat de care este cel mai apropiat, distanta fiind data de masura specifica aleasa în implementarea folosita. Pentru efectuarea experimentelor propuse, am folosit implementarea TiMBL [3] a acestor algoritmi. În plus, am efectuat experimente asemanatoare si cu un clasificator pe baza de arbori de decizie, si anume C4.5 [8]. Arborii de decizie sunt construiti din setul de exemple de antrenament. La fiecare pas este ales un atribut care discrimineaza cel mai bine exemple din clase diferite (prin valorile sale). Grupele obtinute prin diviziunea dupa acest atribut vor fi din nou împartite în grupe mai mici si mai pure, prin alegerea unui nou atribut care discrimineaza cel mai bine exemplele din grupa. Acest proces continua pâna când grupele obtinute au un grad de puritate acceptabil, sau marimea arborelui depaseste un prag ales initial. Rezultatele obtinute folosind C4.5 sunt asemanatoare cu cele obtinute folosind TiMBL, însa C4.5 are capacitatea de a genera reguli expresive, folositoare pentru implementari practice.

Având în vedere ca lucram la nivelul literelor, atributul care trebuie învatat este constituit de litera ambigua. Acesta poate fi oricare din literele ambigue enumerate în Tabelul 1. Pentru limba româna avem 4 perechi de litere ambigue: *s - s, t - t, a - a, i - î*. Literele mari au fost convertite în prealabil în litere mici. Datorita faptului ca datele folosite aplica ortografia cu *î*, nu avem ambiguitatea *a - â*, ci doar ambiguitatea *i - î*. Aceasta nu implica însa o pierdere de generalitate. Conversia între cele doua forme de ortografie este simpla si se poate realiza folosind doar pozitia literei în cuvânt, si prin urmare scrierile diferite nu afecteaza rezultatul algoritmului.

3.3. Atribute

Atributele folosite în orice algoritm de învatare au un impact foarte mare asupra eficacitatii algoritmului. Dupa cum am mentionat si în introducere, nu avem posibilitatea de a folosi procesoare care determina partea de vorbire a cuvintelor, si nici un alt fel de analizoare morfologice sau sintactice. În plus, nu dorim sa ne bazam pe cuvintele

învecinate, deoarece avem un numar limitat de date, si în consecinta exista sansa de a întâlni un numar mare de cuvinte necunoscute. Prin urmare, ne-am decis asupra folosirii unor atribute foarte simple, pentru extragerea carora nu este nevoie de nici un fel procesare speciala. Vom folosi litere învecinate, cu o notatie speciala atribuita spatiilor, virgulelor si punctelor (aceste caractere pot afecta procesul de învatare, fiind considerate caractere speciale de catre C4.5 si/sau TiMBL).

Daca X este litera a carui ambiguitate trebuie rezolvata, atributele folosite sunt N litere la stânga si la dreapta literei ambigue:

$$L_{-N}, L_{-(N-1)}, ..., L_{-1}, X, L_1, L_2, ..., L_{(N-1)}, L_N$$

Acest set de atribute se comporta surprinzator de bine, relativ la acuratete, dupa cum vom arata in cele ce urmeaza.

Dupa cum am mentionat anterior, am ales sa nu ne bazam pe nici un tag obtinut cu procesoare lexicale sau morfologice, ci doar pe informatia care se poate extrage din text neprelucrat. De asemenea, suntem interesati sa gasim posibilitati de generalizare, astfel încât un corpus limitat sa poata fi folosit pentru a genera reguli de reinserare a diacriticelor. În loc de a învata reguli bazându-ne pe cuvinte, dupa cum s-a procedat pâna acum, dorim sa învatam reguli bazate pe litere, pentru ca acestea constituie cele mai mici unitati în limbaj, si ofera posibilitatea învatarii chiar si dintr-o colectie mica de texte.

Pentru fiecare pereche ambigua de litere, parcurgem textul si generam toate exemplele posibile întâlnite în text. Atributele într-un exemplu sunt formate folosind N litere la stânga si la dreapta literei ambigue, si atributul tinta este însasi litera ambigua. Forma generala a exemplelor generate este:

$$L_{-N}, L_{-(N-1)}, ..., L_{-1}, L_1, L_2, ..., L_{(N-1)}, L_N, X$$

unde ca si în exemplul anterior, X este litera ambigua. Prezentam mai jos exemple de vectori de atribute care constituie date de intrare pentru algoritmul de învatare pentru rezolvarea ambiguitatii perechii *s - s*. CO, DO si SP sunt codurile care înlocuiesc virgula, punctul si spatiul.

l , 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.

Învatarea se reduce la detectarea corelatiilor între valorile atributelor care caracterizeaza exemplele de antrenament si valorile atributelor tinta, si utilizarea acestora pentru stabilirea valorii atributului tinta din exemplele de testare.

Numarul de exemple extrase din corpus depinde de perechea de litere. Din întregul set de 3 milioane de cuvinte, am obtinut 2.161.556 exemple pentru perechea ambigua *a - a*, 2.055.147 pentru perechea *i - î*, 1.257.458 exemple pentru *t - t*, si în final 866.964 exemple pentru perechea *s - s*. În fiecare din aceste cazuri, spatiul exemplelor este împartit în doua

clase, date de cele 2 variante ale literei ambigue. Metoda de învatare automata va folosi atributele date pentru a gasi reguli de clasificare a exemplelor în cele 2 clase.

3.4. Rezultate

Precizia cea mai ridicata s-a obtinut pentru o fereastra de 10 litere în vecinatatea literei ambigue (N = 5). Data fiind aceasta observatie, am considerat ca este important sa studiem mai in detaliu acest caz, si sa determinam ratele de învatare pentru cele 4 perechi de litere ambigue. Cu toate acestea, prezentam rezultate pentru ferestre de diverse dimensiuni, pentru comparatie.

Tabelul 2 arata rezultatele obtinute pentru N=5. Preciziile raportate în acest tabel sunt obtinute folosind algoritmul bazat pe învatarea de instante. Am condus experimente cu seturi de antrenament de diverse dimensiuni, variind de la 2.000.000 exemple pâna la 10 exemple, pentru a determina rata de învatare si dimensiunea minima a corpusului necesara pentru a obtine o precizie satisfacatoare. În toate aceste experimente s-au folosit seturi de testare continând 50.000 exemple. Pentru a obtine rezultate cât mai acurate am folosit validare încrucisata folosind 10 seturi diferite de test. Tabelul indica de asemenea baza de comparatie, definita aici ca fiind precizia obtinuta când se foloseste implicit litera cea mai frecventa din fiecare pereche ambigua.

Rezultatele prezentate în Tabelul 2 sunt reprezentate grafic în Figura 1. Este interesant de observat ca cea mai importanta faza a procesului de învatare are loc când se folosesc primele 10 000 exemple. În conformitate cu masuratorile efectuate, a rezultat ca 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 obtine îmbunatatiri de numai 1% este necesar un numar semnificativ de exemple. Tabelul 2 indica de asemenea, în caractere groase, prima precizie care depaseste baza de comparatie, ca o indicatie a dimensiunii minime a setului de antrenament pentru care se observa o forma minima de învatare. Dupa cum se observa din tabel, si numai 1.000 exemple sunt suficiente pentru învatare.

Tabel 2

Rezultate obtinute în rezolvarea ambiguitatii literelor cu diacritice în limba româna

	Pereche ambiguă				
	a - a	a - a(2)	i - î	s - s	t - t
Nr.total exemple	2.161.566	1.369.517	2.055.147	866.964	1.157.458
Baza comparație	74.70%	85.90%	88.205	76.53%	85.81%
Exemple de Antrenament	Precizie obținuta pe date de test (50.000 exemple)				
2,000,000	96.14%	-	99.69%	-	-
1,000,000	95.10%	99.14%	99.58%	-	98.75%
750,000	94.83%	98.97%	99.53%	99.07%	98.63%
500,000	94.57%	98.79%	99.46%	98.86%	98.40%

250,000	94.00%	98.37%	99.28%	98.87%	98.26%
100,000	93.03%	97.56%	98.96%	98.54%	97.81%
50,000	92.10%	96.86%	98.57%	98.13%	97.40%
25,000	90.99%	95.75%	98.11%	97.58%	96.92%
10,000	88.99%	93.75%	97.31%	96.53%	96.20%
5,000	87.56%	92.76%	96.65%	95.61%	95.10%
4,000	86.91%	91.86%	96.49%	94.99%	94.53%
3,000	86.39%	90.99%	96.19%	94.18%	94.30%
2,000	85.81%	89.93%	95.49%	93.47%	93.56%
1,000	83.49%	88.36%	93.78%	92.31%	91.85%
500	80.61%	85.66%	93.07%	90.75%	89.74%
250	77.89%	83.17%	92.75%	87.41%	87.23%
100	74.80%	84.04%	91.41%	82.13%	84.46%
50	72.79%	82.73%	88.05%	86.53%	77.54%
25	72.45%	81.34%	88.15%	78.26%	78.52%
10	73.38%	85.90%	88.20%	75.88%	85.81%

Folosind întregul set de exemple extrase din corpus, rezolvarea ambiguitatii perechii *i* - *î* este aproape 100% corecta. Pentru aceasta diacritica, avem acum o instanta gresita din 300 instante, în timp ce baza de comparatie implica o instanta gresita din fiecare 8 instante, deci o îmbunatatire semnificativa.

Cel mai slab rezultat este obtinut în cazul perechii *a* - *a*. Dupa o analiza a rezultatelor, reiese ca principalul motiv care cauzeaza aceasta precizie scazuta este faptul ca multe substantive în limba româna au forma nearticulata terminata în *a* si forma articulata terminata in *a*. De exemplu, *masa* si *masa* reprezinta forma articulata si respectiv nearticulata a substantivului *masa*. De asemenea, timpuri diferite ale aceluiasi verb se disting numai prin terminatia în *a* sau *a*. Algoritmul de învatare este deci indus în eroare din cauza folosirii acestor litere în contexte identice. O solutie simpla consta în evitarea în procesul de învatare a exemplelor care contin *a* sau *a* la sfârsitul unui cuvânt. Rezultatele obtinute sub aceasta ipoteza simplificatoare sunt raportate in Tabelul 2, in coloana *a*-*a*(2). Dupa cum se arata în tabel, câstigul este de mai mult de 4% in precizie folosind doar aceasta conditie simpla (câstig care se traduce într-o reducere a erorii de 87%).

Am folosit de asemenea si algoritmul de învatare bazat pe arbori de decizie C4.5, cu aceleasi date de antrenament, fara a observa însa nici o îmbunatatire comparativ cu rezultatele raportate în Tabelul 2. Dezavantajul folosirii C4.5 pentru aceasta problema este faptul ca faza de învatare este mult mai lenta decât în cazul folosirii algoritmului TiMBL. Pe de alta parte, C4.5 are capacitatea de a genera reguli expresive. „*Daca* L_1*=e si* L_2*=spatiu atunci s*”(99.5%), „*Daca* L_1*=t si* L_2*=spatiu atunci s*” (98.7%), „*Daca* L_{-4}*=p si* L_{-1}*=v si* L_1*=t si* L_2*=e atunci s*”(95.5%), sunt exemple de astfel de reguli. L_i denota o litera învecinata în pozitia i relativ la litera ambigua. Se observa ca aceste reguli nu tin cont de faptul ca literele folosite în clasificare apartin aceluiasi cuvânt sau nu. Algoritmul de

învatare se bazeaza pur si simplu pe litere, indiferent de cuvântul caruia îi apartin. În consecinta, pseudo-omonimele (cum ar fi *peste* si *peste*), sunt adresate în mod egal de aceasta metoda, pentru ca algoritmul are capacitatea de a se extinde dincolo de cuvinte.

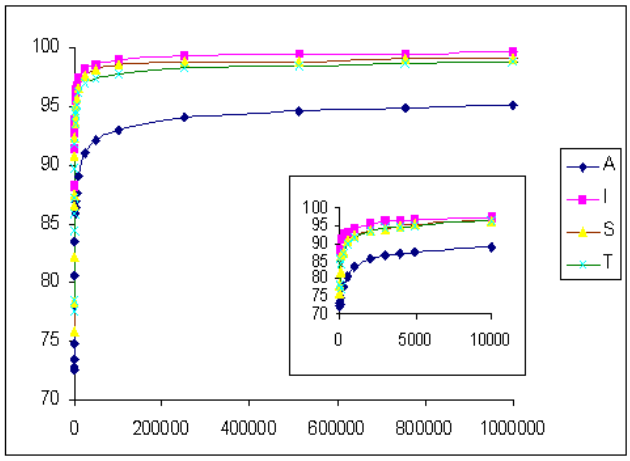


Figura 1. Rate de învatare pentru diacriticele în limba româna. Graficul din mijloc este o reprezentare marita a zonei 0-10.000

3.5. Ferestre de dimensiune diferita

Am efectuat experimente cu ferestre de diverse dimensiuni, pentru a determina dimensiunea contextului care modeleaza cel mai bine problema noastra. Pentru aceasta, am considerat ferestre de dimensiune doi, sase, zece, patrusprezece si 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 corespunzatoare valorii N=5 în tabelul de fata).

Tabel 3

Rezultate comparative obtinute cu ferestre de dimensiuni diferite

Pereche ambiguă	Dimensiune fereastra				
	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 exista suficient context disponibil, o fereastra de dimensiune N=3 poate fi folosita fara a pierde mult din precizie. Însa, dupa cum am specificat si înainte, cea mai ridicata acuratete se obtine pentru o fereastra de zece litere înconjuratoare (N=5).

3.6. Comparatie cu experimente asemanatoare

Rezultatele prezentate în lucrarea de fata se pot compara cu rezultatele raportate de Tufis si Chitu [10], care au folosit tot limba româna în experimentele lor. Tufis si Chitu mentioneaza ca sarcina recuperarii diacriticelor în limba româna este mai dificila decât în alte limbi, deoarece în româna diacriticele sunt mai intens folosite. Dupa cum raporteaza în experimentele lor, numai 60% din cuvintele din limba româna nu au diacritice, comparat cu studii mentionate în [9] care arata ca aproximativ 85% dintre cuvintele limbii franceze se scriu fara accent.

Abordarea prezentata de Tufis si Chitu foloseste dictionare, un analizor morfologic, iar învatarea se face la nivel de cuvinte. Folosind aceste resurse, au obtinut o precizie globala de 97.4%. Nu putem efectua o comparatie directa a rezultatelor noastre, având în vedere ca atât metodele, cât si modul de evaluare, sunt fundamental diferite. Precizia medie de 99% pe care noi o raportam este masurata la nivel de litera, pe când acuratetea raportata in [10] este determinata la nivel de cuvânt[81].

Metodologia noastra depaseste abordarile anterioare, prin faptul ca s-au obtinut precizii si viteze de procesare ridicate fara a folosi nici un fel de resurse aditionale cum ar fi procesoare pentru analiza morfologica sau sintactica sau dictionare. Din aceste motive, algoritmul se poate aplica oricarei limbi, singura cerinta fiind un corpus relativ mic de texte cu diacritice.

4. Alte limbi

Pentru a demonstra generalitatea algoritmului pa care l-am propus, am condus experimente pe texte în alte trei limbi europene care fac uz de diacritice: ceha, poloneza si maghiara. Limbile considerate pentru aceste experimente sunt limbi cu raspândire restrânsa, 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 dupa cum urmeaza: (1) pentru ceha, am folosit arhiva ziarului *Lidovky* si texte literare de *Kafka, Hašek* si *Capek*; (2) pentru maghiara, arhiva furnizata de catre *Digitális Irodalmi Akadémia* si un roman de *Petöfi Sándor*; (3) pentru poloneza, arhiva ziarului *Wiedza i zycie*. Pe lânga acestea, am mai folosit texte aditionale colectate de pe diverse situri, astfel încât sa obtinem un corpus de minim un milion de cuvinte pentru fiecare limba. Asemanator cu procesarea aplicata limbii române, datele au fost convertite în fisiere text, iar literele mari au fost transformate în litere

[81] *Diferența dintre precizia raportată la nivel de literă şi precizia raportată la nivel de cuvânt rezultă practic din diferența de granularitate dintre litere şi cuvinte. Presupunând că un cuvânt conține 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 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.*

mici. În urma acestei etape de pre-procesare, am obtinut un corpus de 1.46 milioane cuvinte pentru ceha, 1.72 milioane cuvinte în maghiara si 2.5 milioane cuvinte în poloneza.

Algoritmii de învatare si atributele folosite în procesul de învatare sunt identice cu cele folosite în experimentele efectuate pe limba româna, raportate în detaliu în sectiunea precedenta. Tabelul 4 prezinta rezultatele obtinute pentru cele trei limbi. Pentru fiecare set de litere ambigue, sunt prezentate în tabel: (1) numarul de exemple obtinute din corpusul limbii respective, (2) baza de comparatie, masurata ca fiind precizia ce se poate obtine daca pentru fiecare set ambiguu se foloseste implicit litera cu frecventa de aparitie cea mai ridicata, si (3) precizia obtinuta prin aplicarea metodei propusa în lucrarea de fata.

Media obtinuta pentru toate patru limbile studiate (cele trei limbi a caror rezultate sunt prezentate in Tabelul 4, si limba româna) este de 98.17%. Precizia medie masurata pe fiecare limba în parte este influentata de marimea setului de date folosit. Textele colectate pentru ceha si maghiara contin aproximativ 1.4-1.7 milioane cuvinte, si prin urmare precizia obtinuta în aceste doua limbi este mai joasa decât pentru poloneza si româna, pentru care am reusit sa colectam un corpus de 2.5-3 milioane cuvinte. Estimam deci posibilitatea cresterii preciziei ca urmare a cresterii dimensiunii corpusului de antrenament.

Tabel 4

Rezultate obtinute în restaurarea diacriticelor în trei limbi europene

Set litere ambigue	Număr exemple	Baza comparatie	Metodă propusă
Cehă			
a á	649,886	75.01%	96.96%
c è	217,570	72.21%	97.08%
d d’	271,070	99.05%	99.86%
e é e	768,051	74.59%	97.02%
i í	504,298	60.43%	96.29%
n ò	439,552	98.97%	99.71%
o ó	566,521	99.08%	99.86%
r ø	319,352	65.55%	97.60%
s š	380,805	84.44%	98.88%
t	387,214	99.05%	99.85%
u ú ù	264,408	80.89%	93.51%
y ý	191,317	65.55%	95.06%
z ž	219,082	66.49%	98.70%
		Medie	97.83%
Maghiară			
a á	1,198,294	73.51%	96.91%
e é	1,306,944	76.34%	96.40%
i í	647,137	89.14%	99.49%
o ó ö õ	678,012	71.15%	96.10%
u ú ü û	207,753	56.00%	97.31%

		Medie	97.04%
Poloneză			
a ¹	1,387,019	88.83%	97.07%
c æ	657,669	91.50%	99.42%
e ê	1,305,584	89.23%	98.47%
l ³	506,041	59.29%	98.80%
n ñ	878,824	96.75%	99.85%
o ó	1,230,389	88.67%	99.87%
s œ	688,677	88.67%	99.83%
z Ÿ ¿	896,909	86.26%	99.73%
		Medie	99.02%

Este interesant de observat ca numarul de diacritice într-o limba nu influenteaza precizia medie obtinuta. Precizia care se obtine în cazul limbii maghiare, care are un total de 5 seturi de litere ambigue, este mai scazuta decât precizia care se obtine pentru limba ceha, care are un numar impresionant de diacritice (treisprezece). Si aceasta cu toate ca datele colectate pentru limba maghiara sunt mai numeroase decât datele colectate pentru limba ceha.

5. Concluzii

Am descris în lucrarea de fata o metoda de restaurare a diacriticelor bazata pe tehnici de învatare la nivelul de litera. Avantajul principal al metodei consta în capacitatea ei de generalizare dincolo de cuvinte. Nu este necesara nici un fel de analiza a textului, si nu se folosesc nici un fel de procesoare de limbaj sau dictionare. Singura cerinta este un corpus relativ mic de texte cu diacritice.

Metoda este folositoare în special pentru limbi pentru care nu sunt disponibile dicţionare electronice de dimensiune adecvate, şi nici procesoare pentru analiză morfologică şi/sau sintactică. Mecanismul de învăţare 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 simplităţii sale, viteza de procesare este foarte mare, de până la 20 pagini de text pe secundă.

Referinte 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-Bcze, 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), vol 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 for machine 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] Tufis, D., Chitu, A. Automatic diacritics insertion in Romanian texts. In Proceedings of the International Conference on Computational 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.

Contributii privind structura statistica de cuvinte în limba româna scrisa*

Adriana VLAD si Adrian MITREA
Universitatea "POLITEHNICA" din Bucuresti
Facultatea de Electronica si Telecomunicatii
B-dul. Iuliu Maniu, 1-3, Bucuresti, România
vadriana@vala.elia.pub.ro

1. Introducere

Aceasta lucrare apartine unui studiu mai larg dedicat de autori descrierii limbii române ca sursa de informatie. Punctul de plecare al acestui studiu a fost presupunerea generala conform careia limba naturala este bine aproximata de un lant Markov ergodic multiplu, cu ordin de multiplicitate mai mare decat 30, [1]. Descrierea acestei surse Markov multiple se realizeaza prin aproximatii succesive. Investigatia noastra statistica pâna în prezent a descris structura de litere, digrame, trigrame, tetragrame, precum si probabilitatile conditionate de o litera precedenta, [2]-[8].

Obiectivul principal ale prezentei lucrari este descrierea sursei de informatie fara memorie având ca simboluri cuvintele limbii române scrise. Aceasta presupune determinarea probabilitatii unui cuvânt (oricare ales), în caz ca aceasta probabilitate exista. Determinarea probabilitatii a însemnat implicit si o verificare a ipotezei de stationaritate a limbii române scrise pe baza structurii de cuvinte; verificarea s-a facut utilizând o procedura similara 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 noastra statistica de a determina probabilitatile cuvintelor a combinat urmatoarele tipuri de inferente statistice: teoria estimarii cu multiple intervale de încredere statistica; test al ipotezei ca probabilitatea apartine unui interval; test de egalitate între doua probabilitati.

Primele doua tipuri de inferente statistice mentionate (teoria estimarii cu multiple intervale de încredere statistica; test al ipotezei ca probabilitatea apartine unui interval) au folosite pentru a decide care este intervalul de încredere statistica "*reprezentativ*" pentru probabilitatea cuvântului investigat în textul natural. Simultan a aparut si o multime "*reprezentativa*" de date *i.i.d.* extrase din textul natural, corespunzatoare cuvântului

* *O parte din acest studiu s-a desfasurat în cadrul unui* ***Grant CNCSIS-MEC*** *(2001-2002) cu tema: "*Descrierea limbii române scrise ca sursa de informatie*"*

investigat (modelul statistic *i.i.d.* presupune ca datele provin din variabile aleatoare independente statistic si identic distribuite).

Ultimele doua tipuri de inferente statistice mentionate (test al ipotezei ca probabilitatea apartine unui interval; test de egalitate între doua probabilitati) ca si intervalele de încredere statistica "*reprezentative*" si multimile de date "*reprezentative*" obtinute în prealabil au fost folosite pentru comparatii matematice între texte naturale. Aceste comparatii matematice (dincolo de valoarea lor ca atare) au avut scopul principal de a vedea daca putem vorbi de un model matematic al sursei de cuvinte pentru limba ca ansamblu, pe domenii ale limbii, pe autori, etc. Comparatiile s-au facut în doua moduri:

- urmarind probabilitatile unui cuvânt (acelasi) în texte naturale diferite;
- urmarind probabilitatile cuvintelor situate pe acelasi rang în texte naturale diferite (se compara probabilitatile asociate unui aceluiasi rang în ierarhiile frecventelor relative).

Rezultatele experimentale au adus probe noi în sprijinul ipotezei de stationaritate a limbii române scrise în cadrul unui aceluiasi domeniu punctând catre unele diferente între domenii diferite.

Investigatia noastra (atât privind intervalele de încredere statistica "*reprezentative*", cât si comparatia matematica dintre texte) a avut în vedere si eroarea statistica de ordinul al doilea. Acest tip de eroare are un rol special în dimensionarea unui nou corpus lingvistic care sa satisfaca acuratetea dorita pentru descrierea modelului matematic (sursa de informatie de cuvinte).

Lucrarea mai contine si un studiu experimental al uneia dintre cele mai cunoscute legi de tipul rang – frecventa, legea lui Zipf. Este analizat si un corolar al acesteia, de interes lingvistic.

Analiza experimentala s-a bazat pe corpusul lingvistic global pe care l-am alcatuit în prealabil pentru studiul structurilor de litere, digrame, trigrame si tetragrame (vezi spre exemplu [6]). Acest corpus este format din 93 de carti în limba româna, scrise cu noua ortografie (introdusa dupa 1993). Cartile reprezinta: literatura scrisa de autori români (11 carti: romane si nuvele), literatura straina tradusa în româna (47 de romane si nuvele), carti stiintifice (drept, medicina, silvicultura, istorie, sociologie, etc.) si altele. Au fost considerate doar cele 31 de litere ale limbii române (A A Â B C D E F G H I Î J K L M N O P Q R S S T T U V W X Y Z) precum si caracterul spatiu (blanc); orice alte simboluri (cifre, elemente de ortografie sau punctuatie) au fost eliminate (suprimate).

Rezultatele experimentale au fost obtinute pe diverse corpusuri organizate pe baza celor 93 de carti:

- Corpusul Mixt Global (**#CMG**) – obtinut prin concatenarea aleatoare a celor 93 de carti; acesta contine un numar de $L_C = 8806433$ cuvinte dintre care $N_C = 202403$ sunt distincte.

- Cele doua jumatati ale Corpusului Mixt Global: prima jumatate (**#1JCMG**) si a doua jumatate (**#2JCMG**); acestea contin un numar de $L_c = 4403217$ cuvinte si respectiv $L_c = 4403216$ dintre care $N_c = 148853$ si respectiv $N_c = 137845$ sunt distincte.
- Corpusul Literar Global (**#CLG**) – obtinut prin concatenarea aleatoare a 58 de carti (romane si nuvele scrise de autori români sau traduse în româna); acesta contine un numar de $L_c = 6255235$ cuvinte dintre care $N_c = 162124$ sunt distincte.
- Cele doua jumatati ale Corpusului Literar Global: prima jumatate (**#1JCLG**) si a doua jumatate (**#2JCLG**); acestea contin un numar de $L_c = 3127618$ cuvinte si respectiv $L_c = 3127617$ dintre care $N_c = 116247$ si respectiv $N_c = 116860$ sunt distincte.
- Corpusul Stiintific Global (**#CSG**) – obtinut prin concatenarea aleatoare a 11 de carti; acesta contine un numar de $L_c = 1049969$ cuvinte dintre care $N_c = 59093$ sunt distincte.

Au fost facute determinari atât pe o singura carte cât si pe grupuri de carti scrise de acelasi autor. Dintre acestea mentionam:

#1. George Calinescu, *Bietul Ioanide,* Editura Minerva, Bucuresti, 1995, ISBN 973-21-0432-5 (*vol 1,* ISBN 973-21-0431-7, pag. 1-214), (*vol. 2,* ISBN 973-21-0433-3, pag. 5-256), (*vol. 3,* ISBN 973-21-0434-1, pag. 5-238).

#2. Radu Anton Roman, *Precum fumul,* Editura Cartea Românească, Bucuresti, 1996, ISBN 973-23-0274-7, pag. 5-283.

#3. Radu Anton Roman, *Zile de pescuit,* Editura Metropol, Bucuresti, 1996, ISBN 973-562-073-1, pag. 11-302.

#4. John le Carré, *Casa Rusia,* Editura Univers, Bucuresti, 1997, ISBN 973-34-0457-8, pag. 9-355.

#5. John le Carré, *Spionul care venea din frig,* Editura Univers, Bucuresti, 1996, ISBN 973-34-0355-5, pag. 9-252, cu ortografie actualizata.

#6. John Le Carré, *Micuta tobosareasa,* Editura Univers, Bucuresti, 1998, ISBN 973-34-0430-6, pag. 7-443, cu ortografie actualizata.

#7. Alexandr Soljenitîn, *Arhipelagul Gulag,* Editura Univers, Bucuresti, (*vol. I,* 1997, ISBN 973-34-0454-3, pag. 7-432), (*vol. II,* 1997, ISBN 973-34-0480-2, pag. 5-474), (*vol. III,* 1998, ISBN 973-34-0497-7, pag. 5-414), cu ortografie actualizata, fara note.

Primul pas în analiza noastra a fost evaluarea frecventelor relative ale cuvintelor din corpusurile mentionate anterior. Tabelul 1 contine primele 55 de cuvinte din ierarhia frecventelor relative din diverse corpusuri.

Un alt rezultat experimental interesant este identificarea unui numar de 162 de cuvinte care se regasesc în toate cele 93 de carti ce alcatuiesc corpusul (fie ca este vorba de literatura, medicina, drept, etc.). Desi sunt doar 162, aceste cuvinte au o pondere importanta în textul global **#CMG** acoperind circa 45% din totalul celor 8806433 cuvinte. Aceste cuvinte comune împreuna cu rangul lor în ierarhie si frecventele lor relative în întreg textul **#CMG** sunt continute în Tabelul 2.

Tabel 1

Ierarhia frecventelor relative în câteva corpusuri
0. Rang; 1. Cuvânt; 2. Frecventa relativa (în %)

	#CMG		#CLG		#1		#4+#5+#6		#6		#CŞG	
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	şi	3,20	şi	3,12	şi	2,65	în	2,55	în	2,58	în	3,47
3	în	2,67	în	2,44	în	2,50	şi	2,39	şi	2,58	şi	3,07
4	să	1,62	să	1,87	cu	1,75	să	1,81	o	1,94	a	2,35
5	a	1,47	la	1,52	o	1,62	o	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	o	1,30	o	1,41	pe	1,39	pe	1,39	pe	1,41	o	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	lui	0,71	al	0,53
20	lui	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	el	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	îi	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	îl	0,26	e	0,30	nici	0,32	fost	0,24
32	nici	0,28	cum	0,28	al	0,26	aŞa	0,30	cum	0,30	după	0,24
33	prin	0,28	sau	0,25	nici	0,25	dacă	0,29	îŞi	0,28	dacă	0,21
34	sa	0,26	fost	0,25	fără	0,25	îl	0,29	îl	0,28	c	0,19
35	sunt	0,25	după	0,24	îi	0,25	charlie	0,29	al	0,28	când	0,19
36	cum	0,25	sa	0,24	avea	0,24	fost	0,28	aŞa	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	joseph	0,24	pot	0,16
40	au	0,23	aşa	0,22	însă	0,22	iar	0,26	sau	0,24	are	0,16
41	îi	0,23	îl	0,22	foarte	0,21	îşi	0,26	ai	0,23	penală	0,16
42	mă	0,21	îşi	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	aşa	0,20	am	0,25	sa	0,22	lui	0,16
45	poate	0,19	sunt	0,20	sunt	0,19	sau	0,24	fără	0,22	acest	0,16
46	aşa	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	îşi	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	gonzalv	0,17	au	0,18	săi	0,19	între	0,13

Tabel 2.

Lista cuvintelor comune în toate cele 93 de carti
1. Cuvânt; 2. Rangul cuvântului în ierarhia frecventelor relative în textul mixt global, #CMG; 3. Frecventa relativa 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	0,05
şi	2	3,20	poate	45	0,19	atunci	94	0,10	una	187	0,05
în	3	2,67	aşa	46	0,18	două	95	0,10	început	188	0,05
să	4	1,62	ar	47	0,18	doar	96	0,10	încât	193	0,05
a	5	1,47	fără	48	0,18	dintre	100	0,10	alte	196	0,04
la	6	1,46	îşi	49	0,17	are	101	0,10	acestea	198	0,04
se	7	1,39	îl	50	0,17	face	102	0,10	facă	199	0,04
cu	8	1,38	le	51	0,17	sub	104	0,09	altă	200	0,04
o	9	1,30	ale	52	0,17	nimic	106	0,09	acelaşi	204	0,04
nu	10	1,28	toate	53	0,17	fel	107	0,09	deşi	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	puțin	109	0,09	printre	220	0,04

că	13	0,97	tot	56	0,16	între	110	0,09	pare	224	0,04
mai	14	0,95	lor	57	0,16	întrun	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	până	59	0,15	i	113	0,08	sus	240	0,04
ce	17	0,66	chiar	60	0,15	săl	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	20	0,49	fie	65	0,14	față	126	0,08	tuturor	264	0,03
dar	21	0,45	ne	66	0,14	unul	127	0,08	măcar	266	0,03
fi	22	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	24	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	nouă	299	0,03
e	26	0,34	numai	72	0,13	ele	138	0,07	acela	302	0,03
el	27	0,34	despre	73	0,12	totul	140	0,07	trebui	307	0,03
al	28	0,33	avea	74	0,12	dată	141	0,07	dintro	330	0,03
când	29	0,33	atât	75	0,12	toți	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	rămâne	371	0,02
prin	33	0,28	unde	80	0,12	orice	155	0,06	alt	373	0,02
sa	34	0,26	întro	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	săi	86	0,11	acesta	168	0,06	alta	404	0,02
dacă	38	0,24	cât	87	0,11	lucru	169	0,06	legătură	415	0,02
după	39	0,24	mare	88	0,11	către	174	0,05	măsură	429	0,02
au	40	0,23	apoi	89	0,11	multe	175	0,05	rândul	601	0,01
îi	41	0,23	ceva	91	0,11	celor	178	0,05			
ea	43	0,21	însă	92	0,10	totuşi	184	0,05			

2. Descrierea structurii statistice de cuvinte. Studiu bazat pe multiple intervale de încredere statistica

Fie un text natural **considerat ca succesiune de cuvinte** pe care îl esantionam cu o perioada suficient de mare astfel încât sa rupem practic dependenta dintre observatiile succesive. Initial în investigatia noastra statistica am considerat aceasta perioada ca fiind de 200 cuvinte. La fiecare moment de esantionare am înregistrat observatia facuta (cuvântul respectiv), conform Fig. 1. Multimea de date obtinute în acest fel contine *N* cuvinte unde $N = L_c / 200$, unde L_c este lungimea textului în cuvinte.

1 201 401

CÂND GAITTANY AMINTI LUI … GAITTANY TACU CACI STIA CA … ÎNSA GAITTANY LIPSIT DE

1.	CÂND	CACI	LIPSIT...
2.	GAITTANY	STIA	DE
...			
200.		TACU	GAITTANY ...

Figura 1. 200 de multimi de date (cuvinte) în model statistic *i.i.d.* obtinute prin esantionare periodica a textului natural

Deplasând originea esantionarii în textul natural apar 200 de astfel de multimi de date experimentale, fiecare în parte de volum N, Fig. 1.

Fiecare multime de N observatii astfel obtinuta satisface modelul statistic *i.i.d.*, model necesar în aplicarea inferentelor statistice utilizate. Independenta este asigurata de marimea perioadei de esantionare; distributia identica este un rezultat al ipotezei de stationaritate a limbii naturale.

Acceptând ipoteza de stationaritate a limbii, toate cele 200 de multimi de date experimentale (compatibile cu modelul *i.i.d.*) extrase din textul natural conform Fig. 1, trebuie sa contina aceeasi informatie despre probabilitatea cuvântului investigat (oricare ar fi acesta).

! Atentie însa, aceste multimi de date nu sunt independente între ele.

Un prim obiectiv al studiului nostru a fost de a vedea daca într-adevar cele 200 multimi de date confirma sau nu aceeasi probabilitate p teoretica (necunoscuta) a cuvântului investigat.

Un raspuns afirmativ ne-ar permite sa obtinem un model matematic pentru sursa de informatie de cuvinte asociata limbii române. Pentru a da un raspuns am extins o procedura statistica pe care am dezvoltat-o în [3]-[8] pentru m-grame. Prin aceasta procedura cele 200 de multimi de date experimentale se compara între ele aplicând repetat un test statistic al ipotezei ca probabilitatea apartine unui interval dat, vezi Anexa 1.

Mentionam ca nu am putut face o comparatie pe baza unui test mai des folosit, anume acela privind egalitatea între doua probabilitati, întrucât multimile de date care se compara nu sunt independente între ele.

Procedura a permis în final determinarea unui interval de încredere statistica optim care a fost denumit în continuare "*reprezentativ*" pentru cuvântul urmarit si textul natural. Simultan a aparut si multimea de date experimentale *i.i.d.* "*reprezentativa*" pentru cuvântul respectiv si textul natural, multime ce va fi folosita în comparatii matematice între texte naturale.

2.1. Intervale de încredere statistica "*reprezentative*" pentru probabilitatile cuvintelor. Metoda de determinare si rezultate experimentale

Scopul acestui subcapitol este de a determina probabilitatea p a unui un cuvânt urmărit.

Fie m_i numarul de aparitii ale cuvântului în multimea i de date experimentale *i.i.d.* de volum N, $i = 1 \div 200$. (Aceste multimi sunt extrase din textul natural conform Fig. 1.)

Aplicând teoria estimarii, fiecare din cele 200 de multimi de date conduce la o estimatie $\hat{p}_i = m_i / N$ a probabilitatii p necunoscute si la un interval de încredere statistica al probabilitatii $I_i = (p_{1,i}; p_{2,i})$, $i = 1 \div 200$. Considerând N suficient de mare astfel încât conditia deMoivre – Laplace sa fie satisfacuta, $Np(1-p) >> 1$, limitele intervalului de încredere statistica (inferioara si superioara) se calculeaza conform relatiei (1), [9], [10]:

$$p_{1,i} \cong \hat{p}_i - z_{\alpha/2}\sqrt{\hat{p}_i(1-\hat{p}_i)/N} \qquad p_{2,i} \cong \hat{p}_i + z_{\alpha/2}\sqrt{\hat{p}_i(1-\hat{p}_i)/N} \qquad (1)$$

unde $z_{\alpha/2}$ este $\alpha/2$ cuantila legii normale de medie 0 si dispersie 1. În determinarile noastre experimentale am lucrat cu un nivel de încredere statistica de 95%; rezulta $z_{\alpha/2} = 1.96$.

Cu alte cuvinte putem spune ca probabilitatea adevarata p se afla în intervalul $I_i = (p_{1,i}; p_{2,i})$, cu o încredere statistica egala cu 0,95.

Într-o prima etapa a analizei noastre, pentru un anumit eveniment urmarit (aparitia unui cuvânt), s-au folosit urmatoarele marimi (a se vedea Fig. 2):

- p^* – frecventa relativa a cuvântului pe întreg textul natural considerat (ceea ce înseamna masurare din date corelate); p^* este raportul între numarul de aparitii ale cuvântului în textul natural si lungimea L_C a textului respectiv (numarul total de cuvinte). Se observa ca p^* este media aritmetica a celor 200 de estimatii. Subliniem ca p^* este o marime importanta pentru orice experimentator.
- $\hat{p}_{\min} = \min_i \hat{p}_i$, $i = 1 \div 200$ – valoarea minima a estimatiilor;
- $\hat{p}_{\max} = \max_i \hat{p}_i$, $i = 1 \div 200$ – valoarea maxima a estimatiilor;
- $\Delta_M = \max_i p_{2,i} - \min_i p_{1,i}$, $i = 1 \div 200$ – reuniunea celor 200 de intervale de încredere statistica;

- $\Delta_M^c = \max_i \hat{p}_i - \min_i \hat{p}_i$, $i = 1 \div 200$ – diferenta maxima între doua estimatii (intervalul de împrastiere al estimatiilor);
- $\delta_M = \max_i \left| \hat{p}_i - p^* \right|$, $i = 1 \div 200$ – diferenta maxima între estimatiile $\hat{p}_i$ si frecventa relativa p^*;
- $\delta_m = \min_i \left| \hat{p}_i - p^* \right|$, $i = 1 \div 200$ – diferenta minima între estimatiile $\hat{p}_i$ si frecventa relativa p^*.

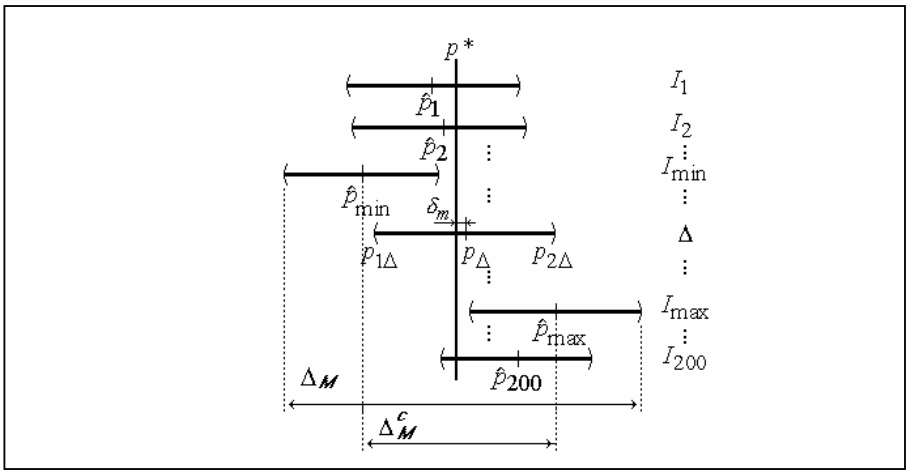


Figura 2. Marimi utilizate în obtinerea intervalului de încredere statistica *"reprezentativ"* pentru probabilitate

Urmatoarele întrebari (probleme) au ghidat analiza noastra teoretica si experimentala:

1. **Cât de largi sunt intervalele Δ_M^c, δ_M si Δ_M?** Intervalele Δ_M^c si δ_M sunt importante în analiza împrastierii estimatiilor în jurul valorii p^*. Intervalul Δ_M ne da o idee despre cel mai mare interval în care s-ar afla *p*, probabilitatea adevarata, banuita ca exista.
2. **Exista valori $\hat{p}_i$ foarte apropiate de p^* si cât de apropiate?**
3. Pentru a raspunde la aceasta întrebare a fost urmarita experimental marimea δ_m. δ_m conduce la estimatia $\hat{p}_i$ care este cea mai apropiata de p^*,

estimatie care va fi în continuare notata cu p_Δ. S-a notat cu Δ intervalul de încredere statistica asociat estimatiei p_Δ conform relatiei (1).

4. **Cât de multe intervale de încredere statistica I_i contin (îmbraca) p^*?** Prin presupunerea de stationaritate, ne asteptam ca un mare numar de intervale de încredere statistica I_i sa se intersecteze, continându-l în acelasi timp pe p^*. Nu ne asteptam la o proportie de 95%, întrucât cele 200 de multimi de date *i.i.d.* nu sunt independente între ele.
5. Putem gasi un interval de încredere statistica pentru probabilitatea adevarata *p*, interval care sa fie acceptat de toate cele 200 de multimi de date experimentale? Se pot gasi mai multe astfel de intervale? Daca ipoteza de stationaritate este adevarata atunci astfel de intervale trebuie sa existe. Daca intervalul Δ (definit mai sus în întrebarea 2) este unul dintre aceste intervale, atunci el va fi preferat de experimentator si va fi considerat ca *"reprezentativ"* pentru probabilitatea cuvântului si textul analizat.

Pentru a întelege metoda dezvoltata de noi care raspunde la aceste întrebari si care conduce la obtinerea intervalului de încredere statistica *"reprezentativ"* exemplificam pentru cuvântul DE în corpusul mixt global **#CMG**.

În Tabelul 3 cele 200 de rânduri corespund celor 200 de multimi de date experimentale *i.i.d.* fiind explicitate atât estimatiile $\hat{p}_i$ cât si intervalele de încredere statistica I_i, $i = 1 \div 200$. Succesiv, fiecare interval I_i a fost considerat ca interval de referinta si am aplicat 199 de teste ale ipotezei ca probabilitatea apartine intervalului mentionat, test descris în Anexa 1. Fiecare dintre cele 199 de teste este aplicat pe o singura multime de date experimentala. (Numarul 199 se explica prin faptul ca nu se testeaza si multimea care a produs intervalul de referinta.)

În primul rând al tabelului 3, intervalul $I_1 = (3{,}89;\ 4{,}27) \times 10^{-2}$ este intervalul de referinta fata de care se aplica testele de apartenenta a probabilitatii. Se testeaza daca probabilitatea cuvântului DE apartine sau nu intervalului I_1 pe baza unei singure multimi de date *i.i.d.*; aceasta înseamna ca verificam succesiv fiecare din restul multimilor de date, anume $i = 2 \div 200$. Acceptarea ipotezei ca probabilitatea cuvântului DE apartine intervalului I_1 este marcata cu "DA" pentru respectiva multime de date (Tabelul 3, rândul 1). În caz contrar, pe pozitia respectiva este completat "NU". Numarul total de multimi de date care trec testele este continut în ultima coloana din dreapta. Aceasta procedura se repeta alegând ca referinta pe rând toate cele 200 de intervale de încredere statistica I_i, $i = 1 \div 200$.

Tabel 3

Determinarea intervalului de încredere statistica *"reprezentativ"* Δ pentru probabilitatea cuvântului DE în corpusul mixt global #CMG. Este îngrosat rândul 3 care corespunde intervalului *"reprezentativ"* Δ

	$\hat{p}_i$	$I_i\ (\times 10^{-2})$	Multime i	Total

i	$\overline{(\times 10^{-2})}$	$p_{1,i}$	$p_{2,i}$	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 obtinut $p^* = 0{,}040986$, iar estimatia cea mai apropiata de p^* a fost $\hat{p}_3 = p_\Delta = 0{,}040992$, $\delta_m / p^* = 0{,}0002$. Pentru estimatia p_Δ se obtine intervalul de încredere statistica 95% $\Delta = (0{,}0391\,;\ 0{,}0429)$. Din rândul 3 se observa ca intervalul Δ trece toate cele 199 de teste ale ipotezei ca probabilitatea cuvântului DE este cuprinsa în interiorul sau. Sunt multe intervale I_i care au compatibilitate cu toate multimile de date *i.i.d.* (în ultima coloana numarul 199 a aparut de 101 ori). Dintre aceste 101 intervale am ales $\Delta = (0{,}0391\,;\ 0{,}0429)$ ca fiind interval de încredere statistica 95% *reprezentativ* pentru probabilitatea cuvântului DE întrucât este usor de determinat de oricare experimentator. Multimea de date *i.i.d.* specificata de indicele $i = 3$ va fi numita multime de date *"reprezentativa"* pentru cuvântul DE în corpusul **#CMG**.

Tabelul 4 contine informatii despre elementele analizei pentru primele zece cuvinte din ierarhia frecventelor relative în corpusul **#CMG**. Exemplificam pentru cuvântul DE care având frecventa relativa $p^* = 4{,}10 \times 10^{-2}$ este pe primul loc în ierarhie. Valoarea p^* este cuprinsa în $N(p^*) = 192$ de intervale de încredere statistica din cele 200 considerate, (coloana 3); reuniunea celor 200 intervale de încredere statistica raportata la p^* este $\Delta_M / p^* = 22{,}24 \times 10^{-2}$, (coloana 4); diferenta maxima între doua estimatii raportata la p^* este $\Delta_M^c \big/ p^* = 13{,}08 \times 10^{-2}$, (coloana 5); diferenta maxima între o estimatie $\hat{p}_i$ si p^* raportata la p^* este $\delta_M / p^* = 7{,}80 \times 10^{-2}$, (coloana 6); diferenta minima între o estimatie $\hat{p}_i$ si p^* raportata la p^* este $\delta_m / p^* = 0.02 \times 10^{-2}$, (coloana 7); largimea intervalului de încredere statistica *"reprezentativ"* Δ raportata la p^* este

$\Delta / p^* = 9{,}23 \times 10^{-2}$, (coloana 8); exista $N(\Delta) = 101$ intervale de încredere statistica la fel de bune ca intervalul Δ *"reprezentativ",* (coloana 9); Aceste $N(\Delta)$ intervale sunt confirmate de toate cele 199 de teste de apartenenta a probabilitatii la interval, prin care s-a facut verificarea stationaritatii.

Numarul relativ mare de intervale de încredere statistica confirmate practic de toate cele 199 de teste de apartenenta a probabilitatii la interval – $N(\Delta)$ din coloana 9 a tabelului 4 – este o sustinere a ideii de stationaritate.

Tabel 4

Rezultate numerice privind marimile din Fig. 2 pentru cele mai frecvente 10 cuvinte în #CMG. Valorile din coloanele 2, 4-8 sunt multiplicate cu 100

Cuvânt	p^*	$N(p^*)$	Δ_M / p^*	Δ_M^c / p^*	δ_M / p^*	δ_m / p^*	Δ / p^*	$N(\Delta)$
1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
DE	4,10	192	22,24	13,08	7,80	0,02	9,23	101
ŞI	3,20	198	25,42	14,91	8,77	0,04	10,45	102
ÎN	2,67	194	24,88	13,43	7,09	0,04	11,43	172
SĂ	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
O	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 si 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 inferentele statistice doar pentru acele cuvinte pentru care am avut suficiente date; anume $N\,p^*(1-p^*) > 20$, unde N este volumul multimii de date *i.i.d.* (forma experimentala pentru conditia DeMoivre - Laplace). Cuvintele au fost sortate în ordine descrescatoare a frecventelor de aparitie p^*. Aceasta sortare a permis organizarea studiului pe clase de frecventa. Am ales ca limite ale claselor urmatoarele valori: 5%, 2%, 1%, 0,5%, 0,2%, 0,1% si 0,05%.

În studiul nostru experimental în aproape toate situatiile (oricare cuvânt urmarit si orice corpus lingvistic investigat) am gasit o estimatie p_Δ practic egala cu p^*. Acest lucru se vede în Tabelul 5, coloana 8 urmarind raportul dintre δ_m si p^*. Pentru toate situatiile analizate am obtinut $\delta_m / p^* \le 2{,}23\%$. Având în vedere ca studiul experimental a condus si la obtinerea de intervale Δ *"reprezentative"* în toate situatiile analizate, rezulta ca aceste intervale de încredere statistica 95% pot fi scrise sub forma:

$$\Delta=(p_{1\Delta}\,;p_{2\Delta})\cong p^{*}\left(1\mp\varepsilon_r\right),\qquad \varepsilon_r\cong 1.96\times\sqrt{(1-p^{*})/(N\,p^{*})}\qquad (2)$$

ε_r este eroarea relativa cu care se determina probabilitatile.

Exemplificam citirea Tabelului 5 pentru corpusul **#CMG** si clasa a doua de frecventa. Exista 8 cuvinte (coloana 3) care au frecventele relative cuprinse între $(0{,}01;\ 0{,}02)$. Aceste 8 cuvinte acopera 11,17% (coloana 4) din totalul aparitiilor de cuvinte din **#CMG,** $L_c=8806433$. Celelalte coloane, 5–9, contin informatii referitoare la marimile din Fig. 2. Astfel coloana 9 contine raportul dintre lungimea intervalului Ä si p^{*} pentru cuvintele existente în clasa respectiva (limita minima si maxima). Acest raport este practic dublul erorii relative, ε_r, în determinarea probabilitatii cuvântului; se observa o precizie relativ buna a determinarilor din aceasta clasa, $\varepsilon_r\le 8.5\times10^{-2}=17\times10^{-2}/2$.

În total în **#CMG** au fost $194=3+8+8+24+59+92$ cuvinte pentru care s-a putut determina intervalul Δ *"reprezentativ"*. Desi cele 194 cuvinte reprezinta o mica pondere din totalul cuvintelor distincte posibile, ele acopera 48,87% din $L_c=8806433$, totalul aparitiilor de cuvinte în corpusul mixt global, **#CMG**.

Tabel 5

Rezultate experimentale organizate pe clase de frecvente relative. Valorile din coloanele 4-9 au fost înmultite cu 100

Clasa de frecvente	Corpus	Nr.	Aco-perire	Δ_M/p^{*}	Δ^{c}_M/p^{*}	δ_M/p^{*}	δ_m/p^{*}	Δ/p^{*}
1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
$2x10^{-2}\le p^{*}<5x10^{-2}$	**#CMG**	3	9,97	22-25	13-15	7-9	0,02-0,04	9-11
	#1JCMG	3	10,04	31-39	18-22	10-13	0,01-0,07	13-16
	#2JCMG	3	9,90	30-41	17-25	9-14	0,01-0,06	13-16
	#CLG	3	9,58	27-33	16-19	8-11	0,03-0,04	11-14
	#1JCLG	3	9,60	39-47	23-27	12-14	0,05-0,10	16-20
	#2JCLG	3	9,55	38-46	22-25	12-13	0,03-0,10	16-20
	#CŞG	4	13,76	52-83	27-52	14-27	0,05-0,37	25-35
$10^{-2}\le p^{*}<2x10^{-2}$	**#CMG**	8	11,17	36-40	21-24	11-14	0,00-0,05	15-17
	#1JCMG	9	12,07	49-59	28-33	14-18	0,00-0,20	21-27
	#2JCMG	8	11,28	46-60	24-37	12-19	0,10-0,16	21-24
	#CLG	9	12,73	39-52	23-30	12-17	0,00-0,09	16-22
	#1JCLG	9	12,69	55-78	30-48	16-25	0,02-0,26	23-30
	#2JCLG	10	13,77	51-75	28-44	14-24	0,00-0,25	23-31
	#CŞG	4	5,36	110-130	67-78	34-48	0,24-0,59	44-50
$5x10^{-3}\le p^{*}<10^{-2}$	**#CMG**	8	6,48	42-61	23-36	12-20	0,00-0,16	19-26
	#1JCMG	8	5,93	60-99	33-62	17-34	0,11-0,29	27-38
	#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 indica si precizia determinarilor (eroarea relativa ε_r) pentru cuvintele analizate. Aceasta precizie este relativ buna pentru determinarile facute pe corpusul mixt global **#CMG** (pentru cuvinte din primele patru clase de frecventa, $\varepsilon_r \le 20{,}5 \times 10^{-2} = 41 \times 10^{-2} / 2$).

Aplicând procedura descrisa în Cap. 2.1 pentru toate corpusurile lingvistice si pentru toate cuvintele care au satisfacut conditia deMoivre - Laplace au rezultat probe în sprijinul ipotezei de stationaritate a limbii române scrise.

Intervalele *"reprezentative"* precum si multimile de date *i.i.d. "reprezentative"* determinate pentru un cuvânt anumit si textul natural considerat au fost în continuare folosite în Cap. 2.2, pentru a analiza daca putem vorbi despre un model matematic al sursei de cuvinte pentru limba ca ansamblu, pentru diverse domenii ale limbii, pentru diversi autori, etc.

Acuratetea în determinarea probabilitatii cuvintelor este data de:

- încrederea statistica (95%);
- erorile relative, ε_r, cu care s-au obtinut intervalele Δ *"reprezentative"* conform Tabelului 5;
- marimea celor doua tipuri de erori statistice care apar în testul de apartenenta a probabilitatii la interval, întrucât acest tip de test .a stat la baza validarii intervalului Δ ca *"reprezentativ"*.

În ceea ce priveste testul de apartenenta a probabilitatii la interval acesta a fost aplicat pentru un prag statistic $\alpha = 0{,}05$. Întrucât testul a fost trecut de fiecare data (fapt pentru care Δ a fost validat ca *"reprezentativ"* este important de a avea un control asupra marimii β, probabilitatea de a accepta date false. Daca am impune valori mici pentru β am avea nevoie de un corpus mai mare. Spre exemplu, conform [6, Tabel 4], daca se doreste $\beta \le 0{,}3$ si $\delta = 0{,}15$ (Anexa 1) pentru a investiga cuvinte din primele patru clase de frecventa am avea nevoie de un corpus de circa 30 de milioane de cuvinte.

2.2 Comparații matematice între diverse texte naturale pe baza structurii de cuvinte

Investigatia noastra (privind stationaritatea) a fost completata cu comparatii matematice privind probabilitatile cuvintelor, pe care le-am organizat dupa urmatoarele criterii:

a) Se verifica daca un acelasi cuvânt are aceeasi probabilitate în cele doua texte naturale care se compara. Aceasta comparatie va fi numita în continuare *comparatie între cuvinte ca atare*.

b) Se verifica daca probabilitatile cuvintelor situate pe un acelasi rang în ierarhia frecventelor relative din cele doua texte sunt egale. Spre exemplu, pe rangul 20 în corpusul literar global, se afla cuvântul DAR, iar în corpusul stiintific cuvântul UN, vezi Tabelul 1. La comparatia între cele doua domenii se va urmari daca probabilitatea celor doua cuvinte (DAR si UN) este aceeasi. În cele ce urmeaza numim acest criteriu *comparatie pe baza rangului*.

Toate comparatiile matematice, atât pe baza criteriului a) cât si pe baza criteriului b) au fost facute folosind urmatoarele teste statistice:

- *T1* – test al ipotezei ca probabilitatea apartine unui interval, (Anexa 1);
- *T2* – test de egalitate între doua probabilitati, (Anexa 2).

Pentru fiecare din cele doua texte naturale care se compara si pentru fiecare cuvânt investigat s-au determinat în prealabil intervalele *"reprezentative"* precum si multimile de date *i.i.d. "reprezentative"*.

Când aplicam testul *T1*, intervalul $(a;b)$ este intervalul *"reprezentativ"* Δ din primul text natural implicat în comparatie, iar multimea $[x_1, x_2, ..., x_N]$ de date experimentale *i.i.d.* este multimea de date *"reprezentativa"* din cel de-al doilea text natural.

Testul a fost aplicat în ambele situatii: corpus1 *versus* corpus2 si corpus2 *versus* corpus1, Tabel 6.

Când aplicam testul *T2* se considera pentru comparatie cele doua multimi de date *i.i.d. "reprezentative"* extrase din cele doua texte naturale pentru cuvintele care se compara.

Toate testele au fost aplicate pentru un prag de semnificatie statistica $\alpha = 0{,}05$. Cu alte cuvinte probabilitatea de a respinge date corecte este mai mica decât 0,05.

Tabel 6

Comparatii între texte naturale pe baza probabilitatii cuvintelor.
Coloanele 4-9 contin numarul de cuvinte rejectate de testele statistice

Texte comparate			*Comparatie între cuvinte ca atare*			*Comparatie pe baza rangului*		
Corpus	Corpus	Nr.	Test *T1*		Test *T2*	Test *T1*		Test *T2*
1	2		1 versus 2	2 versus 1		1 versus 2	2 versus 1	
1	*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. Comparatiile facute în cadrul domeniului literar, când se compara cele doua jumatati de corpus între ele (**#1JCLG** si **#2JCLG**) nu indica diferente între probabilitati indiferent de testul utilizat (*T1* sau *T2*) sau de criteriul utilizat (comparatii pe baza aceluiasi cuvânt sau pe baza aceluiasi rang).

Acelasi rezultat s-a obtinut si când s-au comparat cele doua jumatati ale corpusului mixt global, **#1JCMG** si **#2JCMG**.

Exemplificăm în continuare modul de citire al Tabelului 6.

Primele doua coloane contin corpusurile care se compara între ele.

Coloana 3 indica numarul de cuvinte investigate în comparatii (care au îndeplinit conditia $Np^*(1-p^*) \geq 20$ în ambele texte care se compara).

Rezultatele din coloanele 4, 5 si 6 au fost obtinute aplicând criteriul comparatiilor "cuvintelor ca atare".

Coloanele 4 si 5 arata câte cuvinte nu au trecut testul *T1* de apartenenta a probabilitatii la interval. Coloana 4 se refera la situatia când intervalul fix $(a;b)$ este intervalul Δ *"reprezentativ"* din primul corpus al comparatiei, iar multimea de date *i.i.d.* supusa testului este multimea *i.i.d. "reprezentativa"* din al doilea corpus. Similar, în coloana 5: intervalul fix $(a;b)$ este intervalul Δ *"reprezentativ"* din al doilea corpus al comparatiei, iar multimea de date *i.i.d.* supusa testului este multimea *i.i.d. "reprezentativa"* din primul corpus.

Coloana 6 contine numarul de cuvinte care sunt rejectate de testul *T2* de egalitate între probabilitati.

Coloanele 7, 8 si 9 contin acelasi tip de informatie specificat în coloanele 4, 5 si 6, cu diferenta ca de aceasta data se compara cuvintele care ocupa acelasi rang în loc de cuvintele "ca atare".

Când se compara domenii diferite, spre exemplu literar si stiintific, apar multe diferente marcate de ambele teste *T1* si *T2* si de cele doua criterii de comparatie.

Rezultatele comparatiilor puncteaza unele diferente între domeniile literar si stiintific. Testele nu au indicat diferente când s-au comparat corpusuri organizate dupa aceeasi reguli (jumatatile corpusului mixt global între ele sau jumatatile corpusului literar global între ele); reamintim ca atât corpusul mixt global cât si cel literar global au fost obtinute prin concatenarea aleatoare a cartilor respective.

3. Legea lui Zipf. Studiu experimental

Ierarhiile frecventelor relative ale cuvintelor (prezentate în Cap. 1 si întarite de analiza de stationaritate din Cap. 2) au constituit o baza de plecare pentru studiul nostru experimental asupra legii lui Zipf. În lingvistica legea lui Zipf este una din cele mai cunoscute dependente rang – frecventa. (Aceste dependente rang – frecventa au fost observate de a lungul timpului si în diverse alte domenii: economie, fizica, biologie, demografie, etc. [11], [12].) Obiectivul acestui capitol a fost de a stabili daca si în ce masura (cu ce acuratete) limba româna scrisa satisface legea lui Zipf.

Fie un text (corpus) având o lungime de L_C cuvinte, dintre care N_C sunt distincte. Aceste N_C cuvinte se sortează într-o listă în ordine descrescătoare a numărului de apariții în textul natural. Se notează cu *k* rangul unui cuvânt în listă şi cu $f(k)$ frecvența relativă a acestuia (numărul de apariții raportat la L_C): $f(1) \geq f(2) \geq ... \geq f(N_C)$. (Altfel spus, $f(k)$ este de tipul $p*$ din capitolele precedente). Legea lui Zipf afirmă că produsul dintre rang şi frecvența relativă este constant, [11] – [14].

$$k\, f(k) = A \tag{3}$$

Se observa ca membrul stâng al ecuatiei (3) corespunde realitatii fiind vorba de masuratori efectuate pe texte naturale în timp ce membrul drept corespunde modelului teoretic presupus.

Este stiut din consideratii privind alte limbi naturale ca legea Zipf, apreciata ca foarte simpla si foarte atractiva, functioneaza cu aproximatie pentru o plaja limitata de ranguri, anume nu prea mici si nu prea mari. Astfel un prim pas al studiului nostru teoretic si experimental a fost sa reprezentam grafic dependenta rang – frecventa pe tot corpusul de care am dispus (corpusul mixt global, **#CMG**). Fig. 3 prezinta aceasta dependenta la scara logaritmica ($f(k)$ *versus k*). La o prima vedere am putea spune ca marimea *A* din (3) este aproximativ constanta pentru un interval de ranguri $k \in [k_{\min}; k_{\max}]$ unde $k_{\min} > 50$. Am limitat studiul la acele ranguri pentru care numarul de aparitii ale cuvintelor a fost mai mare decât 50 pentru a beneficia de rezultatele anterioare privind studiul de stationaritate prezentat în Cap. 2. Aceasta face ca rangul $k_{\max}$ sa depinda de corpusul analizat.

Legea lui Zipf este descrisa în numeroase referinte dintre care în limba româna mentionam în special [13] si [14]. Capitolul de fata urmareste determinarea constantei legii atât pe corpusul de ansamblu, **#CMG**, cât si pe diverse texte naturale (grupate dupa autori sau pe subdomenii ale limbii). Se analizeaza si în ce masura comportamentul real se abate de la cel teoretic.

3.1 Elemente teoretice

3.1.1 Determinarea parametrului legii Zipf prin minimizarea erorii p?tratice

Presupunând valabilitatea legii Zipf pentru rangurile $k \in [k_{\min}; k_{\max}]$ ne-am propus sa determinam marimea A din conditia de minimizare pe acest interval a urmatoarei functii (suma patratelor erorilor):

$$g(A) = \sum_{k=k_{\min}}^{k_{\max}} \left[f(k) - \frac{A}{k} \right]^2 \tag{4}$$

Derivând functia $g(A)$ si egalând cu 0 se obtine valoarea marimii A corespunzând minimului:

$$A = \left[\sum_{k=k_{\min}}^{k_{\max}} \frac{f(k)}{k} \right] \Bigg/ \left(\sum_{k=k_{\min}}^{k_{\max}} \frac{1}{k^2} \right) \tag{5}$$

Valorile $k_{\min}$ si $k_{\max}$ sunt la dispozitia experimentatorului. Pentru o evaluare a acuratetei cu care limba naturala verifica legea lui Zipf definim urmatoarele tipuri de erori:

- ε , suma patratelor erorilor pe intervalul $k \in [k_{\min}; k_{\max}]$ si forma ei normata, ε_n :

$$\varepsilon = \sum_{k=k_{\min}}^{k_{\max}} \left[f(k) - \frac{A}{k} \right]^2 \qquad \varepsilon_n = \varepsilon \Bigg/ \left[\sum_{k=k_{\min}}^{k_{\max}} \left(\frac{A}{k} \right)^2 \right] \tag{6}$$

❖ $\varepsilon_{\max}$, eroarea relativă maximă pe intervalul de optimizare $k \in [k_{\min}; k_{\max}]$:

$$\varepsilon_M = \varepsilon_r(k_M) = \max_k \varepsilon_r(k) \qquad \varepsilon_r(k) = \left| f(k) - \frac{A}{k} \right| \Bigg/ \left(\frac{A}{k} \right) \tag{7}$$

3.1.2 Determinarea parametrului legii lui Zipf considerând cazul ideal

Daca acceptam legea lui Zipf ca fiind corecta pe întreg domeniul de ranguri $k \in [1; N_c]$, atunci valoarea constantei A se determina prin rationamentul descris în [13], [14]:

$$A = \frac{1}{c + \ln N_c} \tag{8}$$

unde *c* este constanta lui Euler, egala cu 0,577215 si $N_c > 50$.

Observam ca marimea *A* calculata cu relatia (8) nu depinde decât de numarul N_c de cuvinte distincte din textul analizat. Prin urmare sunt de asteptat unele diferente între evaluarile marimii *A* pe baza datelor experimentale cu relatia (5) si cazul ideal, pur teoretic, relatia (8).

3.1.3 Corolar al legii lui Zipf

Rezultatele experimentale cuprind si verificarea unui corolar al legii Zipf care se refera la determinarea cotei parti, l_s / L_c , pe care o acopera cele mai frecvente *s* cuvinte într-un text de lungime L_c , [13], [14].

$$\frac{l_s}{L_c} = \frac{c + \ln s}{c + \ln N_c} \tag{9}$$

Relatia (9) este valabila pentru un numar de cuvinte $s > 50$.

Observam ca valoarea raportului l_s / L_c nu depinde de marimea *A*. De aceea diferentele existente între diversele moduri de evaluare ale marimii *A* nu vor influenta acest raport. În consecinta ne asteptam la o buna verificare experimentala a acestui corolar.

3.2. Rezultate experimentale si concluzii

Analiza experimentala a legii lui Zipf a început cu corpusul global **#CMG** (vezi Fig. 3) si a continuat pentru comparatie cu o serie de texte naturale incluse în acesta (prezentate în Introducere). Rezultatele experimentale sunt concentrate în Tabelul 7. Pentru fiecare text analizat Tabelul 7 contine în coloanele 2 si 3 numarul total de cuvinte L_c si numarul cuvintelor distincte N_c . În toate textele analizate s-au investigat toate cuvintele cu numar de aparitii mai mare decât 50; acesta determina rangul $k_{\max}$ corespunzator fiecarui text analizat (coloana 4). $k_{\max}$ difera de la text la text; $k_{\min}$ este ales întotdeauna 51. Pentru acest interval de ranguri, $k \in [k_{\min}; k_{\max}]$, s-a determinat cu relatia (5) marimea *A* cuprinsa în coloana 5. Coloanele 6 – 9 contin rezultatele numerice calculate cu relatiile (6) si (7) unde marimea *A* este cea din coloana 5 (determinata din textul natural respectiv). Coloana 9 contine rangul k_M pentru care s-a obtinut eroarea relativa maxima ε_M .

Ne-am pus problema si daca marimea $A = 0{,}0909$ determinata pentru corpusul mixt global **#CMG**, ar putea fi acceptata drept referinta pentru limba româna. De aceea coloanele 10 – 13 contin succesiv marimile din relatiile (6) si (7) unde $A = 0{,}0909$ pentru toate textele naturale analizate. Eroarea relativa maxima ε_M este însotita de rangul corespunzator, k_M .

Tabel 7

Studiu experimental al legii lui Zipf în limba româna scrisa

Text	L_c	N_c	$k_{\max}$	A x10^2	ε 10^6	ε_n x10^2	ε_M x10^2	k_M	ε	ε_n	ε_M	k_M
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

În Fig. 3 sunt prezentate pentru corpusul mixt global doua traiectorii, una experimentala (cu ‘o’) si cea teoretica (cu ‘*’) conform relatiei (3) cu parametrul $A = 0{,}0909$ din coloana 5, Tabelul 7. Se observa o buna concordanta a celor doua curbe pentru $k \in [k_{\min}; k_{\max}]$.

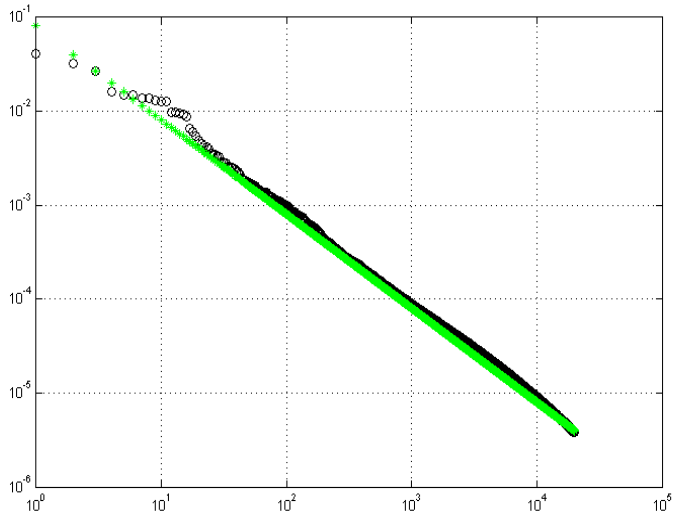


Figura 3. Dependenta rang – frecventa relativa de aparitie a cuvintelor în corpusul mixt global #CMG (scara logaritmica $f(k)$ *versus k*). Curba experimentala marcata cu 'o'; curba teoretica, relatia (3) pentru $A = 0{,}0909$, marcata cu '*'

În cazul ideal, pur teoretic, marimea *A* poate fi determinata cu relatia (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 relatia (8) rezulta $A = 0{,}0795$.

Tabel 8

Valori teoretice, relatia (9), si experimentale ale raportului l_S / L_C în corpusul literar global #CLG

λ	0,1%	0,05%	0,01%
s	104	189	911
l_S / L_C (experimental)	43,69%	49,64%	64,38%
l_S / L_C (teoretic)	41,53%	46,28%	58,78%

Tabelul 8 contine date despre cota parte acoperita de cuvintele pentru care $f(k) \geq \lambda$ unde $\lambda = 0{,}1\%;\ 0{,}05\%; 0{,}01\%$, în textul literar global **#CLG**. S-a folosit relatia (9) unde $N_C = 162124$, iar numarul de cuvinte *s* corespunzator pragului λ este continut în linia 2 a Tabelului. Se observa o concordanta destul de buna între valorile teoretice si cele experimentale.

Nota: Din cele 189 cuvinte din corpusul literar global **#CLG** din care au frecventa relativa mai mare decât 0,05%, doar 156 au îndeplinit conditia deMoivre – Laplace si au fost investigate cu control statistic aparând si în Tabelul 5.

Ca o remarca finala legea lui Zipf poate fi considerata ca valabila si pentru limba româna pentru ranguri nu prea mici si nu prea mari, fapt sustinut de Fig. 3 si datele din Tabelul 7.

4. Concluzii. Perspective

Unul din principalele rezultate obtinute în cadrul acestei lucrari este de a aduce probe noi privind stationaritatea limbii române scrise, de aceasta data pe baza structurii de cuvinte. (Ipoteza de stationaritate este inclusa în presupunerea generala conform careia limbile naturale sunt lanturi Markov multiple ergodice). Analiza stationaritatii s-a facut prin extinderea unei metode dezvoltate de autori pentru studiul structurii statistice de m-grame (litere, digrame, trigrame, tetragrame). În consecinta s-au putut obtine probabilitatile cuvintelor cu intervale de încredere statistica 95% *"reprezentative"*. Aceste intervale pe care le-am numit *"reprezentative"* au avut compatibilitate cu toate multimile de date *i.i.d.*

obtinute prin esantionarea periodica a textului natural. Simultan au rezultat multimile de date *i.i.d. "reprezentative"* pentru cuvântul investigat si textul natural analizat.

O alta contributie consta în procedura de comparatie matematica între texte naturale facilitata de intervalul *"reprezentativ"* pentru probabilitate si de multimile de date *i.i.d. "reprezentative"*. Comparatiile facute între corpusuri organizate în aceeasi maniera (literar *versus* literar sau mixt *versus* mixt) au întarit ideea de stationaritate a limbii si au confirmat modelul matematic prezentat anterior prin intervale de încredere statistica 95% *"reprezentative"* pentru probabilitatile cuvintelor. Au aparut unele diferente între domeniile literar si stiintific.

Rezultatele experimentale dau un plus de semnificatie frecventei relative, p^*, marime de care orice experimentator este interesat. Acest plus de semnificatie este datorat faptului ca în toate situatiile analizate de noi (cuvânt sau text natural) am putut obtine o estimatie a probabilitatii practic egala cu p^*, iar intervalul de încredere statistica asociat acestei estimatii a fost confirmat ca interval *"reprezentativ"* pentru probabilitate.

Lucrarea contine totodata si confirmarea valabilitatii pentru limba româna scrisa a legii lui Zipf (lege de tip rang - frecventa) si a unui corolar al acesteia de interes lingvistic.

Autorii doresc sa multumeasca D-lui dr. ing. Dan TUFIS, membru corespondent al Academiei Române, pentru sprijinul stiintific acordat continuu în studiul limbii române scrise.

Autorii mentioneaza, de asemenea, sugestiile utile primite din partea D-lui Prof. dr. ing. Alexandru Serbanescu de la Academia Tehnica Militara.

Referinte bibliografice

[1] Shannon C. E., "*Prediction and Entropy of Printed English*", Bell Syst. Tech. J., Vol. 30, pp. 50-64, January 1951.

[2]. Adriana Vlad, and A. Mitrea 1997 "Estimating conditional probabilities and digram statistical structure in printed Romanian". In Tufis D, Andersen P. (Eds), *Recent Advances in Romanian Languge Technology*, Bucharest, Ed. Academiei, ISBN 973-27-0626-0, pp. 57-72, http://www.racai.ro/books/awde/ vlad.html.

[3]. Adriana Vlad, A. Mitrea, M. Mitrea, D. Popa, "*Statistical methods for verifying the natural language stationarity based on the first approximation.Case study: Printed Romanian*" în Vol. **VEXTAL'99** (Conferinta Venezia per il trattamento automatico della lingue), Ed. Unipress, ISBN 88-8098-112-9, pp. 127-132, Nov. 22-24, 1999, Venetia; http://byron.cgm.unive.it/events/papers/vlad.pdf.

[4]. Adriana Vlad, A. Mitrea, M. Mitrea, "*Verifying Printed Romanian Language Stationarity Based on the Digram Statistical Structure*", Proceedings of the Romanian Academy, Series A, Vol. I, No. 2/2000, pp. 129-139.

[5] Vlad Adriana, Mitrea A., Mitrea M., *"Two frequency–rank laws for letters in printed Romanian"*, Procesamiento del Languaje Natural, Revista N^{o} 24, Septiembre de 2000, pp. 153-160, ISSN 1135-5948.

[6]. Adriana Vlad, A. Mitrea, M. Mitrea, "*The trigram statistical structure in printed Romanian*", în **ROMJIST** (Romanian Journal of Information Science and Technology), Vol. **4**, No. **3**, 2001, pp. 353-372.

[7]. Adriana Vlad, A. Mitrea, M. Mitrea, "*A Corpus – based Analysis of how Accurately Printed Romanian Obeys Some Universal Laws*", Capitolul 13 în Cartea A Rainbow of Corpora: Corpus Linguistics and the Languages of the World; A. Wilson, P. Rayson, and T. McEnery Editors, Lincom–Europa Publishing House, Munich, 2001, pp. 155-167; lucrarea a fost prezentata la Corpus Lingvistics 2001, Aprilie 2001, Lancaster, Marea Britanie iar rezumatul este cuprins în Proc. of CL2001, pp. 600.

[8]. Adriana Vlad, A. Mitrea si M. Mitrea, "*Contributii privind structura statistica de tetragrame în limba româna scrisa*", in Proc A XXIX-a Sesiune de Comunicari Stiintifice cu Participare Internationala "Tehnologii Moderne în Secolul XXI", Academia Tehnica Militara, nov. 2001, Bucuresti, Sectiunea 9.1, pp. 60-65, ISBN 973-8290-27-9.

[9]. J. Devore, 1987, *Probability and Statistics for Engineering and the Sciences*, 2nd ed., Brooks/Cole Publishing Company, Monterey, Ca.

[10]. Adriana Vlad, B. Badea si M. Mitrea, 1999, *Metode Statistice în Prelucrarea Informatiei. Compendiu si Aplicatii*, Ed. Metropol, Bucuresti, ISBN 973-562-104-5.

[11] Kanter I., Kessler D. A., "*Markov Processes: Linguistics and Zipf's Law",* Physical Review Letters, Vol. 74, No. 22, pp. 4559-4562, May 1995.

[12] Günther R., Levitin L., Schapiro B., Wagner P., *"Zipf's Law and the Effect of Ranking on Probability Distributions"*, Intl. J. of Theoretical Physics, Vol. 35, No. 2, pp. 395-417, 1996.

[13] S. Marcus, Ed.Nicolau, S. Stati, *Introducere în lingvistica matematica,* Ed. Stiintifica, Bucuresti, 1966.

[14] Dinu M., *Personalitatea limbii române,* Ed. Cartea Românească, Bucuresti, 1996.

Anexa 1. Test de apartenenta a probabilitatii la un interval dat – T1

Fie $I=(a;b)$ un interval în care presupunem ca se afla probabilitatea p a unui eveniment urmarit. Dispunem de o multime de date experimentale $[x_1,x_2,...,x_N]$, date care satisfac modelul statistic *i.i.d.*. Ne intereseaza daca datele experimentale $[x_1,x_2,...,x_N]$ confirma ipoteza ca probabilitatea p apartine intervalului $I=(a;b)$, pentru un prag de semnificatie statistica, α, ales.

Procedura de test este urmatoarea:

Se formuleaza cele doua ipoteze statistice, ipoteza nula H_0 si respectiv ipoteza alternativa H_1:

H_0: p apartine intervalului $(a;b)$; $p\in(a;b)$;

H_1: p este în afara intervalului $(a;b)$; $p\notin(a;b)$.

Se alege pragul de semnificatie α (echivalent, nivelul de încredere statistica $1-\alpha$). Se calculeaza estimatia $\hat{p}=m/N$, unde cu m s-a notat numarul de succese ale evenimentului în multimea de date $[x_1,x_2,...,x_N]$. Verificam daca estimatia $\hat{p}$ se afla sau nu în zona de acceptare a datelor. Regiunea de acceptare a datelor este un interval $(c_1;c_2)$ care include $(a;b)$. Intervalul $(c_1;c_2)$ se determina conform relatiei (10), [3]-[8]:

$$\int_{c_1}^{c_2}\frac{1}{\sqrt{2\pi a(1-a)/N}}\exp\left(-\frac{(x-a)^2}{2a(1-a)/N}\right)dx=\int_{c_1}^{c_2}\frac{1}{\sqrt{2\pi b(1-b)/N}}\exp\left(-\frac{(x-b)^2}{2b(1-b)/N}\right)dx=1-\alpha \quad (10)$$

În relatia (10) apar doua functii de densitate de probabilitate corespunzatoare legii normale: de medie a si dispersie $a(1-a)/N$ si respectiv de medie b si dispersie $b(1-b)/N$.

Ipoteza nula H_0 va fi acceptata daca si numai daca estimatia $\hat{p}$ apartine intervalului $(c_1;c_2)$. În caz contrar, $\hat{p}\notin(c_1;c_2)$, datele se resping ca fiind semnificative pentru pragul de semnificatie α ales (se accepta ipoteza H_1).

Ca în orice test statistic, pot sa apara doua tipuri de erori:

Eroarea de tipul (genul) I: Eroarea de a fi respinse date bune, adica sa fie respinsa ipoteza H_0 când ea este corecta. Aceasta situatie apare atunci când estimatia $\hat{p}$ nu satisface testul, adica $\hat{p}\notin(c_1;c_2)$, si totusi probabilitatea adevarata p este în intervalul $(a;b)$. Probabilitatea acestui tip de eroare este mai mica decât α.

Eroarea de tipul (genul) II: Eroarea de a fi acceptate date false, adica sa fie acceptata H_0 când ea este, de fapt, falsa. Aceasta situatie apare atunci când estimatia $\hat{p}$ satisface testul, $\hat{p} \in (c_1; c_2)$, si totusi probabilitatea adevarata p a evenimentului nu apartine intervalului $(a;b)$, $p \notin (a;b)$. Pentru α si N fixate, probabilitatea acestui tip de eroare depinde de valoarea adevarata necunoscuta p, si se calculeaza cu relatia:

$$\beta(p) = \int_{c_1}^{c_2} \frac{1}{\sqrt{2\pi p(1-p)/N}} \exp\left(-\frac{(x-p)^2}{2p(1-p)/N} \right) dx , \qquad p \notin (a;b).$$

$\beta(p)$ este mare atunci când p este la stânga lui a (sau la dreapta lui b), dar foarte aproape de a (respectiv de b). Practic, deranjante sunt situatiile în care $p \le (1-\delta)\cdot a$ sau $p \ge (1+\delta)\cdot b$, si totusi testul este trecut, adica $\hat{p} \in (c_1; c_2)$. Valoarea δ este determinata (prestabilita) de utilizator, în functie de cât de mult deranjeaza aceasta situatie.

În studiul nostru asupra stationaritatii limbii române acest test a fost absolut necesar, vezi Cap. 2. A trebuit sa stabilim daca probabilitatea p a unui anumit cuvânt este aceeasi când dispunem de diverse multimi de date experimentale extrase dintr-un acelasi text natural (unde multimile sunt compatibile cu modelul statistic *i.i.d.*, dar nu sunt independente între ele). Testul a fost folosit si în comparatii între texte naturale.

Anexa 2. Test de egalitate între doua probabilitati – T2

Dispunem de doua multimi de date experimentale în model statistic *i.i.d.*, de volume N_1, respectiv N_2. Notând cu m_1 numarul de succese (aparitii) ale unui eveniment în prima multime de date experimentale, estimatia probabilitatii este $\hat{p}_1 = (m_1 / N_1)$. Similar, pentru a doua multime de date experimentale, estimatia probabilitatii este $\hat{p}_2 = (m_2 / N_2)$. Urmarim sa stabilim daca cele doua estimatii $\hat{p}_1$ si $\hat{p}_2$ provin din aceeasi probabilitate teoretica, respectiv $p_1 = p_2$.

Procedura de test este urmatoarea:

Se formuleaza cele doua ipoteze statistice, ipoteza nula H_0 si respectiv ipoteza alternativa H_1:

H_0: cele doua probabilitati teoretice sunt egale $p_1 = p_2$;

H_1: cele doua probabilitati teoretice sunt diferite $p_1 \neq p_2$.

Se alege pragul de semnificatie statistica α.

Se construieste o valoare de test *z* conform, [9], [10]:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{p_1(1-p_1)}{N_1} + \frac{p_2(1-p_2)}{N_2}}}.$$

Valoarea *z* depinde de datele experimentale prin estimatiile $\hat{p}_1$ si $\hat{p}_2$. În conditiile în care ipoteza H_0 este adevarata *z* provine dintr-o variabila aleatoare a carei lege de repartitie este practic legea normala standard.

Întrucât p_1 si p_2 sunt necunoscute, se considera $p_1 = p_2 \cong \frac{m_1 + m_2}{N_1 + N_2}$.

În aceste conditii valoarea de test *z* devine:

$$z = \sqrt{\frac{N_1 + N_2}{N_1 N_2}} \frac{m_1 N_2 - m_2 N_1}{\sqrt{(m_1 + m_2)(N_1 + N_2 - m_1 - m_2)}}. \qquad (11)$$

Ipoteza nula H_0 va fi acceptata (se va considera ca probabilitatile sunt egale, $p_1 = p_2$) daca si numai daca $|z| \leq z_{\alpha/2}$ ($z_{\alpha/2}$ corespunde pragului de semnificatie statistica α ales; am folosit $z_{\alpha/2}$=1,96). În caz contrar se respinge ipoteza de egalitate a celor doua probabilitati pentru pragul de semnificatie statistica α ales.

Aceasta procedura de test a fost folosita când am comparat între ele diverse texte naturale.

Dezambiguizarea automata a cuvintelor din corpusuri paralele folosind echivalentii de traducere

Dan TUFIS
Institutul de Inteligenta Artificiala
Academia Româna

Rezumat

Corpusurile paralele constituie surse de cunostinte extrem de valoroase, traducerea unui text reprezentând o succesiune de decizii lingvistice pe care traducatorul le ia în vederea asigurarii unei transpuneri cât mai naturale si mai fidele a semnificatiei din textul sursa în textul tradus. Explicitarea si extragerea acestor cunostinte prin metode algoritmice, formalizarea si reutilizarea lor ulterioara constituie provocari ale inteligentei artificiale, subiecte de interes fierbinte în cercetarea actuala. Lucrarea prezinta o serie de contributii în aceasta directie, prezentând mai întâi o metoda originala de identificare a echivalentilor lexicali de traducere a cuvintelor dintr-un corpus paralel (extragând deci un dictionar multilingv) si apoi o metoda extrem de promitatoare pentru identificare automata a diferitelor sensuri ale cuvintelor polisemantice.

Motivatii

Evolutia stiintifica si tehnologica este o sursa permanenta de formare a noi termeni sau a noi sensuri specializate pentru cuvintele existente. În domeniul lexicografiei multilinguale, pastrarea în actualitate a dictionarelor bi- si multilingve fara a apela la tehnologiile informatice, cu precadere cele din sfera ingineriei lingvistice, este aproape imposibila. O serie de studii în domeniul traducerii automate au aratat ca principalele probleme în *acceptabilitatea* traducerilor automate si cu atât mai mult al celor implicând pre- sau post-editare umana, nu sunt legate de probleme de natura sintactica (topica, acorduri, structura frazala) ci ele se regasesc cu precadere în sfera lexicala, mai precis al semanticii lexicale. Evaluarea sistemelor existente de prelucrare a limbajului natural si mai ales a celor de traducere automata (cu variantele ce presupun interventia expertului uman) a condus la identificarea unor puncte sensibile, deficitare (pentru o interesanta trecere în revista a problemelor privind evaluarea sistemelor de prelucrare a limbajului natural si a sistemelor de traducere a se vedea http://www.isi.edu/natural-language/mteval/). De pilda, traducerea gresita a unui cuvânt sau al unei expresii într-o fraza perfecta din punct de vedere sintactic este perceputa de imensa majoritate a consumatorilor de traduceri, cu

precadere de natura stiintifica, ca mult mai grava decât un dezacord gramatical sau vreo alta abatere de la norma gramaticii. S-a invocat pe buna dreptate ca dificultatea majora a prelucrarii automate a limbajului este rezolvarea ambiguitatilor lexicale, a omonimiilor si a polisemiei ce apar în orice text (scris sau vorbit). Spre deosebire de oameni, care de multe ori nici nu constientizeaza aceste fenomene (ele sunt „obturate" fie de contextul textului, fie de cunostintele de „bun simt" ale fiecarei persoane), procesoarele artificiale de limbaj natural încearca rezolvarea ambiguitatilor printr-un proces inteligent de alegere, dintr-un spatiu al tuturor solutiilor posibile în raport cu o modelare a limbajului, a solutiei care respecta cel mai bine restrictiile modelului. Raportarea la modelul limbajului este esentiala întrucât dificultatea procesului de prelucrare este cu atât mai mare cu cât modelul este mai complex: spatiul de cautare a solutiilor poate creste exponential, iar procesul decizional poate deveni nedeterminist sau de complexitate neoperationala.

Rezolvarea algoritmica eficienta a omografiei a cunoscut spectaculoase progrese în ultimii 10-15 ani, dar identificarea automata a sensului pe care îl are un anumit cuvânt polisemantic într-un context dat este înca o problema nerezolvata satisfacator si, prin urmare, un subiect „fierbinte" de cercetare. Problema identificarii sensului cu care este utilizat un cuvânt este vitala în traducerea automata, întrucât se cunoaste faptul ca de foarte multe ori un cuvânt polisemantic dintr-o limba se traduce într-o alta limba prin cuvinte diferite, în functie de sensul considerat. Este interesant de remarcat ca daca un cuvânt polisemantic din limba sursa se traduce printr-un singur cuvânt polisemantic în limba tinta, sau altfel spus toate sensurile cuvântului de tradus se regasesc în cuvântul reprezentând traducerea sa, necesitatea identificarii sensului de utilizare al cuvântului sursa nu mai este obligatorie (cel putin la nivelul fazei de transfer lexical) cu exceptia situatiei în care diferitele sensuri ale cuvântului tinta se realizeaza lingvistic prin structuri de subcategorizare distincte.

În aceasta lucrare vom prezenta în prima parte o metoda de extragere automata a echivalentilor de traducere si vom descrie apoi procedura de discriminare a sensurilor cuvintelor din corpusuri paralele pe baza echivalentilor de traducere.

2. Echivalenti de traducere

1.1. Notiuni preliminare

O pereche de texte în doua limbi diferite L_A si L_B, astfel încât unul reprezinta traducerea celuilalt constituie cea ce se numeste un *bitext*. Un bitext suficient de mare, constituie un corpus paralel. L_A si L_B se numesc echivalenti de traducere. Notiunea de echivalenta de traducere se poate rafina la niveluri subtextuale, de pilda la nivelul paragrafului, al propozitiei sau chiar la nivel lexical, al cuvântului sau al expresiei. În continuare elementul de aliniere lexicala îl vom numi, generic, *atom lexical* sau simplu *atom*. Un bitext în care echivalentii de traducere sunt explicitati se numeste un bitext *aliniat*. Cea mai mica unitate textuala la nivelul careia se realizeaza alinierea defineste *granularitatea*

echivalentilor de traducere. Echivalentii lexicali de traducere (obiectul nostru de interes în aceasta lucrare) depind evident de bitextul din care sunt extrasi iar procesul de extragere a lor devine echivalent cu extragerea unui dictionar bilingv, specific unui anumit domeniu[82]. Extragerea unui dictionar de echivalenti de traducere dintr-un bitext este în fond un proces de explicitare a dictionarului mental folosit de translatorul (sau translatorii) textului original.

Presupozitia fundamentala în încercarea de a alinia corpusurile paralele este ca *aceeasi* semnificatie este exprimata în doua sau mai multe limbi. Definirea identitatii de înteles între doua sau mai multe reprezentari ale (presupus) aceluiasi lucru este o binecunoscuta problema filozofica care ramâne deschisa chiar în domenii mult mai precise decât cel al limbii (de pilda în ingineria software). Prin urmare, notiunea de echivalent de traducere este un concept vag, si pentru operationalizarea sa în domenii ca traducerea automata, terminologie, managementul multilingual al documentelor si altele asemenea avem nevoie de o definitie precisa în termeni direct cuantificabili. Una dintre cele mai larg acceptate definitii a echivalentei de traducere este cea folosita în (Melamed, 2001): „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 abordarilor moderne în extragerea automata a echivalentilor de traducere[83], sprijinite de forta de calcul din ce în ce mai mare a calculatoarelor, utilizeaza metode statistice si pot fi clasificate în doua mari categorii:

- paradigma „*presupune si testeaza*" (Gale, Church, 1991), (Smadja et al., 1996) etc., se bazeaza pe generarea unei multimi de potentiali echivalenti de traducere (spatiul ipotezelor) din care se selecteaza ulterior, pe baza unor teste de independenta statistica, echivalentii de traducere. Selectarea fiecarui echivalent de traducere se face independent de echivalentii extrasi anterior (procesul poate fi considerat ca fiind unul de optimizare locala).
- paradigma „*modelului de limba*" (Brown et all, 1993), (Kupiec, 1993), (Hiemstra, 1997) etc. presupune construirea unui model statistic al bitextului, model ai carui parametri se estimeaza, prin metode de optimizare globala. În aceasta abordare un candidat supus estimarii nu mai este o pereche de atomi lexicali ci o multime de perechi, numita *asignare* (Brown et al., 1993).

Exista sustinatori si critici ai ambelor abordari si o discutie a avantajelor si dezavantajelor lor este prezentata în (Hiemstra, 1997). În esenta, paradigma „*presupune si testeaza*" este mult mai eficienta din punct de vedere computational deoarece presupune investigarea unui spatiu al solutiilor proportional cu N^2, unde N este maximul dintre

---

[82] *Posibilitatea de a genera automat dictionare bilingve în domenii specializate, coroborata cu performantele tot mai bune ale programelor de clasificare automata a textelor, deschide noi perspective traducerii automate si în general prelucrarii multilinguale a textelor.*

[83] *În continuare, daca nu vom specifica altminteri, prin „echivalenti de traducere" vom întelege implicit „echivalenti lexicali de traducere"*

numerele de articole lexicale distincte din cele doua parti ale bitextului, dar echivalentii de traducere cu numar mic de aparitii sunt de obicei pierduti. Paradigma *„modelului de limba”* este extrem de costisitoare din punct de vedere computational întrucât spatiul de cautare al solutiilor este teoretic proportional cu N!, în schimb având potentialitatea identificarii corecte chiar si echivalentilor de traducere cu o singura aparitie în bitext (*hapax-legomena*). În (Brown et al., 1993) sunt prezentati o serie de algoritmi foarte eficienti, bazati pe o serie de ipoteze simplificatoare dar rationale, ce permit o ignorarea unor mari regiuni din spatiul de cautare, regiuni în care este improbabil sa existe solutii acceptabile.

Metoda descrisa aici poate fi încadrata în categoria abordarilor de tip „*presupune si testeaza*”. Algoritmul genereaza mai întâi o lista de candidati si apoi succesiv, alege din aceasta lista perechile cu cele mai mari scoruri de co-ocurenta în regiuni corespondente ale bitextului. Dupa cum se va vedea în continuare, acest algoritm nu are nevoie de un dictionar bilingv initial, dar daca acesta exista, utilizarea sa poate spori substantial viteza si acuratetea prelucrarii.

2.2. Ipoteza corespondentei lexicale 1:1

În general, un cuvânt dintr-un segment ce apare într-o parte a bitextului se traduce în segmentul corespunzator din cea de a doua parte a bitextului tot printr-un singur cuvânt. Daca acest lucru s-ar întâmpla întotdeauna, problema alinierii lexicale a unui bitext ar fi substantial mai simpla decât în realitate. Din pacate ipoteza „*cuvânt la cuvânt”* nu este adevarata în foarte multe cazuri, astfel încât adoptarea ei ca premisa de calcul nu pare foarte promitatoare. Dificultatea poate fi însa ocolita prin considerarea ipotezei conform careia **un articol lexical** dintr-o limba se traduce în cealalta tot printr-**un singur articol lexical**. Asa cum am aratat în sectiunea precedenta, un articol lexical este reprezentat fie de un cuvânt, fie de o secventa de cuvinte (sintagma, compus, expresie). Aceasta formulare, cunoscuta sub numele de „*ipoteza corespondentei lexicale 1:1”*, adoptata ca premisa computationala, simplifica mult problema tinta a alinierii lexicale a unui bitext, dar introduce probleme noi si anume definirea si respectiv recunoasterea automata a articolelor lexicale. Din fericire aceste probleme sunt reductibile la contexte monolingve si au solutii simple si foarte eficiente. Un program capabil sa realizeze recunoasterea automata a articolelor lexicale se numeste *segmentator lexical.* Un segmentator lexical este în general independent de limba, iar functionarea sa este controlata prin resurse specifice (dictionare continând cuvinte, secvente de cuvinte sau expresii regulate definite peste un vocabular limitat). În (Tufis&Barbu, 2001b) este discutata structura resurselor necesare segmentarii lexicale a textelor în limba româna cu ajutorul segmentatorului *MtSeg*, dezvoltat la Universitatea Aix-en-Provence în cadrul proiectului european „Multext”.

Adoptarea „*ipotezei corespondentei lexicale 1:1”* reduce dramatic complexitatea problemei extragerii echivalentilor lexicali, indiferent de paradigma în care este abordata rezolvarea (a se vedea (Tufis&Barbu, 2001b, 2002) pentru detalii). Trebuie mentionat însa ca o segmentare lexicala perfecta (din punctul de vedere al utilitatii ei într-un context multilingv) este practic imposibila din cauza incompletitudinei inerente a oricarui dictionar

frazal. În (Tufis 2001b, Tufis&Barbu2002) se arata cum poate fi surmontata aceasta incompletitudine a resurselor necesare segmentarii lexicale.

2.3. Etape de preprocesare

2.3.1 Alinierea frazal?

Înainte de extragerea propriu-zisa a echivalentilor de traducere, un corpus paralel este supus unor prelucrari preliminare, de aducere a bitextului într-o forma normalizata. Dupa ce fiecare parte a bitextului a fost supusa segmentarii lexicale, urmeaza etapa de aliniere la nivelul propozitiei a corpusului paralel. Pentru acest scop, am utilizat o varianta putin modificata a algoritmului prezentat si documentat (Gale&Church, 1993). În (Tufis&Barbu, 2001b) este descris procesul de aliniere la nivel de fraza si furnizate exemple si statistici pentru diferite perechi de limbi prezente în corpusul paralel multilingv „1984", continând traduceri în sase limbi ale romanului omonim al lui George Orwell. Acolo aratam ca, în marea majoritate a cazurilor, traducerile din limba engleza s-au realizat în celelalte limbi pastrând corespondenta de 1:1 la nivelul frazei[84], cu alte cuvinte, aproape întotdeauna o fraza din textul englezesc a fost tradusa ca o singura fraza în celelalte limbi reprezentate în corpusul paralel. Algoritmul de aliniere la nivelul frazei poate depista si acele cazuri în care traducerea s-a realizat fara pastrarea corespondentei 1:1. Astfel, au fost cazuri în care doua fraze sursa au fost traduse printr-o singura fraza, sau invers, când o fraza din limba engleza a fost tradusa prin 2 sau chiar 3 fraze în celelalte limbi. În cele ce urmeaza, indiferent de tipul de aliniere (1:1, 2:1, 1:2 etc.) vom numi portiunile aliniate la nivelul frazal, *unitati de traducere* (UT).

Ratiunea acestei etape de prelucrare consta în intuitia comuna ca elementele lexicale aflate în relatie de echivalenta de traducere se regasesc în frazele ce se constituie în unitati de traducere. Pe de alta parte, procesul de aliniere la nivelul frazei este mult mai simplu, pentru ca în general indiferent de perechile de limbi considerate intr-un bitext ordinea frazelor dintr-o limba este pastrata în cealalta limba. Aceasta ipoteza, operationalizata de un algoritm de optimizare dinamica de genul celui descris în (Gale&Church, 1993), permite printre altele si identificarea portiunilor netraduse într-una din limbi (alinieri de tipul N:0 sau 0:M).

O alta ipoteza simplificatoare pentru procesul identificarii echivalentilor lexicali de traducere se bazeaza pe observatia ca în marea majoritate a traducerilor, categoriile gramaticale din limba sursa se conserva în limba tinta (Melamed, 2001). Cu alte cuvinte, un verb se traduce de obicei printr-un verb, un substantiv printr-un substantiv s.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 doua limbi au categorii gramaticale diferite. Melamed, distinge si o a treia categorie de perechi de traducere, tipul I, perechile de traducere incomplete, rezultate ca urmare a unei segmentari lexicale partiale si a

[84] *Notiunea de fraza este luata aici în sensul ei larg, al unei propozitii sau fraze (enunt terminat cu un semn de punctuatie din categoria celor finale: punct, punct si virgula, doua puncte, semnul exclamarii, semnul întrebarii, trei puncte).*

utilizarii „ipotezei de aliniere lexicala 1:1”. Consideratiile lui Melamed referitoare la distributia celor trei tipuri de traduceri lexicale sunt foarte bine confirmate de experimentele noastre, în ciuda faptului ca textul nostru este un text literar în timp ce textul sau este un text politic (dezbaterile din Parlamentul Canadian) continând traduceri literale, mult mai putin afectate de personalitatea literara a traducatorului. Ceea ce este demn de remarcat este ca perechile de tip P nu contin categorii gramaticale arbitrare, si ca de la o pereche de limbi la alta, se pot identifica regularitati în alternanta categoriilor gramaticale la traducere (de ex. participiu-adjectiv, gerunziu-substantiv, gerunziu-adjectiv). Astfel de regularitati 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 identificarii rapide si precise a categoriei gramaticale (si eventual al altor trasaturi morfologice sau lexicale) pentru atomii lexicali dintr-un bitext impune o alta prelucrare preliminara, respectiv etichetarea morfo-lexicala, prelucrare pe care o prezentam în sectiunea urmatoare.

2.3.2 Etichetatea morfo-lexical? ?i lematizarea

Etichetarea morfo-lexicala este procesul prin care fiecarui articol lexical dintr-un text arbitrar i se atribuie un cod morfo-lexical unic dintr-o multime specifica articolului lexical respectiv, numita clasa sa de ambiguitate. Codul morfo-lexical reprezinta o reprezentare compacta, si de obicei standardizata, a proprietatilor morfologice si lexicale ce caracterizeaza aparitia unui atom lexical într-un text. Clasa de ambiguitate a unui atom lexical reprezinta multimea tuturor interprtarilor posibile în orice context legal al atomului respectiv. De exemplu cuvântul "*urâti*" are cel putin 8 interpretari posibile putând fi substantiv, adjectiv sau verb. Lema sa poate fi una dintre "urât" (substantiv sau adjectiv), "a urâti" sau "a urî" (verb).

urâți	urâți	Vmnp	(inf.: A *urâți* înseamnă a face să devină urât)
urâți	urâți	Vmis3s	(ind., perf.simplu, sing., pers. 3: El *urâți* totul în viața ei)
urâți	urâți	Vmm-2s	(imp., sing: Prietene, nu *urâți* singurul lucru frumos din viața lui!)
urâți	urî	Vmip2p	(ind., prez., pl., pers. 2: De pomană îi *urâți* pe ei, ceilalți sunt de vină)
urâți	urî	Vmsp2p	(subj., prez., pl., pers. 2: Voi ar trebui să *urâți* tot ce e împotriva vieții)
urâți	urî	Vmm-2p	(imp., sing: Nu-i *urâți* pe apărătorii planetei!)
urâți	urât	Afpmp-n	(adj., masc. pl., neart. : Doi câini *urâți* şi răi păzeau intrarea.)
urâți	urât	Ncmp-n	(subs. com., masc. pl., neart.: Nişte *urâți* m-au băgat în sperieți.)

Asadar, clasa de ambiguitate a cuvântului "*urâti*" este multimea (Vmnp, Vmis3s, Vmm-2s, Vmm-2p, Vmip2p, Vmsp2p, Afpmp-n, Ncmp-n), iar etichetarea morfo-lexicala a acestui cuvânt înseamna a alege, în functie de contextul aparitiei sale, unul si numai unul

dintre cele 8 coduri reprezentând interpetarea contextuala a cuvântului. În cercetarile anterioare am dezvoltat o metoda statistica de etichetare morfo-lexicala (Tufis, 1999), numita etichetarea cu doua niveluri si modele de limba combinate (TT-CLAM: tiered-tagging with combined language models), bazata pe programul TnT al lui Thorsten Brants (Brants, 2000) de prelucrare a modelelor markov cu legaturi ascunse de ordin 2 (3-gram HMM), program ce poate fi descarcat de la adresa www.coli.uni-sb.de/~thorsten/tnt/. Abordarea TT-CLAM a aratat ca texte arbitrare în limba româna pot fi etichetate morfo-lexical în mod corect în peste 98.5% din cazuri si ca atunci când de interes este numai categoria gramaticala, procentul de etichetare corecta depaseste 99.5%. Metoda TT-CLAM s-a dovedit independenta de limba, rezultate mai bune decât în alte abordari fiind raportate în literatura de specialitate pentru limbi foarte diferite de limba româna: limba maghiara (Varadi, 2002, Oravecz *et al.*, 2000, Tufis *et al.*, 2000) limba germana (Hinrics&Truskina, 2002), Slovene (Erjavec, 2002).

Lematizarea este procesul prin care o forma flexionata a unui articol lexical (cuvânt sau expresie) este redusa la forma normala de dictionar. Lematizarea se poate realiza fie printr-un proces de analiza morfologica fie prin cautarea într-o baza de date lexicale, continând cuvinte în forma flexionata însotite de analiza lor morfologica si de forma lema. Lematizarea se realizeaza în acest caz prin identificarea în baza de date a lemei pentru care forma flexionara si analiza morfo-lexicale sunt identice cu cele din textul de lematizat, care desigur a fost în prealabil etichetat. Pentru limba româna, noi am experimentat cu ambele metode si datorita vitezei mult superioare, am optat pentru varianta a doua.

In figura de mai jos este exemplificat rezultatul prelucrarilor preliminare discutate în aceasta sectiune (segmentare lexicala, aliniere frazala, etichetare morfo-lexicala si lematizare) pentru începutul bitextului Englez-Român din corpusul multilingv „1984”. Prima linie arata ca în limba română, fraza cu identificatorul Oro.1.2.2.1, reprezinta traducerea a doua 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 urmatoare, specifica pentru fiecare articol lexical din fiecare limba tipul sau (TOK, LSPLIT, DATE, ABR etc.), forma ocurenta, lema, codul morfo-lexical si categoria gramaticala (ultimele 3 separate prin caracterul „\”).

```
<link targets="Oro.1.2.2.1; Oen.1.1.1.1 Oen.1.1.1.2">
        (<S FROM="Oro.1.2.2.1">                    (<S FROM="Oen.1.1.1.1">
          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         senina   senin\Afpfsrn\A      TOK   bright  bright\Af\A
          …                                         …
          …                                         </S>
          …                                         <S FROM="Oen.1.1.1.2">
          …                                         …
        </S>                                        </S>
```

))

Figura 1: Bitext preprocesat pentru extractia echivalentilor lexicali de traducere

O descriere a principiilor de codificare morfo-lexicala, în conformitate cu recomandarile EAGLES poate fi gasita în Erjavec and Ide (1998). Codificarea specifica pentru limba româna, conforma cu standardul respectiv este pe larg descrisa în (Tufis *et al.*, 1997).

2.4. Un prim algoritm de extragere automata a echivalentilor lexicali de traducere

Exista, asa cum am vazut mai sus, mai multe ipoteze simplificatoare care permit tinerea sub control a complexitatii problemei extragerii automate a echivalentilor de traducere. Nici una dintre aceste ipoteze nu este valabila în cazul general, dar situatiile în care ele nu sunt adevarate sunt suficient de rare astfel încât adoptarea lor nu altereaza valoarea rezultatelor. Trebuie subliniat faptul ca ipotezele simplificatoare folosite de noi, discutate anterior si rezumate în continuare, în general nu afecteaza precizia (corectitudinea) dictionarelor bilingve extrase ci completitudinea lor. Altfel spus, o serie de perechi corecte (echivalenti de traducere reali), desi prezente în bitext, pot sa nu sunt gasite. Precizia si completitudine (în limba engleza acesti termeni sunt *precision* si *recall*) se definesc în mod standard astfel:

PREC=(numar de echivalenti corect extrasi)/(numar total de echivalenti extrasi)

COMP=(numar de echivalenti corect extrasi)/(numar total de echivalenti existenti în bitext)

Mai trebuie precizat si faptul ca ipotezele simplificatoare enumerate mai jos nu împiedica recuperarea ulterioara a echivalentilor negasiti din cauza adoptarii acestor ipoteze de lucru. În (Tufis, 2000) sunt discutate metode de recuperare a unor echivalenti de traducere ce nu respecta ipoteza „echivalentei lexicale 1:1”.

- ipoteza „echivalentei lexicale 1:1”; ea sta la baza majoritatii abordarilor cunoscute: (Kay & Röscheisen, 1993), (Brew & McKelvie, 1996), (Hiemstra, 1997), (Tiedemann, 1998), (Ahrenberg et al., 2000), (Melamed, 2001), etc. Asa cum am aratat mai devreme, un articol lexical identificat corespunzator de un segmentator lexical adecvat diminueaza considerabil efectul contrazicerii acestei ipoteze;
- un articol lexical polisemantic ce apare de mai multe ori în aceeasi unitate de traducere este folosit cu acelasi înteles; aceasta presupozitie este explicit utilizata de (Melamed, 2001) si implicit de toti cercetatorii amintiti mai sus;

- un articol lexical dintr-o parte a unitatii de traducere UT poate fi aliniat unui articol lexical în cealalta parte a UT doar daca cele doua articole au categorii gramaticale compatibile; în majoritatea cazurilor compatibilitatea categoriilor gramaticale se reduce la identitate, dar cum am specificat anterior, este posibil sa se defineasca corespondente compatibile (de pilda, verbele la participiu si gerunziu din limba engleza sunt destul de frecvent traduse în limba româna ca adjective sau substantive, si reciproc).
- Desi ordinea cuvintelor nu este un invariant al traducerii, ea nu este nici arbitrara; când doua sau mai multe perechi de articole lexicale candideaza la statutul de echivalenti de traducere, iar alte criterii de evaluare nu permit departajarea lor, atunci este preferata perechea continând articolele cele mai apropiate în pozitiile lor relative. Aceasta euristica este, de asemenea, folosita de [Ahrenberg et al., 2000].

Pe baza bitextului preprocesat asa cum s-a prezentat în sectiunea precedenta, primul pas al algoritmului este de a delimita spatiul de cautare al solutiilor. Acest lucru se realizeaza prin constructia unei liste a tuturor candidatilor posibili (în conformitate cu ipotezele de lucru amintite mai sus). Aceasta lista, pe care o notam cu TECL (Translation Equivalence Candidates List) contine la rândul ei o multime de sub-liste (câte una pentru fiecare categorie gramaticala luata în considerare). Fiecare sublista contine perechi de forma <token$_S$ token$_T$> unde *token$_S$* si *token$_T$* sunt articole lexicale de categorii gramaticale compatibile si care au aparut în partile corespunzatoare ale aceleiasi unitati de traducere. Fie TU^j cea de a j^a unitate de traducere (translation unit). Prin colectarea tuturor articolelor lexicale apartinând aceleiasi categorii gramaticale POS_k (pastrând ordinea lor relativa si eliminând duplicatele) se construiesc pentru fiecare TU^j multimile ordonate L^{Sj}_{POSk} si L^{Tj}_{POSk}. Pentru fiecare POS_i, fie TU^j_{POSi} produsul cartezian $L^{Sj}_{POSi} \otimes L^{Tj}_{POSi}$. Atunci, definim lista de corespondente în unitatea de traducere TU^j ca fiind CTU^j (correspondences in the j^{th} translation unit):

$$CTU^j = \bigcup_{i=1}^{no.of.pos} TU^j_{POSi}$$

Cu aceste notatii, si presupunând ca bitextul de intrare contine *n* unitati de aliniere, atunci TECL se defineste astfel:

$$TECL = \bigcup_{j=1}^{n} CTU^j$$

TECL contine desigur foarte mult „zgomot" si 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 filtrata pe baza unor functii scor ce supun fiecare TEC la o analiza a ipotezei statistice de independenta a asocierii articolelor lexicale. Pentru a prezenta functiile scor pe care le-am utilizat în experimentele noastre, vom mai defini o serie de notatii:

- TEC = <T_S T_T> $\in$ TECL, un potential echivalent de traducere definit ca perechea formata din articolul lexical sursa T_S si posibila sa traducere T_T în limba tinta;
- n_{11} = numarul de ocurente ale <T_S T_T> din TECL;
- n_{12} = numarul de perechi <T_S $\neg T_T$> din TECL în care T_S a fost asociat cu un articol lexical diferit de T_T;
- n_{21} = numarul de perechi <$\neg T_S$ T_T> din TECL în care T_T a fost asociat cu un articol lexical diferit de T_S;
- n_{22} = numarul de perechi <$\neg T_S$ $\neg T_T$> din TECL ce nu contin nici pe T_S si nici pe T_T;
- n_{1*} = numarul de perechi <T_S *> din TECL în care apare T_S indiferent cu cine este asociat;
- n_{*1} = numarul de perechi <* T_T> din TECL în care apare T_T indiferent cu cine este asociat;
- n_{2*} = numarul de perechi <$\neg T_S$ *> din TECL în care T_S nu apare;
- n_{*2} = numarul de perechi <* $\neg T_T$> din TECL în care T_T nu apare;
- n_{**} = numarul de perechi <* *> din TECL;

Tabela de contingenta din figura de mai jos ilustreaza aceste notatii:

	T_T	$\neg T_T$	
T_S	n_{11}	n_{12}	n_{1*}
$\neg T_S$	n_{21}	n_{22}	n_{2*}
	n_{*1}	n_{*2}	n_{**}

$n_{1*}=n_{11}+n_{12}$, $n_{2*}=n_{21}+n_{22}$

$n_{*1}=n_{11}+n_{21}$, $n_{*2}=n_{12}+n_{22}$

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

Figura 2: Tabela de contingenta pentru un potential echivalent de traducere <T_S T_T>

Pentru ordonarea potentialilor echivalenti de traducere în vederea filtrarii (eliminarea candidatilor cei mai putin plauzibili) am realizat experimente folosind 4 functii de calcul al scorului de echivalenta: MI (*informatia mutuala*), DICE, LL (log likelihood) , and χ^2 (chi-patrat). Folosind notatiile de mai sus, aceste functii-scor se definesc în felul urmator:

(1) $$MI(T_T, T_S) = \log_2 \frac{n_{**} * n_{11}}{n_{1*} * n_{*1}},$$

(2) $$DICE(T_T, T_S) = \frac{2n_{11}}{n_{1*} * n_{*1}},$$

(3) $$LL(T_T, T_S) = 2 * \sum_{j=1}^{2} \sum_{i=1}^{2} n_{ij} * \log \frac{n_{ij} * n_{**}}{n_{i*} * n_{*j}} \text{ si}$$

(4) $$\chi^2(T_T, T_S) = n_{**} \sum_{j=1}^{2} \sum_{i=1}^{2} \frac{\left(n_{ij} - \frac{n_{i*} * n_{*j}}{n_{**}} \right)^2}{n_{i*} * n_{j*}}$$

Figura 3: Functii-scor pentru evaluarea unui potential echivalent de traducere <T_S T_T>

O formula mai simpla de calcul pentru $\chi^2(T_T, T_S)$ este urmatoarea:

(4') $$\chi^2(T_T, T_S) = \frac{n_{**}(n_{11} * n_{22} - n_{12} * n_{21})^2}{(n_{11} + n_{12}) * (n_{11} + n_{21}) * (n_{21} + n_{22}) * (n_{21} + n_{22})}$$

Filtrarea potentialilor echivalenti de traducere se face în raport cu un prag numeric impus scorului calculat cu una dintre functiile de mai sus. Toate perechile ce obtin un scor mai mare decât pragul ales sunt considerate plauzibile si vor fi supuse unor prelucrari suplimentare iar celelalte sunt eliminate. Orice metoda de filtrare statistica va elimina multi echivalenti falsi de traducere, dar pe lânga acestia si un numar de perechi corecte. Alegerea pragului de scor s-a facut având ca obiectiv minimizarea numarului de perechi corecte dar eliminate în mod gresit si a numarului de perechi incorecte acceptate ca urmare a scorului superior pragului de selectie. Dupa mai multe experimente, cele mai bune rezultate s-au obtinut folosind functia de scor LL cu limita pragului de acceptanta egala cu 9.

Într-o prima varianta, algoritmul nostru de extragere a echivalentilor de traducere, având unele asemanari cu algoritmul iterativ prezentat în (Ahrenberg et al. 2000), implementa o strategie de selectie indiferenta la locul si pozitia în corpus a articolelor lexicale aparând în perechea TEC analizata la un anumit moment. O diferenta majora fata de algoritmul descris în (Ahrenberg et al. 2000) este ca în programul nostru calculul diferitelor probabilitati (mai exact al estimatilor de probabilitate) si al scorurilor (testul t) devine nenecesar, conducând la o viteza de prelucrare cu cel putin un ordin de marime mai mare. Pornind de la lista filtrata a potentialilor echivalenti de traducere, algoritmul selecteaza în mod iterativ cei mai plauzibili candidati (vezi mai jos) si apoi îi sterge din lista initiala. Algoritmul se opreste dupa un numar prestabilit de iteratii sau mai devreme în cazul în care lista candidatilor s-a golit sau daca nici un candidat nu mai indeplineste conditia de selectie.

În iteratia *k* a algoritmului se construieste o matrice de contingenta (TBLk) pentru fiecare categorie gramaticala (POS) având dimensiunile S_m* T_n unde S_m si T_n reprezinta numarul de articole lexicale din limba sursa respectiv tinta care mai exista în lista de

candidati la pasul *k* (Figura 4). Liniile si coloanele tabelei sunt indexate cu articolele lexicale (având aceeasi categorie gramaticala) din limba sursa respectiv limba tinta. Fiecare celula (i,j) a matricii reprezinta numarul de ocurente in lista de candidati a perechii $<T_{Si}, T_{Tj}>$.

	T_{T1}	…	T_{Tn}	
T_{S1}	n_{11}	…	n_{1n}	n_{1*}
…	…	…	…	…
T_{Sm}	n_{m1}	…	n_{mn}	n_{m*}
	n_{*1}	…	n_{*n}	n_{**}

$$n_{ij} = occ(T_{Si}, T_{Tj}); \; n_{i*} = \sum_{j=1}^{n} n_{ij} \; ; \; n_{*j} = \sum_{i=1}^{m} n_{ij} \; ; \; n_{**} = \sum_{j=1}^{n} \left(\sum_{i=1}^{m} n_{ij} \right).$$

Figura 4: Matricea de contingenta la pasul *k*

Conditia de selectie la pasul k a multimii de echivalenti de traducere este exprimata de relatia (5):

$$TP^k = \left\{ < T_{Si}\, T_{Tj} > \mid \forall p, q \, (n_{ij} \geq n_{iq}) \wedge (n_{ij} \geq n_{pj}) \right\} \tag{5}$$

Conditia de mai sus constituie esenta algoritmului iterativ (numit în (Tufis&Barbu, 2002) algoritmul BASE) si ea spune ca pentru a selecta perechea $<T_{Si}, T_{Tj}>$ drept echivalent de traducere, numarul de asocieri ale lui T_{Si} cu T_{Tj} trebuie sa fie mai mare sau cel putin egal decât numarul de asocieri ale lui T_{Si} cu orice alt T_{Tp} ($p \neq j$) si simultan numarul de asocieri ale lui T_{Tj} cu T_{Si} trebuie sa fie mai mare sau cel putin egal decât numarul de asocieri ale lui T_{Tj} cu orice alt T_{Sq} ($q \neq i$). Toate perechile selectate în TP^k sunt eliminate din lista de candidati (ceea ce în matricea de contingenta pentru pasul k+1 implica punerea pe 0 a contoarelor de ocurenta pentru perechile selectate anterior). Daca T_{Si} este tradus în mai multe moduri (fie pentru ca are sensuri ce se lexicalizeaza diferit în limba tinta, fie pentru ca în limba tinta se folosesc diferiti sinonimi pentru T_{Tj}) restul traducerilor sale va fi extras în iteratiile urmatoare. Algoritmul discutat este schitat 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));  ## relatia (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 echivalentilor de traducere

2.5. Un algoritm îmbunatatit de extragere automata a echivalentilor lexicali de traducere

Una dintre principalele deficiente ale algoritmului BASE este vulnerabilitatea la ceea ce (Melamed, 2001) numeste *asociatii indirecte*. Daca $<T_{Si}, T_{Tj}>$ are un scor de coocurenta ridicat iar T_{Tj} apare (dintr-un motiv sau altul) de mai multe ori împreuna cu T_{Tk}, s-ar putea ca si perechea $<T_{Si}, T_{Tk}>$ sa primeasca un scor ridicat. Desi, asa cum observa si Melamed, in general, asociatiile indirecte au un scor mai mic decât cele directe (corecte), ele pot obtine totusi scoruri mai mari decât multe alte perechi corecte ce n-au legatura cu T_{Si} iar acest lucru nu numai ca genereaza echivalenti de traducere gresiti, dar va elimina din competitie si echivalenti corecti. Prin urmare asociatiile indirecte afecteaza atât precizia cât si completitudinea procesului. Pentru a slabi aceasta sensibilitate în implementarea algoritmului BASE a fost nevoie de stabilirea unei limite inferioare de ocurenta pentru fiecare articol lexical luat în consideratie. Aceasta limita, conduce inevitabil la eliminarea din spatiul de cautare a solutiilor a mai mult de 50% dintre echivalentii de traducere[85]. Deficienta algoritmului BASE se explica prin faptul ca scorurile de coocurenta sunt calculate în mod global fara a verifica daca atomii lexicali ai unei perechi evaluate sunt sau nu prezenti în unitatile de traducere prelucrate.

Pentru diminuarea influentei asociatiilor indirecte fara a mai impune un prag de ocurenta, algoritmul BASE a fost modificat astfel încât ierarhizarea si alegerea celor mai probabili echivalenti de traducere se realizeaza în contextul local al fiecarei unitati de traducere (desi scorurile lor se calculeaza tot la nivelul întregului bitext). Cu aceasta modificare, noul algoritm (BETA) se apropie de algoritmul „competitive linking" al lui Melamed (Melamed, 2001). Candidatii proveniti din unitatea de traducere curenta sunt analizati prin prisma scorului lor de coocurenta si cel cu scorul cel mai mare este selectat. În baza ipotezei corespondentei lexicale 1:1, dintre candidatii ramasi sunt eliminati toti aceia care contin unul din articolele lexicale ale perechii câstigatoare. Dintre candidatii care ramân dupa aceasta filtrare, se alege din nou cel cu scorul cel mai bun si iar se elimina candidatii continând unul dintre articolele lexicale din perechea selectata. Procesul se

[85] *Pierderea unui asa mare numar de echivalenti de traducere desigur nu surprinde, întrucât una din legile distributionale ale lui Zipf (celebra lege „rang-frecventa") prognosticheaza acest lucru.*

repeta pâna când nici un echivalent de traducere nu mai poate fi extras din unitatea de traducere curenta, caz în care algoritmul trece la prelucrarea urmatoarei unitati de traducere.

Eliminarea pragului de ocurenta a îmbunatatit substantial completitudinea si calitatea dictionarelor de traducere (o detaliata comparatie a performantelor si o analiza cantitativa si calitativa a dictionarelor extrase automat este furnizata în (Tufis&Barbu, 2002)) dar a ridicat problema decelarii între candidatii cu una sau doua aparitii, pentru care scorul de coocurenta este statistic nesemnificativ. În acest caz, criteriul frecventei a fost înlocuit cu o combinatie între un scor de similaritate ortografica si un scor de proximitate relativa. Functia de similaritate ortografica folosita de noi, COGN(T_S, T_T), este o varianta a functiei **XXDICE** descrisa în (Brew&McKelvie, 1996). Astfel, daca T_S este un sir de *m* caractere $\alpha_1\alpha_2 \ldots \alpha_m$ and T_T un sir de *n* caractere $\beta_1\beta_2 \ldots \beta_n$ se construiesc doua noi siruri T'_S si T'_T prin inserarea în T_S si T_T a unui numar minim de caractere speciale astfel încât în final sirurile T'_S si T'_T au aceeasi lungime p (max (m n)≤p<m+n) si un numar maxim de caractere pozitional identice. Fie α_i un caracter din T'_S si β_i un caracter din T'_T care se potrivesc si sunt puse în corespondenta. Fie $\delta(\alpha_i)$ numarul de caractere speciale consecutive ce preced imediat caracterul α_i si $\delta(\beta_i)$ numarul de caractere speciale ce preced imediat caracterul β_i . Fie *q* numarul de caractere care se potrivesc în cele doua siruri. Cu aceste notatii, masura de similaritate COGN(T_S, T_T) se defineste astfel:

$$COGN(T_S, T_T) = \begin{cases} \dfrac{\sum_{i=1}^{q} \dfrac{2}{1+|\delta(\alpha_i)-\delta(\beta_i)|}}{m+n} & \text{if } q > 2 \\ 0 & \text{if } q \le 2 \end{cases} \qquad (6)$$

Limita de relevanta a scorului de similaritate a fost empiric gasita a fi 0.42. Aceasta valoare este dependenta într-o oarecare masura de pereche de limbi considerata în procesul de extragere a echivalentilor de traducere. Implementarea efectiva a testului de similaritate incluse si o serie de normalizari ale sirurilor testate (eliminarea unor afixe, reducerea consoanelor duble, ignorarea distinctiei create de diacritice etc.) normalizari care depind de morfologia fiecarei limbi în parte.

Cel de al doilea criteriu de evaluare a plauzabilitatii unui candidat este scorul de proximitate, DIST(T_S, T_T) definit dupa cum urmeaza:

Daca((<T_S, T_T>∈ $L^{Sj}_{posk} \otimes L^{Tj}_{posk}$)&($T_S$ este al *n*-lea element în L^{Sj}_{posk})&(T_T 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[86] decât DIST(T_S, T_T), astfel încât scorul de similaritate are precedenta asupra celui de proximitate. Functia DIST(T_S, T_T) este invocata doar atunci când COGN(T_S, T_T)=0 (deci

[86] *Motivatia se bazeaza pe intuitia conform careia daca în doua fraze (în limbi diferite) ce reprezinta una traducerea celeilalte, apar cuvintele asemanatoare din punct de vedere ortografic, atunci este foarte rezonabil a presupune ca ele au si aceeasi semnificatie, adica sunt* cogneti.

când atomii lexicali nu prezinta similaritete ortografica) si perechea <T_S, T_T> nu reprezinta o pereche singulara în corpus (hapax-legomena), sau când mai multe perechi candidate au obtinut acelasi scor de similaritate.

Algoritmul BETA este schitat mai jos:

```
procedure BETA(bitext;dictionary) is:
   dictionary={};
   TECL(k)=build-cand(bitext);
   for each POS in TECL do
      for each TU^i_POS in TECL do
         finish=false;
         loop
            best_cand = get_the_highest_scored_pairs(TU^i_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^i_POS =
            remove_pairs_containing_tokens_in_best_pairs(TU^i_POS);
            endif;
         until {(TU^i_POS={})or(finish=true)}
      endfor
   endfor
   return dictionary
end

procedure filtered(best_cand) is:
   result = get_best_COGN_score(best_cand);
   if (cardinal(result)=0)&(non-hapax(best_cand))then
      result = get_best_DIST_score(best_cand);
   else if cardinal(result)>1
          result = get_best_DIST_score(best_cand);
        endif
   endif
   return result;
end
```

Din corpusul paralel multilingv „1984” am extras 6 bitexte continând textul în limba engleza si traducerea în una din cele 6 limbi amintite. Fiecare bitext a fost prelucrat conform celor prezentate în acest capitol si au fost extrase 6 dictionare bilingve, din care s-a obtinut si un dictionar multilingv în 7 limbi (cele 6 plus engleza). În (Tufis&Barbu, 2002) este furnizata o analiza contrastiva cu alte sisteme de acest tip a vitezei de prelucrare. Timpul mediu de extragere a unui dictionar bilingv din corpusul paralel multilingv „1984”

(circa 110.000 de cuvinte în fiecare limba) este 3 minute. Esantioane ale acestor dictionare pot fi consultate la adresa: http://www.racai.ro/~tufis/BilingualLexicons/AutomaticallyExtractedBilingual Lexicons.html.

3. Dezambiguizarea sensurilor lexicale folosind echivalentele de traducere

3.1. Ambiguitatea limbajului natural

Este binecunoscut faptul ca una dintre cele mai dificile probleme în prelucrarea automata a limbajului natural este ambiguitatea sa inerenta. Ambiguitatea se manifesta la toate nivelurile traditionale ale analizei de limbaj: nivelul fonetic si/sau lexical, sintactic, semantic sau discursiv. Ambiguitatea prezenta pe fiecare nivel genereaza exploziv ambiguitati pe nivelurile urmatoare. De pilda, omofonia sau omografia prezenta pe primul nivel la nivelul unuia sau al mai multor cuvinte va produce secvente lexicale diferite (combinatia tuturor interpretarilor posibile la acest nivel) pentru intrarea fazei de analiza sintactica. Fiecare secventa poate conduce, din pricina unor ambiguitati de natura structurala, la interpretari sintactice multiple, dupa cum o serie de secvente lexicale vor putea fi abandonate pe motivul contrazicerii unor restrictii postulate de modelul sintactic al limbii prelucrate. Fiecare dintre interpretarilor sintactice posibile, poate la rândul ei sa conduca la multiple interpretari semantice, în virtutea multiplelor sensuri pe care le poate avea fiecare element frazal al unei analize sintactice. Desigur, interpretarea semantica poate elimina unele structuri sintactice generate în faza anterioara pe baza încalcarii unor restrictii semantice (valabile în orice univers de discurs sau specifice unor domenii discursive de interes). În sfârsit, analiza de discurs, în care contextul interpretativ transcende limita propozitiei, ambiguitatile ramase se presupun a putea fi rezolvate prin utilizarea restrictiilor pragmatice motivate fie de principii generale ale dialogului (coeziune, coerenta), fie de natura bine precizata a unui univers de discurs (modelata prin cunostinte extra-lingvistice despre entitatile universului de discurs). De pilda, în (Cristea&Dima, 2001) rezolvarea anaforelor, proces tipic analizei de discurs, este modelata în termenii identificarii cailor de accesibilitate a entitatilor mentionate în discurs („vene ale discursului"), care la rândul lor sunt formal definite pe baza principiilor generale al coeziunii si coerentei unui text.

Rezulta din cele spuse pâna aici ca identificarea si rezolvarea timpurie, la fiecare nivel de prelucrare, a ambiguitatilor este un imperativ al oricarui demers computational privind prelucrarea limbajului natural. Si cum cuvântul (sau mai exact spus, atomul lexical) este elementul primar în prelucrarea limbajului o mare parte a eforturilor de cercetare este îndreptata spre nivelul lexical al prelucrarilor. Metodele de etichetare morfo-lexicala (tagging), printre care etichetarea cu doua niveluri si modele de limba combinate - amintita în capitolul 2, permit rezolvarea cu mare acuratete a ambiguitatilor categoriale si intracategoriale. De pilda cuvântul *vin* poate fi atât substantiv cât si verb (ambiguitate

categoriala), iar ca verb, el contine ambiguitatea intracategoriala de persoana, numar si mod („indicativ + persoana I + numar singular", „conjunctiv + persoana I + numar singular" sau „indicativ + persoana III + numar plural"). Un program de etichetare morfo-lexicala „instruit" corect pentru limba româna este capabil sa rezolve, în contextul aparitiei sale, astfel de ambiguitati morfo-lexicale.

Curentul lexicalist predominant în modelarea sintactica a limbajului natural presupune precizarea în descrierea de dictionar a fiecarui cuvânt a proprietatilor si restrictiilor sale distributionale sau colocationale relevante pentru analiza sintactica. Pe baza acestor descrieri lexicalizate si a contextului local, multe din potentialele ambiguitati structurale pot fi eliminate, înaintea unei costisitoare analize sintactice, prin tehnici cunoscute sub numele de analiza sintactica partiala (*partial parsing* sau *shallow parsing*).

Un cuvânt omograf, chiar dupa ce a fost corect clasificat din punctul de vedere al categoriei sale gramaticale si al proprietatilor sale distributionale sau colocationale, poate ramâne ambiguu din punct de vedere semantic. Identificarea sensului cu care este utilizat cuvântul polisemantic într-un context dat este desigur de mare interes. Exista însa diferite grade de rafinare a notiunii de sens, iar natura aplicatiei pentru care identificarea sensului este necesara poate impune o acceptie a notiunii de sens diferita de cea utilizata într-un dictionar explicativ. Sa luam, de pilda, problema traducerii automate. Întrucât în imensa majoritate a cazurilor rezultatul traducerii este destinat uzului uman, ceea ce este important este ca în textul tradus sa nu apara ambiguitati suplimentare fata de cele din textul sursa. Cu alte cuvinte, daca o analiza algoritmica evidentiaza în limba sursa o serie de ambiguitati si pornind de la premiza ca textul este admisibil pentru vorbitorii nativi ai limbii textului sursa, de cele mai multe ori este nenaturala o traducere ce încearca sa evite total ambiguitatea identificata. La nivel lexical, aceasta revine la a spune ca daca diferitele sensuri ale unui cuvânt din limba sursa nu se lexicalizeaza prin cuvinte diferite în limba tinta, este neproductiva o încercare a diferentierii sensului contextual, atâta timp cât indiferent care ar fi el, traducerea cuvântului respectiv în limba tinta este aceeasi. Sau cu alte cuvinte, sensurile unui cuvânt din limba sursa ce se regasesc împreuna într-un cuvânt al limbii tinta nu necesita obligatoriu diferentierea pentru traducere. De exemplu, cuvântul englezesc „bottle" are în Wordnet1.5 (Fellbaum, 1998) doua sensuri (ca substantiv) anume de vas de sticla sau plastic cilindric cu un gât îngust si fara mâner, respectiv cantitatea de substanta continuta într-un astfel de vas. Ambele sensuri se regasesc în cuvântul românesc „sticla" (care însa include si alte sensuri lexicalizate în engleza prin cuvîntul „glass"). În acest caz, a încerca eliminarea ambiguitatii la traducerea textului „He drank only a bottle of beer" în limba româna, de pilda prin utilizarea unei parafraze de genul „El bause doar continutul unei sticle de bere", este nenecesara. Orice vorbitor al limbii române va gasi traducerea „El bause doar o sticla de bere" mult mai naturala si desigur nu va avea dificultati a în întelege despre ce este vorba.

Acelasi gen de consideratii se pot face si în raport cu ambiguitatile sintactice pure. Celebrul exemplu „ I saw the Statue of Libery flying over New York" contine cel putin 4 ambiguitati, dar daca de pilda rezolvarea omografului *saw* (am vazut / tai cu fierastraul)

este esentiala în traducere, rezolvarea ambiguitatii structurale poate fi lasata în sarcina mintii celui ce citeste textul: „Am vazut Statuia Libertatii zburând deasupra New York-ului ", caci daca cititorul englez nu are dificultati în a întelege cine si cum zbura, e plauzibil ca nici cititorul român (de exemplu) ne le va avea. Aceasta nu înseamna ca nu exista ambiguitati structurale a caror nerezolvare prealabila sa nu conduca la traduceri hazlii sau chiar incomprehensibile. Ideea este ca metodele formale de analiza a limbajului, modelabile algoritmic, expliciteaza de multe ori ambiguitati greu de constientizat de omul obisnuit, iar luarea în considerare a factorului uman poate simplifica mult prelucrarile automate. Reconsiderarea conceptului de traducere automata în acceptiunea clasica (MT) în favoarea unor concepte mai realiste de tipul HAMT (human assisted machine translation) sau MAHT (machine assisted human translation) a relevat faptul ca, în numeroase ocazii, posteditarea umana a unui text tradus automat introduce ambiguitati care, desi nu sunt sezizabile usor la lectura, pot fi totusi puse în evidenta de algoritmii de analiza.

Cercetarile moderne în domeniul dezambiguizarii automate, în context, a sensurilor cuvintelor sunt motivate si de alte aplicatii informatice, cum ar fi clasificarea dupa continut a volumelor mari de texte, regasirea mai precisa a documentelor electronice, rezumarea automata a textelor, extragerea de cunostinte din texte, crearea de ontologii. Aceasta directie de cercetare, identificata în literatura engleza prin acronimul WSD (Word Sense Disambiguation) constituie de câtiva ani obiectul unor conferinte specializate si chiar a unei competitii de evaluare (SENSEVAL, ajunsa la a treia editie) a solutiilor propuse de specialisti din întreaga lume.

Primii care au sugerat ideea ca, pentru obiectivele WSD, sensurile ce trebuie diferentiate sunt cele care se lexicalizeaza într-o alta limba prin cuvinte diferite au fost Resnik and Yarowsky (1997). Intuitiv, se poate presupune ca, daca un cuvânt din limba sursa se traduce în limba tinta în mai multe feluri si aceste traduceri nu sunt sinonimice, atunci trebuie sa existe o motivatie conceptuala. Analizând un numar suficient de mare de limbi si de texte, e plauzibil, afirmau cei doi specialisti, sa identificam diferentierile lexicale semnificative care delimiteaza sensurile unui cuvânt. Aceste sensuri sunt numite de cei doi „*sensuri tari*". Inabilitatea de a identifica corect sensurile tari este principala sursa a erorilor inacceptabile în orice aplicatie multilinguala. Utilizarea textelor paralele pentru WSD (Gale *et al.,* 1993), (Dagan *et al.,* 1991), (Dagan and Itai, 1994), în scopul identificarii proprietatilor semantice a lexemelor si a relatiilor dintre ele (Dyvik, 1998) a folosit implicit sau explicit notiunea de „sens tare". Mai recent, pe baza echivalentilor de traducere extrasi din corpusul „1984" prin procedura noastra, descrisa în capitolul precedent, Ide (1999) a aratat ca diferentele 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 (2000) au folosit în schimb traducerile unor propozitii izolate în limba engleza efectuate de vorbitori nativi ai limbilor tinta, dar în mare concluziile studiului lor au fost aceleasi cu ale lui Ide. În ambele studii amintite referinta pentru limba engleza a fost WordNet (Miller *et al*., 1990) si desi rezultatele lor sunt promitatoare, mai ales pentru sensurile tari, ele se bazeaza pe o multime prestabilita de sensuri. Date fiind divergentele semnificative între distinctiile de sens realizate în dictionarele (monolingve)

existente, precum si inexistenta unui acord general asupra gradului de rafinare a descrierilor de sens în practica lexicografica internationala, raportarea la un inventar prestabilit de sensuri, cel putin din perspectiva prelucrarii automate a limbajului, nu pare a fi o solutie optima. În continuare, vom prezenta o abordare alternativa, detaliata în (Ide *et al.* 2001, Ide *et al.* 2002).

3.2. Discriminarea automata a sensurilor lexicale: metodologia

Metoda pe care o vom descrie este menita 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 pilda un roman), un cuvânt sa fie folosit în toate sensurile sale, metoda desigur va identifica, prin analiza textuala descrisa în continuare, doar acel sens sau acele sensuri cu care este folosit cuvântul respectiv în textul prelucrat. La limita prin prelucrarea unor texte foarte diferite este posibil teoretic sa fie identificate toate sensurile atestate ale unui anumit cuvânt.

Din punct de vedere metodologic, studiul nostru s-a bazat pe corpusul paralel multilingv „1984" si pe dictionarul multilingv extras din acest corpus. Cele 7 limbi ale experimentului nostru fac parte din patru familii: germanica (engleza), romanica (româna), slavica (bulgara, ceha si slovena) si ugro-finica (estoniana, maghiara). Desi corpusul contine un text beletristic, textul orwelian ca si traducerile sale în celelalte limbi nu sunt foarte stilizate si, ca atare, ofera un esantion rezonabil de limba moderna, comuna. Mai mult, traducerile textului original, efectuate de translatori avizati (unii dintre ei fiind apreciati scriitori), reflecta riguros originalul: pentru mai mult de 95% din textul englezesc o fraza sursa este tradusa în celealte limbi tot ca o singura fraza. Tipurile de alinieri frazale existente în corpusul „1984" sunt prezentate în tabela de mai jos si discutate în (Tufis&Barbu, 2001b):

Estoniana-Engleza			**Maghiara-Engleza**			**Romanian-Engleza**		
Tip	**Nr.**	**Proc**	**Tip**	**Nr.**	**Proc**	**Tip**	**Nr.**	**Proc**
3-1	2	0.030321%	7-0	1	0.014997%	3-1	3	0.046656%
2-2	3	0.045482%	4-1	1	0.014997%	2-4	1	0.015552%
2-1	60	0.909642%	3-1	7	0.104979%	2-3	3	0.046656%
1-3	1	0.015161%	3-0	1	0.014997%	2-2	2	0.031104%
1-2	100	1.516070%	2-1	108	1.619676%	2-1	85	1.321928%
1-1	**6426**	**97.422681%**	1-6	1	0.014997%	2-0	1	0.015552%
1-0	1	0.015161%	1-5	1	0.014997%	1-5	1	0.015552%
0-2	1	0.015161%	1-2	46	0.689862%	1-3	14	0.217729%
0-1	2	0.030321%	**1-1**	**6479**	**97.165573%**	1-2	259	4.027994%
			0-4	**1**	0.014997%	1-1	6047	94.043551%
			0-2	**3**	0.044991%	0-3	2	0.031104%
			0-1	19	0.284943%	0-2	2	0.031104%
						0-1	10	0.155521%

Bulgară- Engleză			Cehă- Engleză			Slovena- Engleza		
2-2	2	0.030017%	4-1	1	0.015029%	3-3	1	0.014970%
2-1	23	0.345190%	3-1	2	0.030057%	2-1	48	0.718563%
1-2	72	1.080594%	2-1	109	1.638112%	1-5	1	0.014970%
1-1	**6558**	**98.424134%**	1-3	2	0.030057%	1-2	53	0.793413%
0-1	8	0.120066%	1-2	81	1.217313%	**1-1**	**6572**	**98.383234%**
			1-1	6438	96.753832%	1-0	2	0.029940%
			0-1	21	0.315600%	0-1	3	0.044910%

Figura 6: Distributia tipurilor de aliniere frazala în corpusul paralel ”1984”

Alinierile de tipul N:M reprezinta situatiile în care M fraze din limba engleza au fost traduse cu N fraze în limba respectiva. Un caz particular îl reprezinta situatiile de omisiune în traducere (0:M) sau de inserare de text fara corespondent în original (N:0).

3.3. Experimentul initial

Textul original „1984” contine 7.069 leme diferite, iar dictionarul multilingv extras prin metoda descrisa în prima parte a acestei lucrari contine 1.233 de intrari. Aceste intrari au fost retinute respectând conditia ca un articol lexical din limba engleza sa aiba traduceri (eventual multiple) în cât mai multe limbi tinta. Conditia impusa dictionarului multilingv este foarte restrictiva, având în vedere ca majoritatea dictionarelor bilingve extrase automat contin între 6000 si 7000 de intrari. Intrari tipice (partiale) în dictionarul multilingv sunt ilustrate în figura 7. O informatie suplimentara, ce nu apare în exemplificarea din figura 7 este multimea tuturor unitatilor de traducere din corpusul paralel în care cuvântul englezesc a fost tradus prin echivalentii sai listati în dictionar. Dintre aceste intrari, au fost selectate 845 pentru care s-au gasit una sau mai multe traduceri în toate limbile. Dintre acestea, s-a ales o multime de 33 de substantive, acoperind toate gamele de frecventa si ambiguitate, cu care s-a realizat experimentul ale carui rezultate au fost validate de experti umani (Ide, *et al.,* 2001).

Engleză	**Categorie**	**Bulgară**	**Cehă**	**Estoniană**	**Maghiară**	**Română**	**Slovenă**
…	…	…	…	…	…	…	…
finally	R	íredr?	nakonec konečně	lõpuks viimaks	végül	în_cele_din_urmă până_la_urmă	končen nazadnje
…	…	…	…	…	…	…	…
wealth	N	áîarnnnâî	bohatství	jõukus	jólét	avuție	blaginja

		áëraî		rikkus	gazdagság	bogăție	bogastvo

Figura 7: Exemple de echivalentii de traducere identificati
în corpusul paralel „1984”

Pentru fiecare substantiv din acest esantion au fost extrase toate frazele englezesti în care apare, împreuna cu toate frazele corespunzatoare din celelalte limbi si pentru fiecare ocurenta a sa a fost construit un vector binar reprezentând toate traducerile posibile ale cuvântului respectiv. O valoare 1 în pozitia *n* a acestui vector semnifica faptul ca acea ocurenta a fost tradus prin cuvântul ce reprezinta a *n*-a traducere posibila. O valoare 0 semnifica faptul ca a *n*-a traducere posibila nu a fost folosita. De pilda pentru substantivul *„wealth”* (vezi figura 7) au fost depistate 11 traduceri posibile (2 în bulgara, estoniana maghiara româna si slovena, 1 în ceha). Un vector asociat oricarei ocurente a lui *wealth* va avea prin urmare 11 pozitii. Astfel, daca a *m*-a aparitie în textul original al romanului „1984” a cuvântului *wealth* are atasat vectorul 10101010101 acest lucru semnifica faptul ca în varianta bulgareasca el a fost tradus cu *áîarnnnâî*, în cea ceha cu *bohatství*, în cea estoniana cu *rikkus*, în cea maghiara cu *gazdagság*, în cea româna cu *bogatie* iar în cea slovena cu *bogastvo*. Vectorii astfel definiti au fost prelucrati cu un algoritm de clasificare de tip aglomerativ (Stolcke, 1996), 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 clasele rezultate prin dezambiguizarea manuala efectuata, independent de 2 vorbitori nativi ai limbii engleze. Dezambiguizarea manuala a fost efectuata utilizând numerotarea sensurilor din WordNet 1.6.

Pentru a putea compara rezultatele produse de dezambiguizatorii umani (numiti în continuare adnotatori) cu cele produse de algoritmul nostru, datele au fost normalizate în felul urmator: pentru fiecare adnotator si pentru algoritm fiecare din cele 33 de cuvinte a fost reprezentat printr-un vector binar de lungime $n(n-1)/2$, unde n este numarul de ocurente ale cuvântului în tot corpusul. Pozitiile în vector reprezinta o asignare de tip “DA/NU” indicând daca ocurenta respectiva a fost clasificata la fel de catre adnotatori, respectiv algoritm. Rezultatele acestui prim experiment sunt rezumate în tabelul din figura 8 indicând procentul de acord între clasificarile propuse de algoritm si cele ale fiecarui adnotator, acordul dintre cei doi adnotatori si acordul dintre toti 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: Concordanta între diferite clasificari

3.4. Cel de-al doilea experiment

Rezultatele primului experiment au aratat ca metoda discriminarii sensurilor folosind echivalentii de traducere este foarte competitiva, acuratetea procesului fiind comparabila (si uneori superioara) cu performantele obtinute de alti cercetatori ce au folosit ca referinta acelasi dictionar (Wordnet). Mai mult, diferentele de acord asupra clasificarii dintre cei 2 adnotatori pe de o parte si dintre fiecare adnotator si algoritm pe de alta parte este de numai 10-13%, ceea ce din nou este foarte competitiv în raport cu scorurile obtinute în alte experimente.

Pentru a valida aceste rezultate empirice, în cea de a doua faza a experimentului a fost luat în considerare un numar dublu de substantive (76) dintre cele „dificile", adica cu grad de ambiguitate mare, atât din clasa celor abstracte cât si a celor concrete (de exemplu, "thought", "stuff", "meaning", "feeling" respectiv "hand", "boot", "glass", "girl" etc.). Am ales acele substantive care au aparut cel putin de 10 ori în corpus (pentru a elimina efectul de „insuficienta a datelor") si în plus care au cel putin 5 traduceri în cele 6 limbi tinta. Restrictia de 10 aparitii a aparut din pragul înalt de confidenta pe care l-am impus procesului de extragere a echivalentilor de traducere:

$$LL(T_T, T_S) = 2 * \sum_{j=1}^{2} \sum_{i=1}^{2} n_{ij} * \log \frac{n_{ij} * n_{**}}{n_{i*} * n_{*j}} \geq 18$$

În plus, pentru adnotarea manuala au fost cooptati înca doi vorbitori nativi ai limbii engleze, astfel încât fiecare dintre ocurentele cuvintelor selectate a fost etichetata, în mod independent, de 5 clasificatori: 4 adnotatori si algoritmul discutat aici. În tabela din figura 9 sunt rezumate datele si rezultatele de acord între cei 4 adnotatori:

Nr. de cuvânte	76
Nr. ocurențe	2399
Număr mediu de ocurențe-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 toți adnotatorii	345
Numărul mediu de sensuri pe cuvânt	4.53
Procent de acord între adnotatori	
Full agreement (4/4)	54.27
75% agreement (3/4)	28.13
50% agreement (2/4)	16.92
No agreement	0.66

Figura 9: Datele experimentului si acordul între 4 adnotatori umani independenti

Rezultatele produse de algoritmul de clasificare si clasificarile realizate de adnotatori prin asignarea sensurilor din Wordnet1.6 au fost de data aceasta normalizate în mod diferit, prin ignorarea etichetei puse de adnotatori si considerând doar clasele rezultând din aceasta etichetare. Pentru a clarifica acest aspect sa urmarim modul în care doi dintre adnotatori au dezambiguizat cele 7 ocurente ale cuvântului "youth":

Ocurența nr.	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Adnotatorul 1	3	1	6	3	6	3	1
Adnotatorul 2	2	1	4	2	6	2	1

Figura 10: Acordul de clasificare pentru cuvântul " youth" între 2 adnotatori umani independenti

Acordul între cei doi adnotatori este doar de 43% (doar ocurentele 2, 5 si 7 au asignate sensuri consensuale); totusi, ambii adnotatori au clasificat ocurentele 1, 4, and 6 ca având acelasi sens, desi primul le-a etichetat cu sensul 3 din Wordnet, în timp ce al doilea le-a etichetat cu sensul 2. Daca însa ignoram eticheta clasificarea celor 3 ocurente este consistenta, în sensul ca ambii adnotatori au decis ca ele au acelasi sens. Acordul de clasificare se dubleaza în acest caz[87], iar datele sunt mult mai usor de comparat cu rezultatele produse de algoritm.

În acest al doilea experiment am luat în considerare determinarea momentului optim de oprire a clasificarii aglomerative. În primul experiment, am folosit o distanta minima predefinita, pentru determinarea numarului de sensuri între care se realizeaza discriminarea. Aceasta solutie nu tinea însa cont de proprietatile individuale ale cuvintelor (numarul maxim de sensuri, prescrise de Wordnet, frecventa de aparitie a cuvântului, numarul mediu de traduceri pe care le-a primit cuvântul în corpus). Noul algoritm de clasificare a fost modificat astfel încât sa-si calculeze un numar optim de clase[88], optimalitatea fiind judecata în raport cu numarul mediu de clase identificate de adnotatori Drept criteriu de oprire am folosit distanta minima dintre clasele existente la fiecare pas de aglomerare. La un pas de aglomerare, clasele cu cea mai mica distanta relativa sunt reunite într-o clasa mai mare. Procesul începe cu fiecare ocurenta într-o clasa distincta si se opreste

[87] *Singurul dezacord ramas consta în faptul ca Adnotatorul 1 considera ocurentele 3 si 5 ca având acelasi sens, în timp ce Adnotatorul 2 atribuie un sens diferit ocurentei 3—în fapt, realizând o discriminare mai fina între sensurile celor doua ocurente.*

[88] *În principiu, limita superioara a numarului de sensuri pe care îl poate avea un cuvânt englezesc într-un text este data de numarul de sensuri listate în Wordnet; dupa cum era de asteptat însa nu exista în corpusul nostru nici un exemplu în care vreun cuvânt polisemantic sa fi aparut cu toate sensurile din WordNet.*

când distantele relative între clasele existente este „suficient" de mare. Distanta dintre doua clase se calculeaza pe baza vectorilor caracteristici (centroizii) ai celor doua clase (evident depinzând de cuvânt, de numarul de ocurente si de numarul de sensuri ale cuvântului clasificat):

$$dist(v_1, v_2) = \sqrt{\sum_{i=1}^{n} (v_1(i) - v_2(i))^2}$$

Cele mai bune rezultate în discriminarea automata au fost obtinute pe cale experimentala, impunând drept criteriu de oprire a algoritmului conditia:

$$\frac{mindist(k) - mindist(k+1)}{mindist(k+1)} < 0.12$$

în care *mindist(k)* reprezinta distanta minima între clasele existente la pasul *k* de aglomerare.

Pentru medierea opiniilor adnotatorilor am definit o adnotare de referinta reprezentând clasificarea majoritara între cei 4. În cazul egalitatii de voturi, adnotatorul care a fost în cele mai multe cazuri de aceeasi opinie cu majoritatea a impus clasa. Folosind aceasta clasificare mediata si raportând-o la clasificarea produsa de algoritm pentru cele 76 de cuvinte, am analizat diferentele de clasificare, considerate ca fiind erori. Marea majoritate a erorilor de clasificare pentru cele 2399 de ocurente au aparut în cazul cuvintelor pentru care distributia sensurilor este foarte inegala; ca urmare am adaugat algoritmului o faza suplimentara de postprocesare, în care clasele cu un numar mult mai mic de ocurente decât clasa cu cele mai multe ocurente au fost incorporate în ultima. Raportul minim între numarul de ocurente al celei mai mari clase si numarul de ocurente din clasele potential absorbabile în cea dintâi a fost ales empiric ca fiind 10^{89}. Motivatia acestei euristici, consta în constatarea facuta de mai multi cercetatori în domeniul lingvisticii corpusului (fapt sugerat chiar de Zipf cu peste 50 de ani în urma) ca utilizarea frecventa a unui cuvânt într-un text omogen tinde sa-i pastreze sensul.

Cu aceasta noua euristica încorporata, algoritmul de clasificare a atins cifra de 74,6% acord cu clasificarea mediata. În (Ide *et al.* 2002) sunt prezentate alte variante ale algoritmului care au condus prin evaluarea empirica la versiunea sa finala. Clasele produse de fiecare pereche de clasificatori (om sau masina) au fost evaluate printr-un algoritm ce calculeaza alinierea claselor astfel încât intersectia lor sa fie maximala. Diferentele dintre doua clase astfel aliniate au fost considerate dezacorduri de clasificare. Scorul de acord a fost calculat ca fiind raportul dintre suma numarului de ocurente comune pentru fiecare clasa aliniata si numarul total al ocurentelor cuvântului respectiv. In tabela din figura 10 este exemplificat modul de calcul al acordului dintre clasificarea produsa de algoritm si clasificarea mediata a adnotatorilor pentru cuvântul *movement*. Acesta a aparut în text de 40 de ori. Atât algoritmul cât si cei patru adnotatori au identificat 4 sensuri distincte în care

[89] *10 este o valoare precauta; experimente viitore, de mai mare anvergura ar putea furniza argumente pentru coborârea acestui prag.*

acest cuvânt a fost utilizat. Asa cum se vede din figura 10, cea mai numeroasa clasa (clasa 1) contine în clasificarea mediata 28 dintre cele 40 de aparitii ale cuvântului *movement*, în timp ce clasa corespondenta creata de algoritm contine doar 25 de ocurente. Dintre acestea, 24 sunt comune cu cele din clasa 1 a clasificarii mediate. În conformitate cu definitia anterioara a scorului de acord, pentru acest exemplu algoritmul a produs rezultatul corect în 85% din cazuri.

CLASA	1	2	3	4	Σ
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 produsa de algoritm pentru cuvântul *movement*

3.5. Rezultate

Rezultatele obtinute cu ultima varianta a clasificatorului în cel de-al doilea experiment sunt sintetizate in tabelul din figura 12. Tabelul indica procentul de acord între diverse clasificari: 1, 2, 3, 4, reprezinta clasificarile realizate de adnotatorii umani, M reprezinta clasificarea mediata a clasificatorilor umani, A reprezinta clasificarea produsa de algoritm, iar B este referinta de baza (baseline) care presupune toate ocurentele unui cuvânt ca având acelasi 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						77.3

Figura 12: Acorduri între diverse clasificari

Tabela arata ca acordul între adnotatorii umani comparat cu cel dintre algoritm si adnotatorii umani (cu exceptia unuia dintre ei (4), pe care îl suspectam ca a vazut clasificarile celorlalti trei si în consecinta si-a revizuit unele decizii) nu difera substantial. Acest lucru demonstreaza (cel putin în raport cu datele experimentului nostru) ca dezambiguitarea automata este comparabila ca acuratete cu cea efectuata de adnotatori

umani. Diferenta fundamentala consta în faptul ca programul a terminat în circa 2 minute clasificarea pentru care adnotatorilor le-au trebuit între 4 si 5 saptamâni.

Experimentul descris a evaluat dezambiguizarea automata a cuvintelor englezesti pornind de la traducerea lor în celelalte 6 limbi. Aceasta directionare a fost impusa doar de disponibilitatea pentru limba engleza a textului dezambiguizat de experti umani (vorbitori nativi ai limbii engleze). Întrucât algoritmul de clasificare nu depinde în nici un mod de limba pentru care se realizeaza dezambiguizarea (limba tinta) si nici de limbile martor în raport cu care se face acest proces, rezulta ca exact acelasi procedeu descris pâna aici poate fi folosit pentru dezambiguizarea cuvintelor românesti folosind echivalentii lor de traducere în engleza, bulgara, ceha, estoniana, maghiara si slovena, ori pentru dezambiguizarea cuvintelor bulgaresti pe baza echivalentilor lor de traducere în celelalte 6 limbi. Întrucât sensul este (în principiu) un invariant al traducerii, nu pare a se justifica si pentru celelalte limbi efortul de adnotare umana facut pentru limba engleza. Este rational a presupune ca rezultate similare (raporturi relative) s-ar obtine indiferent de limba tinta si de limbile martor.

Sa mai mentionam si faptul ca exista o anumita corelatie (factorul Spearman - 0.51) între numarul de sensuri în Wordnet ale unui cuvânt si nivelul de acord între diferitele clasificari ale ocurentelor sale. Cele mai scazute scoruri de acord au fost obtinute pentru “line” (29 sensuri), “step” (10), position (15), “place” (17) si “corner” (11). Acorduri perfecte s-au obtinut pentru majoritatea cuvintelor cu mai putin 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 si de algoritm, a fi fost folosite cu un singur sens în tot textul. Pe de alta parte, gradul de acord pentru câteva cuvinte cu mai putin de 5 sensuri (“rubbish” (2), “rhyme” (2), “destruction” (3) si “belief” (3)) a fost semnificativ mai mic decât media pentru toate perechile de clasificari (adnotator-adnotator, adnotator-algoritm). Concluzia a fost ca pentru unele cuvinte, distinctiile de sens sunt atât de fine în Wordnet, încât chiar vorbitorii nativi (si cu atât mai mult algoritmul de clasificare) nu pot face diferentieri sistematice de sens ale diferitelor ocurente ale acestor cuvinte. O astfel de hiperdiferentiere a sensurilor este în imensa majoritate a cazurilor irelevanta pentru aplicatiile de prelucrare a limbajului natural.

4. Concluzii

Rezultatele experimentelor noastre arata ca acuratetea discriminarii sensurilor pe baza echivalentilor de traducere extrasi din corpusuri paralele este comparabila cu cea produsa de adnotatori umani. Întrucât abordarea noastra este complet automatizata ea poate fi folosita la crearea de volume mari de texte, având discriminate sensurile cuvintelor polisemantice. Utilizarea expertilor umani este prohibitiva sub aspectul costului si al timpului de realizare a unei asemenea sarcini, iar procentajul suplimentar de acuratete, presupus de activitatea umana, este prea mic pentru a justifica procedurile manuale.

Metoda pe care am descris-o în aceasta lucrare nu eticheteaza clasele de ocurente ale unui cuvânt cu un numar de sens ales dintr-un inventar prescris de sensuri iar majoritatea aplicatiilor de prelucrare a limbajului natural (de pilda clasificarea textelor, regasirea informatiilor, rezumarea automata etc.) nici nu au nevoie de aceasta informatie suplimentara; pentru aceste tipuri de aplicatii este suficient a decide ca doua sau mai multe ocurente ale unui cuvânt sunt folosite în acelasi sens sau nu. O etichetare conventionala a sensurilor identificate pentru un anumit cuvânt ar putea sa se bazeze pe frecventa sensurilor respective (sensul 1 corespunzând clasei cu cele mai multe ocurente). Evident o astfel de etichetare depinde de registrul lingvistic al textului pe baza caruia se identifica sensurile distincte.

O directie foarte promitatoare (Tufis, 2002b) , (Tufis&Cristea, 2002a,b) o constituie utilizarea metodologiei prezentata aici în constructia si validarea ontologiilor multilingve de tip EuroWordNet. Folosind echivalentii de traducere si clasificarea ocurentelor echivalente din punct de vedere al sensului se poate verifica daca proiectia interlinguala a doua sau mai multe dictionare semantice este corecta. Aceasta presupune ca sensurile cuvintelor extrase ca echivalenti de traducere ai cuvintelor englezesti dezambiguizate sa fie puse în corespondenta cu acelasi concept interlingual apartinând Indexului Interlingual (ILI - vezi Tufis&Cristea, 2002b – în acest volum). În cazul contrar (echivalentii de traducere sunt pusi în corespondenta cu concepte interlinguale diferite) este fie vorba de o eroare propriu-zisa de proiectie conceptuala într-unul sau mai multe dintre dictionarele semantice aliniate ori conceptele interlinguale sunt atât de apropiate semantic încât se poate propune unificarea lor într-un concept mai general cu lexicalizare în mai multe limbi. Aceasta este esenta conceptului de „soft-clustering" definit în comunitatea EuroWordNet. Fata de identificarea prin metode statistice a conceptelor interlinguale foarte apropiate semantic, analiza prin metoda echivalentilor de traducere si a discriminarii sensurilor a proiectiilor sensurilor facute de lexicografi profesionisti peste o multime de sensuri conceptualizate în ILI este mult mai robusta. Experimentele preliminare discutate în (Tufis, 2002b) au aratat ca în diferite limbi pentru care se realizeaza o ontologie lexicala multilingva (bulgara, ceha, greaca, româna, sârba, turca) exista dificultati identice de proiectie conceptuala a sensurilor unor cuvinte din limbile considerate. Faptul ca aceleasi concepte interlinguale creeaza acelasi tip de dificultate în proiectia sensurilor unor cuvinte apartinând unor limbi foarte diferite indica cu claritate ca acele concepte trebuie generalizate.

Un alt aspect care merita subliniat este ca metodologia prezentata aici, corelata cu existenta a tot mai multor dictionare semantice de tip Wordnet, ce adera la principiul EuroWordNet de aliniere la Indexul Interlingual, va permite dezvoltarea de corpusuri adnotate semantic (de tipul SemCor) pentru orice limba. Tranzitivitatea relatiilor de tip „EQ-SYN" folosite în proiectia sinseturilor unui wordnet monolingv peste ILI, corelata cu echivalenta de traducere (relatie tot între sensuri) extrasa dintr-un corpus paralel, în care textul dintr-una din limbi este adnotat semantic, permite importul adnotarilor în toate celelalte limbi. Deoarece limba din care se importa adnotarea semantica nu este relevanta pentru aceasta procedura, rezulta ca eforturile depuse de-a lungul timpului în crearea celor

câteva corpusuri cu adnotare semantica pentru limbile „mari" pot fi valorificate pentru orice alta limba în care exista (sau se creeaza) traduceri ale textelor din corpusurile adnotate. Mai mult, se poate imagina crearea unui consortiu multilingv care sa aleaga un corpus paralel în cât mai multe limbi cu scopul de a-l adnota semantic. Prin adnotarea independenta, în fiecare limba, a unor portiuni distincte din corpusul paralel, folosind o metodologie de genul celei prezentate în aceasta lucrare (si desigur având un dictionar semantic multilingv de tip EuroWordNet) adnotarile sectiunilor monolingve vor putea fi importate în sectiunile corespunzatoare ale tuturor celorlalte texte monolingve, în final putându-se obtine adnotarea semantica, consistenta, a întregului text din fiecare limba a corpusului paralel.

Multumiri

Rezultatele prezentate în aceasta lucrare sunt rodul mai multor proiecte internationale de cercetare desfasurate la Institutul de Inteligenta Artificiala, alaturi de colegii Ana Maria Barbu, Eduard Barbu, Radu Ion, Catalin Mititelu, Octavian Popescu. De asemenea, colaborarea cu Nancy Ide de la Universitatea Vassar din Poughkeepsie, SUA, si cu Tomaz Erjavec de la Institutul „Jozef Stefan" din Ljubljana, Slovenia, parteneri în proiectele amintite, a fost si este extrem de productiva. Tuturor le aduc aici cuvenitele multumiri.

Referinte bibliografice

Ahrenberg, L., M. Andersson, M. Merkel. (2000). "A knowledge-lite approach to word alignment", in (Véronis, 2000: 97-116)

Brants, T.(2000) "TnT – A Statistical Part-of-Speech Tagger", in *Proceedings of the Sixth Applied Natural Language Processing Conference, ANLP-2000,* April 29 – May 3, 2000, Seattle, WA

Brew, C., McKelvie, D. (1996). "Word-pair extraction for lexicography", http:///www.ltg.ed.ac.uk/ ~chrisbr/papers/nemplap96

Brown, P., Della Pietra, S. A., Della Pietra, V. J., Mercer, R. L. (1993). "The mathematics of statistical machine translation: parameter estimation" in *Computational Linguistics*19(2): 263-311.

Carletta, J. (1996). Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, 22/2, 249:254.

Cristea, D., Dima, G. E. (2001). „An Integrating Framework for Anaphora Resolution", Journal on *Information Science and Technology, Romanian Academy Publishing House*, Bucharest, vol. 4, no. 3, 273:292.

Dagan, I., Itai, A. (1994). Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics,* 20/4, 563:596.

Dagan, I., Itai, A., Schwall, U. (1991). Two languages are more informative than one. *Proceedings of the 29th Annual Meeting of the ACL,* 18-21 Berkeley, California, 130:137.

Dunning, T. (1993), "Accurate Methods for the Statistics of Surprise and Coincidence" in *Computational Linguistics*19(1),61:74.

Dyvik, H. (1998). Translations as Semantic Mirrors. *Proceedings of Workshop Multilinguality in the Lexicon II, ECAI 98*, Brighton, UK, 24:44.

Erjavec T., Ide, N. (1998). "The Multext-East corpus". In *Proceedings of First International Conference on Language Resources and Evaluation*, Granada, Spain, 971:974

Erjavec. T. (2002) "An Experiment in Automatic Bi-Lingual Lexicon Construction from a Parallel Corpus", Proceedings of the 7th TELRI International Seminar on Corpus Linguistics, Dubrovnik, Croatia (forthcoming).

Fellbaum C. (1998) Wordnet:An Electronic Lexical Database, MIT Press, 423p.

Gale, W.A., K.W. Church, (1991). "Identifying word correspondences in parallel texts". In Fourth *DARPA Workshop on Speech and Natural Language*, 152:157

Gale, W.A., K.W. Church, (1993). "A Program for Aligning Sentences in Bilingual Corpora". In In *Computational Linguistics*, 19(1), 75:102

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.

Hinrics, H., Trushkina, J. (2002): "Forging Agreement: Morphological Disambiguation of Noun Phrases", Proceedings of the Workshop on Treebanks and Linguistic Theories 2002, Sozopol, Bulgaria, 1:18.

Ide, N. (1999). Cross-lingual sense determination: Can it work? *Computers and the Humanities,* 34/1-2, 223:234.

Ide, N., Erjavec, T., and Tufis, D. (2001). Automatic sense tagging using parallel corpora. *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium,* Tokyo, 83:89.

Ide, N., Erjavec, T., Tufis, D. (2002).: „Sense Discrimination with Parallel Corpora" in Proceedings of the SIGLEX Workshop on Word Sense Disambiguation: Recent Successes and Future Directions. ACL2002, July Philadelphia, 56:60

Hiemstra, D. (1997). "Deriving a bilingual lexicon for cross language information retrieval". In *Proceedings of Gronics,* 21:26

Kay, M., Röscheisen, M. (1993). "Text-Translation Alignment". In *Computational Linguistics*, 19/1, 121:142

Kupiec, J. (1993). "An algorithm for finding noun phrase correspondences in bilingual corpora". In *Proceedings of the 31st Annual Meeting of the Association of Computational Linguistics*, 17:22

Miller, G. A., Beckwith, R. T. Fellbaum, C. D., Gross, D. and Miller, K. J. (1990). WordNet: An on-line lexical database. *International Journal of Lexicography,* 3/4, 235:244.

Melamed, D. (2001). "*Empirical Methods for Exploiting Parallel Texts*", MIT Press, 373p.

Resnik, P. and Yarowsky, D. (2000). Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. *Journal of Natural Language Engineering*, 5(2), 113:133.

Resnik, P. and Yarowsky, D. (1997). 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., 79:86.

Smadja, F. (1993). "Retrieving Collocations from Text:Xtract". In *Computational Linguistics*, 19/1, 142:177

Smadja, F., K.R. McKeown, and V. Hatzivassiloglou (1996). "Translating collocations for bilingual lexicons: A statistical approach". *Computational Linguistics*, 22/1, 1:38

Stolcke, A. (1996). Cluster 2.9. http://www.icsi.berkeley.edu/ftp/global/pub/ai/stolcke/software/cluster-2.9.tar.Z.

Tiedemann, J. (1998). "Extraction of Translation Equivalents from Parallel Corpora", In *Proceedings of the 11th Nordic Conference on Computational Linguistics*, Center for Sprogteknologi, Copenhagen, 1998, http://stp.ling.uu.se/~joerg/

Tufis, D., Barbu, A.M., Patrascu, V., Rotariu, G., Popescu, C. (1997) "Corpora and Corpus-Based Morpho-Lexical Processing " in D. Tufis, P. Andersen (eds.) "Recent Advances in Romanian Language Technology", Editura Academiei, 35:56

Tufis, D., Mason, O., (1998). "Tagging Romanian Texts: a Case Study for QTAG, a Language Independent Probabilistic Tagger". In Proceedings of *the First International Conference on Language Resources & Evaluation (LREC),* Granada, Spain, 589:596.

Tufis, D. (1999). "Tiered Tagging and Combined Classifiers" In F. Jelinek, E. Nöth (eds) Text, Speech and Dialogue, Lecture Notes in Artificial Intelligence 1692, Springer, 29:33

Tufis, D., Ide, N. Erjavec, T. (1998). "Standardized Specifications, Development and Assessment of Large Morpho-Lexical Resources for Six Central and Eastern European Languages". In *Proceedings of First International Conference on Language Resources and Evaluation (LREC)*, Granada, Spain, 233:240

Tufis, D. (2000) “Using a Large Set of Eagles-compliant Morpho-Syntactic Descriptors as a Tagset for Probabilistic Tagging. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC),* Athens, Greece, 1105:1112

Tufis, D., Dienes, P., Oravecz, C., Váradi T. (2000). “Principled Hidden Tagset Design for Tiered Tagging of Hungarian” *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC),* Athens, Greece, 1421:1426

Tufis, D., Barbu, A.M., (2001a). “Extracting multilingual lexicons from parallel corpora” in Proceedings of the ACH/ALLC 2001, New York University, 42:46

Tufis, D., Barbu, A.M. (2001b). "Automatic Learning of Translation Equivalents" in "Romanian Journal on Information Science and Technology", Romanian Academy, vol.4, no. 3-4, 325:351.

Tufis, D., Barbu, A.M., (2002). „Revealing translators knowledge: statistical methods in constructing practical translation lexicons for language and speech processing”, in *International Journal of Speech Technology*. Kluwer Academic Publishers, no.5, 199:209.

Tufis, D.(2001a). “Building an ontology from a large Romanian dictionary of synonyms by importing Wordnet relations”, RACAI Research report, June, 68 pp.

Tufis, D. (2001b). “Partial translations recovery in a 1:1 word-alignment approach”, RACAI Research report, June, 32pp.

Tufis, D. (2002a). ”A cheap and fast way to build useful translation lexicons” in *Proceedings of the 19th International Conference on Computational Linguistics, COLING2002,* Taipei, China, 246:251

Tufis, D. (2002b). “Interlingual alignment of parallel semantic lexicons by means of automatically extracted translation equivalents”, Proceedings of the 7th TELRI International Seminar on Corpus Linguistics, Dubrovnik, Croatia (forthcoming)

Tufis, D., Cristea, D. (2002a). ”Methodological issues in building the Romanian Wordnet and consistency checks in Balkanet”, In Proceedings of LREC2002 Workshop on Wordnet Structures and Standardisation, Las Palmas, Spain, 35-41

Tufis, D., Cristea, D. (2002b). ”RO-BALKANET - ontologie lexicalizata, în context multilingv, pentru limba româna”, în acest volum.

Varadi, T.(2002) The Hungarian National Corpus, Proceedings of LREC2002, Las Palmas, Spain, 385:389

Véronis, J. (ed), (2000). Parallel Text Processing. Text, Speech and Language Technology Series, Kluwer Academic Publishers Vol. 13, 2000

Yarowsky, D., Florian. R. (1999). Taking the load off the conference chairs: towards a digital paper-routing assistant. *Proceedings of the Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora,* 220:230.

Zipf, G.K., (1936). "The Psycho-biology of Language: an Introduction to Dynamic Philology". Routledge, London, UK

Referentialitate si cursivitate în relatie cu structura de discurs

Dan CRISTEA

Universitatea "Al.I.Cuza" Iasi, Facultatea de Informatica
Academia Româna, Institutul de Informatica Teoretica - Filiala Iasi

1. Introducere

În ultimii 25 de ani s-a studiat enorm pentru a se întelege ce anume face dintr-un text (considerat o secventa de propozitii sintactic corecte) sa fie un discurs, asadar de ce un discurs e coerent si ce elemente îi atribuie coeziune. Dintre teoriile computationale ale discursului, trei au avut o influenta covârsitoare asupra dezvoltarilor ultimilor ani din acest domeniu: teoria structurilor retorice, teoria starilor atentionale si teoria centrelor.

Dezvoltata initial din perspectiva generarii textelor, teoria structurilor retorice (*rhetorical structure theory*, de aici încolo RST), a fost elaborata de Mann si Thompson între 1986 si 1987 ca o teorie a organizarii textelor [Mann, Thompson, 1988, Hovy, 1988, Scott, de Souza, 1990]. Ea caracterizeaza structura de discurs în termeni de relatii ce leaga parti componente ale textului. Unitatea elementara de discurs în RST este, de regula, o propozitie ce, la nivel semantic, formuleaza o predicatie. O structura de discurs este descrisa de o **schema**. Ea grupeaza o secventa de unitati, sau de unitati si scheme, sau o secventa de scheme. Dintr-un anumit punct de vedere o schema poate fi asemuita cu o regula a unei gramatici, ea relevând structura de constituenti a unui compus. O schema consta dintr-o **relatie** (27 în RST) care leaga doua sau mai multe întinderi de text, fiecare dintre ele având, la rândul lor, o structura (constituentii schemei). Un discurs este fie o unitate, care este o întindere de text elementar, fara structura, fie o schema (un text mai lung decât o singura unitate si care manifesta o structura). Relatiile pot fi de doua tipuri: **hipotactice** – daca argumentele sunt constituenti neegali ca importanta si **paratactice** (sau echinucleare) – daca constituentii pe care-i agrega sunt egali ca importanta. Între constituentii uniti de relatiile hipotactice exista întotdeauna unul singur mai important, numit **nucleu**, ceilalti fiind numiti **sateliti**. La relatiile paratactice, prin conventie, se considera ca toti constituentii sunt nucleari. Satelitul este, în general, mai susceptibil de a fi schimbat sau eliminat complet decât nucleul, fara ca, prin aceasta, întelesul discursului sa se modifice. Dimpotriva înlocuirea sau stergerea nucleului este o optiune mult mai drastica, care poate duce la denaturari ale întelesului. Relatiile hipotactice sunt în general cele intentionale, în care o întindere de text comunica un scop si o alta exprima un subscop ce

completeaza, dezvolta etc. scopul principal. Pe de alta parte, relatiile paratactice sunt, în general, de natura informationala, simetrice, neputându-se stabili daca si care componenta predomina.

În RST accentul este pus pe performanta retorica: prin ce mijloace un scriitor (sau vorbitor) reuseste sa convinga un cititor (ascultator) de intentiile 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 îmbunatatirii setului de relatii propus initial de teorie. Într-adevar pare extrem de convenabil, inclusiv din punct de vedere computational, sa vedem discursul reprezentat ca un arbore, în care nodurile terminale sa reconstituie, în secventa lor, textul. Cu toate acestea RST nu aduce nici o lumina în privinta vreunei legaturi care ar exista între structura si referentialitate. RST este deci o teorie asupra structurii globale a discursului.

Teoria starilor atentionale (*atentional state theory*, AST) [Grosz, Sidner, 1986] reprezinta o dezvoltare a liniei de cercetare în discurs dominata de Barbara Grosz si Candace Sidner asupra manierii în care focarul ori centrul de discurs (*focus* în engleza) se modifica pe parcursul derularii textului si a recunoasterii intentiilor comunicate de discurs [Grosz, 1981, Sidner, 1983]. Grosz si Sidner nu cred ca varietatea atât de mare a intentiilor ce pot fi comunicate de un discurs poate fi condensata într-un numar fix de sabloane retorice exprimate ca relatii, ca în RST sau tentative similare acesteia. Teoria se doreste a fi un model formal, care se distanteaza de detaliile ce ar putea fi asociate participantilor la discurs. Realizând proiectii corespunzatoare utilizatorului de limbaj, însotite de detalii specifice, ea s-ar putea regasi atât în constructia unui sistem automat cît si într-o teorie psihologica, ambele consumatoare de limbaj natural. Desi recunoaste însemnatatea mesajului transmis de un discurs, teoria nu adreseaza problema întelesului discursului si 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 oricarei tentative de a aborda problema construirii sensului.

Conform lui Grosz si Sidner intentiile joaca rolul principal în explicarea structurii discursului, în timp ce dinamica atentiei joaca rolul principal în explicarea interpretarii discursului. Structura discursului are trei componente distincte, dar strâns corelate:

- o **structura lingvistica**, care face ca una sau mai multe propozitii, exprimari (*utterance* în engleza) sa fie agregate într-un segment de discurs iar limitele dintre segmente sa fie indicate de expresii lingvistice, intonatie, schimbari ale timpului si aspectelor verbelor. Segmentul de discurs are însa o definitie recursiva: un segment poate îngloba alte segmente, acestea pe altele s.a.m.d.;
- o **structura intentionala**, care face sa vedem discursul ca având un scop global (scopul discursului – SD), care este scopul fundamental al vorbitorului/scriitorului la emiterea discursului si fiecare segment al sau are un scop al segmentului (scopul segmentului de discurs – SSD) care este un subscop al scopului segmentului din care face el parte. Dintr-un punct de vedere intuitiv, SSD specifica cum contribuie respectivul subsegment la

realizarea scopului segmentului din care face el parte. Teoria admite ca nu exista o lista finita de scopuri ale discursului, care sa faca posibila o comparatie cu lista categoriilor gramaticale, de exemplu. Conform teoriei, doua relatii structurale sunt suficiente pentru a compune structura discursului: **relatia de dominare** (daca SSD_1 domina SSD_2 atunci SSD_2 contribuie la SSD_1, sau SSD_2 este intentionata sa satisfaca partial SSD_1) si **relatia de satisfacere-precedenta** (SSD_1 satisface-precede SSD_2 daca SSD_1 trebuie satisfacut înainte de SSD_2);

- o **stare atentionala**, prin care se asociaza fiecarui segment al discursului un spatiu al entitatilor aflate în centrul atentiei. Starea atentionala este o proprietate a discursului iar nu a participantilor la discurs. Ea reprezinta o trasatura dinamica a discursului, pastrând **obiecte, proprietati si relatii** ce sunt importante la fiecare moment al parcurgerii discursului. Starea atentionala e modelata printr-un **set de spatii ale centrelor atentiei**, în timp ce schimbarile ce pot avea loc în starea atentionala sunt restrictionate de un set de reguli de tranzitie care arata conditiile de adaugare si stergere a spatiilor. Colectia tuturor spatiilor centrelor de atentie ce sunt disponibile în fiecare moment al interpretarii unui discurs formeaza o structura a atentiei ce are dinamica unei **stive** si care ar fi capabila sa explice procesele implicate în procesarea discursului, inclusiv accesibilitatea referentiala: domeniul în care trebuie cautate entitatile de discurs referite în segmentul corespunzator starii atentionale aflate în vârful stivei este cel al starilor aflate în stiva.

Structura recursiva a segmentului de discurs din AST permite si aici acceptarea unei reprezentari arborescente, în cadrul careia cele doua relatii între segmente, de dominare si de satisfacere-precedenta, nu sunt altceva decât relatiile topologice normale pe orice structura de arbore: cea dintre parinte si orice fiu al sau si, respectiv, cea de ordine dintre frati. AST se constituie într-o teorie globala asupra structurii si a coeziunii discursului.

Cercetatori precum Moser si Moore [1996] sau Marcu [1999] pun în evidenta similaritati semnificative între AST si RST, inclusiv în ceea ce priveste maniera de reprezentare prin arbori a structurii de discurs, ceea ce permite combinarea puterii de reprezentare, mai fine în RST, datorita proliferarii relatiilor, cu implicatiile pe care structura le poate avea asupra referentialitatii, puse în evidenta de AST. Utilizând structura de segmente si stiva, ca mecanism de prelucrare, AST propune o maniera de a rezolva accesibilitatea referintelor anaforice printr-o transparenta pe verticala, de sus în jos, de-a lungul starilor atentionale ce se afla la un moment dat în stiva. Reprezentarea prin segmente din AST are însa o slabiciune: modelul stiva nu poate reflecta relatia de dominare atunci când scopul dominat corespunde unui segment care apare în text înaintea celui care domina [Ide, Cristea, 2000]. Sa remarcam ca defectul este unul de granularitate pentru ca identificarea segmentului dominat ce precede pe cel dominator cu însusi segmentul dominator elimina problema. AST nu e, asadar, capabila sa reprezinte segmente având o

granularitate oricât de fina: coborând de la o granularitate grosiera la una fina, exista o limita dincolo de care ne putem astepta la grave contradictii.

Teoria centrelor (*centering*, CT) [Grosz *et al*., 1995, Brennan *et al*., 1987] furnizeaza explicatii convingatoare asupra contextelor ce permit utilizarea pronumelor pentru realizarea referintelor si asupra ce anume face un discurs sa fie coerent. CT nu se aplica însa dincolo de limitele unui segment (vazut în acceptiunea din AST). Avem de a face, asadar cu o teorie locala asupra coeziunii si coerentei. Desi nu este definita riguros în teorie, în toate exemplele autorilor unitatea elementara a structurii lingvistice este fraza (*utterance*, exprimare). Abordari ulterioare întrevad posibilitatea de a considera o segmentare mai fina, la nivel de propozitie (v. [Kameyama, 1998] de exemplu). Noi vom considera drept unitate a structurii de discurs acelasi tip de întindere lexicala ca si în cazul RST, adica acea întindere ce la nivel sintactic este o propozitie iar la nivel semantic – o predicatie. Fiecare unitate de discurs u_n ce intra în compozitia unui segment este caracterizata de o lista de **centre anticipatoare** (*forward-looking*) notata $C_f(u_n)$. Centrele listei $C_f(u_n)$ sunt entitati semantice ce corespund, la nivelul textului, expresiilor referentiale cuprinse în unitatea u_n. Spunem ca o expresie referentiala **realizeaza** un centru. Elementele acestei liste sunt ordonate pentru a reflecta importanta relativa în u_n. Criteriile de ordonare a elementelor listei C_f, sunt, în forma originara a teoriei, de natura sintactica, desi alte abordari le diferentiaza în functie de limba (v. de exemplu [Walker *et al*., 1994] pentru japoneza, [deEugenio, 1990, de Eugenio, 1998] pentru italiana, sau [Strube, Hahn, 1996] pentru germana). Pentru limba engleza, autorii CT dau urmatorul criteriu: subiect > obiect-direct> obiect-indirect > complemente > adjuncti. Elementele listei $C_f(u_n)$ sunt acele entitati despre care se vorbeste în unitatea u_n si deci despre care e cel mai probabil ca se va continua sa se vorbeasca si în unitatea urmatoare, u_{n+1}, daca aceasta apartine aceluiasi segment ca si u_n. Cel mai bine plasat element al listei $C_f(u_n)$ se numeste **centru principal** si se noteaza $C_p(u_n)$. Fiecarei unitati îi este asociat un unic **centru retroactiv** *(backward-looking*), notat $C_b(u_n)$. Prin conventie, centrul retroactiv al primei unitati a segmentului este considerat centrul principal, în timp ce, pentru toate celelalte unitati ale segmentului, el este cel mai bine plasat element al listei C_f a unitatii precedente care este de asemenea realizat si în unitatea curenta.

Teoria face o clasificare a tranzitiilor posibile între unitati consecutive, din punctul de vedere al invariantei ori nu a centrelor retroactive si al identificarii ori nu a lor cu centrele principale. Astfel, cu exceptia cazului în care între unitati succesive ale aceluiasi segment nu exista centre comune, urmatoarele patru tipuri de tranzitii sunt posibile:

CONTINUARE (*continuing*, CON): $C_b(u_{n+1}) = C_b(u_n)$ si $C_b(u_{n+1}) = C_p(u_{n+1})$, corespunzând situatiei în care atât în u_n cât si în u_{n+1} se vorbeste despre aceeasi entitate si este de asteptat ca si în unitatea urmatoare sa se vorbeasca despre ea.

RETINERE (*retaining*, RET): $C_b(u_{n+1}) = C_b(u_n)$ dar $C_b(u_{n+1}) \neq C_p(u_{n+1})$, a carui interpretare este ca, desi atât în u_n cât si în u_{n+1} se vorbeste despre

aceeasi entitate, este de asteptat ca în unitatea urmatoare sa se vorbeasca despre o alta.

SCHIMBARE LINA (*smooth-shifting*, SSH): $C_b(u_{n+1}) \neq C_b(u_n)$ dar $C_b(u_{n+1}) = C_p(u_{n+1})$, cu semnificatia ca în u_n si în u_{n+1} nu se vorbeste despre aceeasi entitate si este de asteptat ca în unitatea urmatoare sa se vorbeasca despre entitatea despre care s-a vorbit ultima oara.

SCHIMBARE ABRUPTA (*abrupt-shifting*, ASH): $C_b(u_{n+1}) \neq C_b(u_n)$ si $C_b(u_{n+1}) \neq C_p(u_{n+1})$, cu semnificatia ca în u_n si în u_{n+1} nu se vorbeste despre aceeasi entitate si este de asteptat ca în unitatea urmatoare sa se vorbeasca despre o alta entitate decât ultima mentionata.

Nucleul CT este concentrat în doua reguli, prima enuntând o constrângere asupra formei de realizare a centrelor prin pronume, iar cea de a doua formulând preferinte asupra secventelor de tranzitii ale centrelor. Regula a doua, cea care se refera la coerenta, formuleaza presupunerea ca anumite secvente produc o încarcare inferentiala în ascultator mai mare decât altele:

Regula 2: Secventele de continuari sunt preferabile secventelor de retineri, care sunt preferabile secventelor de schimbari line, iar acestea sunt preferabile secventelor de schimbari bruste: CON > RET > SSH > ASH.

Daca ne abtinem de a penaliza CT, ca teorie locala, asadar aplicabila la întinderea unui segment, pe motivul fragilitatii notiunii de segment, care are o definitie recursiva (un segment este constituit din alte segmente), slabiciune mostenita de la AST, atunci apare naturala tentativa de a largi aplicabilitatea CT la întregul discurs, într-o maniera recursiva, pe chiar aceasta structura de segment, definita, ea însasi, recursiv. Teoria nervurilor propune o astfel de generalizare.

Teoria nervurilor (*veins theory*, VT) [Cristea *et al*., 1998], preluând de la RST diferentierea data de nuclearitate între argumentele relatiilor retorice dar ignorând, ca si în AST, numele acestora, releva o structura "ascunsa" în arborele de discurs, numita **nervura**. Fara a nega structura lingvistica a segmentelor de discurs, cât si pe cea intentionala a relatiilor dintre scopurile comunicate de segmente si care, prin echivalarea de care am amintit ([Moser, Moore, 1996, Marcu, 1999]), poate fi recuperata din structura de arbore proprie analizelor RST, VT corecteaza defectul de accesibilitate al AST înlocuind modelul accesibilitatii în stiva cu accesibilitatea de-a lungul nervurilor arborelui de discurs si explicând naturaletea unor referinte la distanta 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 globala a coeziunii discursului. VT generalizeaza totodata partea din CT relativa la încarcarea inferentiala (regula a doua), extinzând concluziile ei la întregul discurs, prin aceasta VT constituindu-se si într-o teorie globala a coerentei.

În sectiunea urmatoare sunt prezentate argumente lingvistice în favoarea teoriei. Sectiunea 3 prezinta definitiile teoriei, sectiunea 4 enunta conjectura VT relativa la

referentialitate, iar sectiunea 5 – conjectura VT referitoare la coerenta. Sectiunea 6 descrie rezultate experimentale în sprijinul presupozitiilor VT, sectiunea 7 prezinta o proprietate de granularitate, iar ultima sectiune este dedicata concluziilor si prezentarii unor aplicatii ale VT.

2. Intuitiile VT

Notiunea de nervura s-a nascut sintetizând observatiile asupra modului în care se aliniaza referintele pe o reprezentare arborescenta a discursului. Considerând organizarea ierarhica data de structura de arbore si principiul compozitionalitatii, care permite ca unitati de discurs aflate la distanta sa fie frati sub aceeasi relatie, aceste observatii au fost urmatoarele (pentru simplificarea exprimarii vom spune ca "o unitate A refera o unitate B" si vom întelege "o expresie referentiala apartinând unei unitati A refera o entitate de discurs introdusa de (sau referita dintr-o) unitate B"; de asemenea vom nota cu u_1, u_2, u_3 – unitati de discurs iar cu R, R_1, R_2 – relatii. Atunci când apar ca argumente ale unei relatii, unitatile de discurs vor purta un indice ridicat ^{n} sau ^{s}, cu semnificatia de nucleu sau satelit):

- un satelit sau un nucleu poate referi un frate nuclear aflat la stânga: în combinatii $u_1^n\ R\ u_2^s$, sau $u_1^n\ R\ u_2^n$, u_2 poate referi u_1;

Ex. 1

1. Ion a plecat de acasa fara umbrela

2. desi dimineata r aflase la radio ca va ploua.

Subiectul vid (notat *r)* din unitatea 2, un satelit al unitatii 1, refera entitatea [**Ion**][90] introdusa de expresia referentiala *Ion* din prima unitate.

- un nucleu poate referi un satelit al sau aflat la stânga: în combinatii $u_1^s\ R\ u_2^n$, u_2 poate referi u_1. Astfel, în exemplul:

Ex. 2

1. Ion i-a dat Mariei o floare.

2. Pentru ca r s-a simtit frustrata,

3. sotia lui s-a suparat.

unitatea 2 este un satelit al unitatii 3. Pe cine desemneaza pronumele vid (notat *r*) din 2, pe [**Maria**] sau pe [**sotia lui Ion**]? Într-o interpretare incrementala a textului, la sfârsitul receptionarii celei de a doua unitati avem tendinta de a asocia, prea timpuriu, subiectul vid [**Mariei**]. Dupa citirea unitatii 3 are loc însa o reconsiderare a legarii $r \rightarrow$ [**Maria**] si o

[90] *Vom nota prin [**text**] entitatea de discurs introdusa/referita de expresia referentiala* text.

identificare a expresiei referentiale *sotia lui* cu subiectul vid din 2, ambele indicând entitatea [**sotia lui Ion**].

- un satelit drept al unui nucleu *u* nu e accesibil dintr-un alt frate drept, nuclear sau satelit, al lui *u*: în combinatii $(u_1^{\,n}\ R_1\ u_2^{\,s})^{n}\ R_2\ u_3^{\,n}$ sau $(u_1^{\,n}\ R_1\ u_2^{\,s})^{n}\ R_2\ u_3^{\,s}$, u_3 poate referi u_1 dar nu u_2.

Ex. 3

1. Ion i-a marturisit Mariei ca o iubeste.

2. El n-a fost niciodata casatorit

3. si a trait pâna la 40 de ani lânga mama sa.

4. Ea, dimpotriva, a fost maritata de doua ori.

Secventa 2-3-4 ofera o explicatie la 1. Secventa 2-3 se afla într-o relatie de CONTRAST (o relatie paratactica) fata de 4, iar 3 aduce o completare la 2. Structura este deci urmatoarea: $u_1^{\,n}\ R_1\ ((u_2^{\,n}\ R_2\ u_3^{\,s})^{n}\ R_3\ u_4^{\,n})^{s}$ în care R_3 este relatia CONTRAST. Pentru cei mai multi cititori, *ea* din unitatea 4 trebuie sa fie [**Maria**], iar nu [**mama lui Ion**], desi [**mama lui Ion**] este entitatea cea mai recent referita, din pozitia unitatii 4, cu care pronumele feminin se potriveste în numar si gen. Motivul preferarii Mariei în locul mamei este acela ca cititorul recunoaste unitatea 4 ca fiind într-o relatie de CONTRAST cu unitatea 2 (relatie pusa în evidenta prin *dimpotriva*), ceea ce face ca cele doua unitati sa fie percepute ca fiind adiacente. Apropierea lor nu este însa una liniara, ci ierarhica, pe structura. Unitatea 3 este închisa la referinta din unitatea 4.

- un nucleu blocheaza accesibilitatea dintr-un satelit drept spre un satelit stâng: în combinatii $(u_1^{\,s}\ R_1\ u_2^{\,n})^{n}\ R_2\ u_3^{\,s}$, u_3 poate referi u_2 dar nu u_1.

Ex. 4

1. Înca înainte cu un an de terminarea mandatului sau de presedinte al firmei

2. dl. W. Ross începuse masinatiile pentru falimentarea acesteia.

**3. De altfel, circulau vorbe ca l-ar fi obtinut fraudulos.*

În acest exemplu 1 si 3 sunt sateliti ai lui 2 (1 este o circumstantiala a lui 2, în timp ce unitatea 3 da o explicatie la purtarea necinstita a lui Ross). Referinta *l*=[**mandatul de presedinte al firmei al lui Ross**] se deduce cu dificultate, ceea ce face ca întregul discurs sa fie defectuos. Dimpotriva, în urmatoarea varianta, discursul câstiga în cursivitate:

Ex. 5

1. Dl. W. Ross începuse masinatiile pentru falimentarea firmei al carei presedinte era

2. înca înainte cu un an de terminarea mandatului sau.

3. De altfel, circulau vorbe ca l-ar fi obtinut fraudulos.

În Ex. 5 unitatea 2 este un satelit al lui 1, iar 3 – un satelit al lui 2 (aici *de altfel* anunta o paranteza la informatia asupra mandatului de presedinte). Referinta *l*=[**mandatul de presedinte al firmei al lui Ross**] poate fi recuperata acum fara dificultate.

Motivatia acceptarii Ex. 5 si rejectarii Ex. 4, consta nu în departarea liniara mai mare a anaforului de antecedent în Ex. 4 decât în Ex. 5, ci în faptul ca în Ex. 4, spre deosebire de Ex. 5, accesul anafor-antecedent se face dinspre un satelit catre un alt satelit, între ei interpunându-se un nucleu. Sa remarcam ca Ex. 4 este reparat daca se elimina aceasta referinta:

Ex. 6

1. Înca înainte cu un an de terminarea mandatului sau de presedinte al firmei

2. dl. W. Ross începuse masinatiile pentru falimentarea acesteia.

3. De altfel, circulau vorbe ca el ar fi fraudulat alegerile.

3. Definitiile teoriei

Intuitia fundamentala care sta la baza dezvoltarilor unificatoare asupra structurii de discurs si accesibilitatii în VT este ca distinctia specifica RST dintre nuclee si sateliti constrânge plaja de referenti asupra carora pot fi rezolvati anaforii; cu alte cuvinte, distinctia nucleu-satelit, corelata cu o structura de discurs, induce pentru fiecare unitate de discurs un domeniu de accesibilitate referentiala imediata pentru anaforii pe care-i contine. Mai precis, pentru fiecare anafor *x* apartinând unei unitati de discurs *u*, VT avanseaza ipoteza ca *x* poate fi rezolvata cu usurinta examinând doar un subset al multimii entitatilor de discurs care preced *u*. Daca antecedentul lui *x* este plasat într-o unitate de discurs aflata în afara domeniului lui *u* atunci legatura anafor-antecedet este refacuta cu greutate, sau pentru realizarea ei e nevoie de mijloace referentiale tari, cum sunt, de exemplu, numele proprii.

Mai mult decât atât, aceeasi corelatie nuclearitate-structura, aplicata întregului discurs, permite generalizarea CT dincolo de granitele unui segment, ceea ce face posibila aplicarea concluziilor CT asupra coerentei la întregul discurs.

VT se bazeaza, în mare masura, pe aceleasi elemente ale structurii de discurs ca si RST:

– unitatile de baza ale discursului sunt întinderi de text (în engleza – *text span*) ce nu se intersecteaza. Dupa cum am precizat mai sus, noi le vom asimila cu propozitii, la nivel semantic fiecare continând o predicatie (careia îi corespunde o reprezentare evenimentiala sau situationala);

- structura unui discurs este reprezentata ca un arbore. Spre deosebire de RST, dar fara a reduce generalitatea, în VT vom considera arborii de discurs ca fiind binari (fiecare nod are exact doi descendenti) (pentru argumentatie, v. [Marcu, 2000] si [Cristea, Webber, 1997]);
- principiul secventialitatii [Cristea, Webber, 1997]: secventa de noduri de pe frontiera terminala a arborelui corespunde secventei de unitati de discurs ce compune textul[91];
- principiul compozitionalitatii [Marcu, 2000]: o relatie ce se aplica între doua întinderi de text se aplica, de asemenea, si între subîntinderile nucleare ale întinderilor aflate în relatie;
- la fel ca în RST, nuclearitatea nodurilor este importanta, nodurile fiind clasificate în nuclee (cele mai importante) si sateliti (cele mai putin importante);
- nodurile terminale ale arborelui reprezinta unitatile de discurs, în timp ce nodurile neterminale reprezinta relatii retorice între întinderi adiacente de text. Spre deosebire de RST, în VT nu intereseaza numele relatiilor, ceea ce conteaza fiind topologia arborelui, nuclearitatea nodurilor si etichetarea nodurilor terminale;
- între fiii fiecarui nod intermediar al arborelui exista cel putin un nod nuclear. Nodul radacina, prin conventie, e considerat satelit.

În vizualizarea arborilor vom reprezenta nodurile neterminale prin dreptunghiuri fara nume, pe cele terminale – prin ovaluri etichetate, iar nodurile nucleare vor fi subliniate (v. Figura 1). În definitiile ce urmeaza vom folosi urmatoarele conventii de notare:

- *mark*(α) este o functie care întoarce sirul α în care fiecare simbol este marcat (de exemplu, este pozitionat între paranteze);
- *unmark*(α) este functia inversa lui *mark*(), ce îndeparteaza toate marcajele atasate simbolurilor din expresia α (ex. *unmark*(*mark*(α)) = α);
- *simpl*(α) este functia care elimina toate simbolurile marcate din expresia argumentului α (ex. *simpl*(*mark*(α)) = *r*, sirul vid si *simpl*($\alpha \cdot$ *mark*(β) $\cdot$ γ)) = $\alpha \cdot \gamma$);
- *seq*(α, β) este o functie de secventiere, care întoarce acea permutare a concatenarii simbolurilor din α si β data de citirea de la stânga la dreapta a nodurilor corespunzatoare simbolurilor din α si β pe frontiera terminala a

[91] Unitățile de discurs întrerupte nuanțează acest principiu. Astfel într-un discurs precum următorul: *O dată,*[1]*/ când treceau unul pe lângă altul pe coridor,*[2]*/ ea îi aruncase o privire piezişă*[1]*/ care parcă-l străpunsese*[3]*/ şi pentru o clipă fusese cuprins de o groază oarbă.*[4]*/* (G. Orwell, *1984*), unitatea 1 este întreruptă de unitatea 2.

arborelui. Functia mentine marcajele asupra simbolurilor, daca acestea exista, $seq(r, \beta) = \beta$; si $seq(\alpha, seq(\beta)) = seq(seq(\alpha), \beta) = seq(\alpha, \beta)$;

- $H(n)$ si $V(n)$ reprezinta expresiile *head* si nervura (în engleza – *vein*) ale unui nod n;
- $pref(u, \alpha)$ retine prefixul expresiei simbolice α pâna la simbolul u inclusiv, o eticheta de nod terminal.

Teoria nervurilor calculeaza doua expresii pe care le ataseaza fiecarui nod al structurii.

3.1 Expresia *head* a unui nod al arborelui

Intentia expresiei *head* a unui nod al arborelui de discurs este de a pune în evidenta secventa celor mai importante unitati de discurs din întinderea de text acoperita de nod. Ea este o secventa de etichete de unitati, dupa cum urmeaza:

Definitii

1. Expresia *head* a unui nod terminal este eticheta sa;
2. Expresia *head* a unui nod neterminal este data de concatenarea, în ordinea aparitiei lor în arbore de la stânga la dreapta, a expresiilor *head* ale descendentilor sai nucleari.

Definitiile expresiilor *head* sugereaza un proces de calcul care se propaga de jos în sus în arborele de discurs. Cele mai importante unitati de discurs sunt proiectate în sus pâna în primul nod satelit întâlnit.

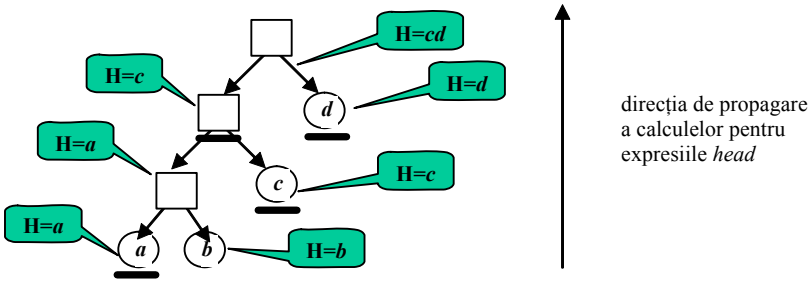


Figura 1: Calculul expresiilor *head*

3.2 Expresia nervurii unui nod al arborelui

Expresia **nervurii** unui nod intentioneaza sa surprinda **secventa unitatilor de discurs care sunt semnificative pentru a sintetiza**[92], **în contextul întregului text, întinderea de text** (în engleza – *text span*) **acoperita de nod**. Pentru orice nod al structurii, expresia nervurii este formata din cele mai importante unitati din întinderea acoperita de nod, împreuna, eventual, cu alte unitati din afara acestei întinderi.

Definitiile care urmeaza, datorita recursivitatii lor, vor face posibila considerarea contextului dat de totalitatea textului din exprimarea "a întelege, în contextul întregului text, întinderea *s*" marginit la întinderea de text acoperita de nodul parinte al celui corespunzator întinderii *s*. Cu alte cuvinte, la fiecare nivel al structurii, cu exceptia radacinii, adica acolo unde exista doua noduri fii sub un nod parinte, cu întinderile celor doua noduri fii însumând întinderea nodului parinte, expresia nervura a parintelui contine deja informatia care permite întelegerea/ rezumarea întinderii acoperite de el în contextul global. Coborârea pentru întelegerea/rezumarea subîntinderii acoperite de nodul curent al definitiei (unul dintre cele doua noduri fii) înseamna adaugarea si/sau stergerea unei secvente noi/subsecvente la/din secventa de etichete contribuita de nervura parintelui, în functie de polaritatea si pozitia specifica a întinderii corespunzatoare nodului fiu curent în întinderea nodului parinte. În continuare, întinderea întregului text, o constanta pentru orice subîntindere, va fi numita **contextul total**. În figurile 2-6, nodurile curente – cele vizate de definitiile curente de nervura – apar în gri. Ele sunt notate simultan cu un dreptunghi si un oval pentru a sugera ca pot fi atât noduri interioare (neterminale), cât si noduri terminale.

Definitii

1. Expresia nervurii radacinii este egala cu expresia sa *head*.

Expresia nervurii nodului radacina, conform intentiei generale a nervurii unui nod, ar trebui sa fie formata din cele mai semnificative unitati de discurs necesare întelegerii/rezumarii întinderii acoperite de nod (în cazul de fata – întregul text) în contextul total. Cum contextul este aici egal cu textul în totalitatea lui, el poate fi lasat la o parte în descriere, ceea ce ne lasa cu definitia expresiei *head* a nodului radacina.

2. Pentru fiecare nod nuclear, al carui parinte are nervura *v*:

---

[92] *Prin sinteza, sau rezumatul, unei întinderi de text se întelege un text mai scurt care reda ideea principala a textului supus sintezei. Indiferent daca este realizat prin parafrazare sau prin punerea cap la cap a unor subsecvente ale întinderii originale [Mani, 2001], orice rezumat trebuie sa fie comprehensibil, adica trebuie sa poata fi înteles prin el însusi (printre altele, de exemplu, rezumatul trebuie sa contina toate elementele care sa permita rezolvarea anaforilor). Adesea însa, atunci când întinderea este decupata dintr-un context mai larg, pentru ca rezumatul sa fie comprehensibil, el trebuie sa contina si elemente din afara întinderii si care apartin contextului. Avem de a face, în acest caz, cu o sinteza a unei întinderi de text* ***în contextul*** *unei întinderi mai vaste. Sa mai observam ca, în multe privinte, "a sintetiza" e analog cu "a întelege", pentru ca ceea ce ne ramâne dupa lectura unui text este o sinteza a lui.*

a. daca nodul nu are un frate nenuclear în stânga, atunci expresia nervurii este *v* (v. Figura 2);

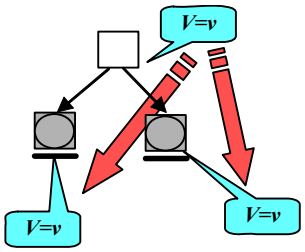


Figura 2: Expresia nervurii unui nod nuclear fara frate satelit în stânga

Definitia exprima faptul ca secventa de unitati necesara întelegerii/ rezumarii, în contextul total, a unei întinderi nucleare de text ce are ca frate în structura o alta întindere nucleara necesita aceeasi secventa de unitati ca si cea necesara întelegerii/rezumarii, în contextul total, a reuniunii celor doua întinderi. Cu alte cuvinte, o întindere nucleara ce este frate, în structura, întinderii nucleare curente este esentiala întelegerii/rezumarii întinderii curente.

b. daca nodul are un frate nenuclear în stânga de *head h*, atunci expresia nervurii lui este *seq*(*mark*(*h*), *v*) (v. Figura 3);

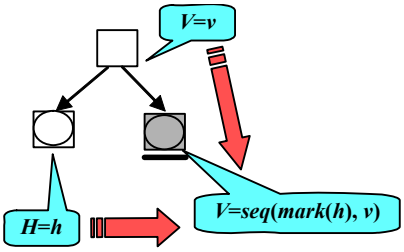


Figura 3: Expresia nervurii unui nod nuclear având
un frate satelit în stânga

Secventa de unitati necesara întelegerii/rezumarii, în contextul total, a unei întinderi nucleare de text ce are ca frate stâng în structura o întindere nenucleara necesita, suplimentar fata de secventa necesara întelegerii în contextul total a întinderii acoperita de

nodul parinte (comunicata de expresia nervura a nodului parinte) si secventa *head* a întinderii frate stângi (adica cele mai importante unitati din întinderea stânga). Considerarea, în expresia nervurii întinderii nucleare curente, a expresiei *head* a întinderii nenucleare frate stângi, corespunde, prin prisma definitiei 2a, cu atribuirea întinderii stângi a calitatii de a se comporta ca un nucleu. Marcarea contributiei satelitului frate stâng prin functia *mark*() face însa aceasta revizuire a nuclearitatii lui, una cu valoare temporara, dupa cum se va dovedi mai jos, în definitia 3b.

3. Pentru fiecare nod nenuclear de *head h*, al carui parinte are nervura *v*:

a. daca nodul este descendentul stâng al parintelui sau, atunci expresia nervurii este *seq*(*h, v*);

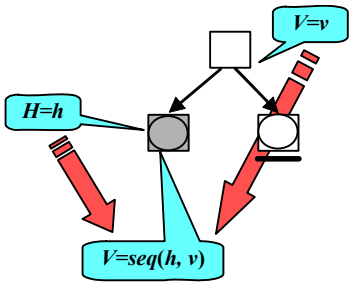


Figura 4: Expresia nervurii unui nod satelit stâng

Definitia exprima faptul ca pentru a întelege/rezuma, în contextul total, o întindere nenucleara de text ce este descendent stâng, în structura, nodului parinte, la secventa de unitati necesara întelegerii/rezumarii contextului total (contribuita de expresia nervura a parintelui) trebuie adaugate cele mai importante unitati din întinderea proprie (contribuite de expresia *head* proprie). Sa observam ca în expresia nervurii nodului parinte, care mosteneste expresii *head* ale nodurilor superioare, nu poate razbate influenta unui fiu satelit al sau, deci numai includerea *head*-ului fiului satelit, direct în expresia nervurii sale poate completa aceasta influenta.

b. daca nodul este descendentul drept al parintelui sau, atunci expresia nervurii lui este *seq*(*h, simpl*(*v*)).

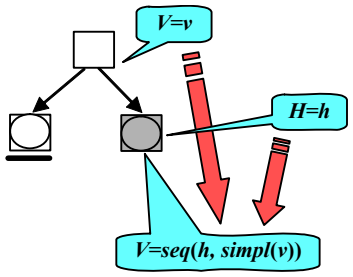


Figura 5: Expresia nervurii unui nod satelit drept

Pentru a întelege, în contextul total, o întindere nenucleara de text ce este descendent pe dreapta al nodului parinte, la secventa de unitati necesara întelegerii/rezumarii contextului total (contribuita de expresia nervura a parintelui) si din care s-au sters unitatile marcate trebuie adaugate cele mai importante unitati din întinderea proprie (contribuite de expresia *head* proprie). În acest fel, daca expresia nervura a nodului parinte nu contine unitati marcate (prin contributia definitiei 2b), atunci expresia nervura a unui satelit drept nu difera de expresia nervura a aceluiasi satelit ce ar fi fost pozitionat pe stânga (conform definitiei 3a). Daca însa nervura parintelui contine unitati marcate, atunci acestea dispar din expresia nervurii satelitului drept. Cum, conform definitiei 2b, unitatile marcate pot fi contribuite doar de un satelit stâng, frate al celui mai apropiat ascendent nuclear al întinderii curente, urmeaza ca definitia curenta exprima o proprietate de blocare a accesibilitatii dinspre un satelit plasat în dreapta unui nucleu catre un satelit plasat în stânga sa (v. Figura 6).

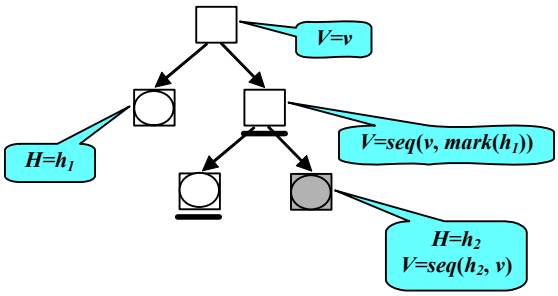


Figura 6: Simplificari în calculul expresiei nervura a unui satelit drept:

$V=seq(h_2, simpl(seq(v, mark(h_1))) = seq(h_2, seq(v)) = seq(h_2, v)$

Daca semnificatia expresiei nervurii unui nod oarecare din structura este particularizata la un nod terminal, obtinem: **expresia nervurii unei unitati de discurs reprezinta secventa unitatilor de discurs care sunt semnificative pentru a întelege/rezuma, în contextul întregului text, însasi unitatea de discurs în cauza**. Printre altele, aceasta înseamna ca expresia nervurii unei unitati de discurs este suficienta pentru a interpreta toate referintele anaforice continute în unitate.

4. Relatia dintre structura de discurs si referentialitate

Ipoteza pe care o avansam este ca exista doua tipuri de procese de rezolutie anaforica: **evocative** (sau **imediate**) si **post-evocative** (sau **inferentiale**). Procesele evocative, cele mai frecvente, sunt rapide si pot fi realizate prin orice mijloace de evocare referentiala, inclusiv cele fragile (de tipul subiectelor vide si pronumelor). Ele dau textului fluenta si-l fac coeziv. Cele post-evocative sunt mult mai putin frecvente decât cele evocative, necesita o încarcare inferentiala mai mare pentru procesarea lor si utilizeaza mijloace referentiale tari (nume proprii, substantive comune articulate).

Vom asocia spatiul de cautare al proceselor evocative unui **domeniu de accesibilitate referentiala evocativa** sau **imediata** (*domain of evocative accessibility – dea*) pe baza definitiei nervurii si al urmatoarelor observatii:

- **natura semantica a relatiei anaforice** [Halliday, Hassan, 1976]: o relatie anaforica are doi termeni: anaforul si antecedentul. Anaforul este reprezentat de o expresie referentiala a carei natura este textuala. Natura semantica a relatiei anaforice trebuie înteleasa ca rasfrângându-se asupra antecedentului care nu trebuie identificat cu o anumita expresie referentiala ce precede în text anaforul ci cu o reprezentare a acesteia într-un plan semantic în asa fel încât semnificatia anaforului se construieste din cea a antecedentului însusi iar nu a semnificatiei lui. În cazul particular al unui lant co-referential acest lucru înseamna ca antecedentul este "realizat" repetat în text în aceeasi entitate de discurs. Expresiile co-referentiale "ancoreaza" în diverse pozitii ale textului entitatea de discurs.
- **dinamica incrementala a interpretarii discursului**: un discurs este un text în procesul citirii ori ascultarii lui de catre un subiect (om sau masina). Când citirea/ascultarea unui text s-a terminat discursul este încheiat si ceea ce ramâne este o reprezentare a lui în memoria subiectului. De asemenea, la un moment dat pe parcursul interpretarii unui text, anumite elemente ale discursului pot fi plasate privilegiat în sfera atentiei [Grosz, Sidner, 1986, Sidner, 1983, Walker, 1996], iar trecerea de la o unitate de discurs la urmatoarea poate produce schimbari în structura memorata ce configureaza sfera atentionala.

– **natura cognitiva comuna a anaforei si a cataforei**: dintr-un punct de vedere cognitiv, toate referintele anaforice se fac dinspre expresii referentiale (entitati textuale) catre entitati ale discursului (entitati semantice) deja introduse de discursul trecut. Acest lucru înseamna ca într-o limba în care textul se noteaza de la stânga spre dreapta nu exista referinte anaforice spre dreapta. Distinctia dintre anafora si catafora, devine, în aceasta viziune care încearca sa reconstituie procesele cognitive ce stau la baza întelegerii textelor (cu sau fara scopul simularii lor pe masina), inutila. În aceeasi maniera în care, într-o anafora, un antecedent este o entitate de discurs propusa de o expresie referentiala ce precede anaforul si pe care anaforul o refera apoi, pronumele ce precede un nume într-o catafora propune o reprezentare, mai saraca, pe care numele o refera si o completeaza în acelasi timp [Cristea, Dima, 2001]. Acest lucru atribuie interpretarii discursului o unica directionalitate, care corespunde axei timpului lecturii, si care este cea a desfasurarii liniare a textului (pentru limbile europene, de exemplu, de la stânga la dreapta). Relatia de referentialitate trebuie deci sa se proiecteze pe aceasta axa, dinspre entitati "noi" catre entitati "vechi", mereu catre înapoi pe axa timpului lecturii.

Ex. 7

1. Pentru ca φn-a vrut sa-si lase tata singur,

2. Ion a renuntat la concediu.

Expresia referentiala vida de pe pozitia de subiect a unitatii de discurs 1 propune o entitate de discurs caracterizata cel mult de o descriere [**type** human] (contribuita, cel mai probabil, de surse de cunoastere de natura pragmatica: cineva care nu poate sa-si lase tatal singur trebuie sa fie o persoana). Apoi, substantivul propriu *Ion*, din unitatea 2, refera entitatea construita precedent si o completeaza pâna la o reprezentare: [**type** human**, name** Ion].

Corelarea definitiei nervurii cu observatiile de mai sus, conduce la definirea domeniului de accesibilitate referentiala evocativa ca fiind format din toate unitatile de discurs care preced unitatea în care se gaseste expresia referentiala (si din care au fost îndepartate eventualele marcaje, ce îndeplineau un rol de memorie temporara):

$$dea(u) = pref(u, unmark(V(u))).$$

Definitia *dea* formalizeaza prima conjectura a VT (sau a coeziunii), care pune în legatura accesibilitatea referentiala imediata de structura de discurs: antecedentii expresiilor referentiale dintr-o unitate de discurs *u* se gasesc, cu precadere, printre entitatile de discurs ancorate în unitatile ce preced pe *u*, inclusiv *u*, în expresia nervurii acesteia.

Paul Cornea [1998] vorbeste despre recodificarea sensului si memorizarea. El pune în evidenta trei tipuri de memorie, ce apar, de altfel, la mai multi cercetatori

[Kinntsch, Van Dijk, 1975, Schank, Abelson, 1977, Walker, 1996]: memoria imediata, memoria de scurta durata (de termen scurt – MST) si cea de lunga durata (de termen lung – MLT). Memoria imediata este un sistem de stocaj senzorial al informatiilor, retinerea urmelor din ultima jumatate de secunda. MST conserva câteva secunde informatia. Lungimea acestei memorii pare a fi de 7±2 semne (cuvinte, cifre, litere – functie de context, v. si [Miller, 1956]; alti cercetatori apreciaza acest "empan" mijlociu la 13-15 cuvinte, la un lector lent fiind de 8 cuvinte, la unul rapid – de 16-20, de ex. [Richadeau, 1969] – citat în [Cornea, 1998] p. 166).

Constructia structurii de discurs se face dinamic, în actul lecturii. Sa ignoram un posibil proces de multi-interpretare ce poate duce la sintetizarea simultana a mai multor constructii alternative din care sa se selecteze, în urma unui proces de dezambiguizare, una sau mai multe structuri arborescente finale. Arborele însusi poate fi considerat rezumat în diverse grade, conform capacitatii de memorare a subiectului. Daca unitatea curenta este u_n, sa notam AR_n arborele de structura rezumat, la momentul prelucrarii unitatii u_n. Nervura acesteia, culeasa pe AR_n, este $V(u_n)$, iar domeniul ei de accesibilitate imediata $dea(u_n)$. Noi credem ca MST poate fi considerata o fereastra de lungime 7±2 semne în directa legatura cu $dea(u_n)$: fie 7±2 unitati din aceasta secventa, fie tot atâtea structuri evenimentiale – ca reprezentari ale unitatilor de discurs, fie înca numai simboluri (cuvinte etc.) culese din acest sir de unitati. Tranzitarea la urmatoarea unitate, u_{n+1}, înseamna înlocuirea memoriei de scurta durata $dea(u_n)$ cu $dea(u_{n+1})$. Acest lucru duce uneori la o simpla prelungire a domeniului de accesibilitate precedent, alteori la o alterare a lui prin stergerea unor unitati si adaugarea altora, de fiecare data domeniul încheindu-se cu unitatea curenta. MST este asadar o proiectie a unui sir de unitati de discurs (sau de microstructuri suportate de unitati) decupate din structura dinamica curenta. Modificarile ce apar în sirul MST reflecta schimbarile de focalizare, în parcurgerea discursului. Componenta acestui sir este influentata de uitare (deci de un proces de abstractizare) si de modificarea de interes curenta în parcurgerea discursului. Când interesul s-a mutat pe o alta axa, componenta nervurii si, de aici, a domeniului de accesibilitate imediata sunt si ele actualizate. Includerea sau excluderea din MST a unor unitati de discurs în ritmul citirii, pentru ca *dea* evolueaza eliminând unele unitati si "redesteptând" altele "uitate", amintesc de procesele de "chemare" în sfera atentiei ale memoriei *cash* a lui Walker [Walker, 1996]. Pe de alta parte, structura memorata (rezumata) a discursului este pastrata în MLT si folosita pentru aducerea în prim plan a unitatilor de interes curent ce au fost temporar retrogradate de o comutare a atentiei într-o alta directie. Procesele evocative se desfasoara asadar în memoria de scurta durata. Pe de alta parte, procesele post-evocative sunt procese de rezolutie anaforica de natura inferentiala, ce presupun un anumit efort de regasire a unei entitati de discurs într-o zona a memoriei de lunga durata sau evoca entitati ale cunoasterii generice din sfera culturala a subiectului. Noi credem ca aceste procese se dezvolta tot pe structura de discurs dezvoltata deja, iesind din *dea*, când rezolutia a esuat acolo.

Dintr-un punct de vedere ce se concentreaza asupra relatiei dintre referintele anaforice si structura de discurs, celor doua tipuri de procese anaforice le corespund **referinte evocative**, respectiv **post-evocative** (sau **inferentiale**). Diferenta dintre ele este

ca primele apar când lantul retroactiv al unitatilor ce ancoreaza expresii aflate în relatii referentiale intersecteaza domeniul de accesibilitate referentiala imediata al unitatii anaforului în cel putin înca un punct decât unitatea anaforului, pe când în cazul referintelor post-evocative nu exista aceasta intersectie dubla. În [Cristea *et al*., 2000, Cristea, 2000] referintele evocative sunt, mai departe, detaliate în **directe** si **indirecte**.

În referintele directe a doua unitate de intersectie este cea mai recenta liniar unitate ce ancoreaza aceeasi entitate de discurs ca si anaforul (în cazul relatiei de co-referinta) sau o entitate corelata functional cu aceasta (în cazul unei relatii de referinta functionala). În referintele indirecte intersectia *dea* cu lantul co/func-referential se realizeaza într-o unitate mai departata decât cea mai recenta liniar de unitatea anaforului. În referintele inferentiale lantul retroactiv al legaturilor anaforice al anaforului nu intersecteaza *dea* (în Figura 7 lantul legaturilor anaforice este reprezentat punctat, iar *dea* printr-o linie groasa).

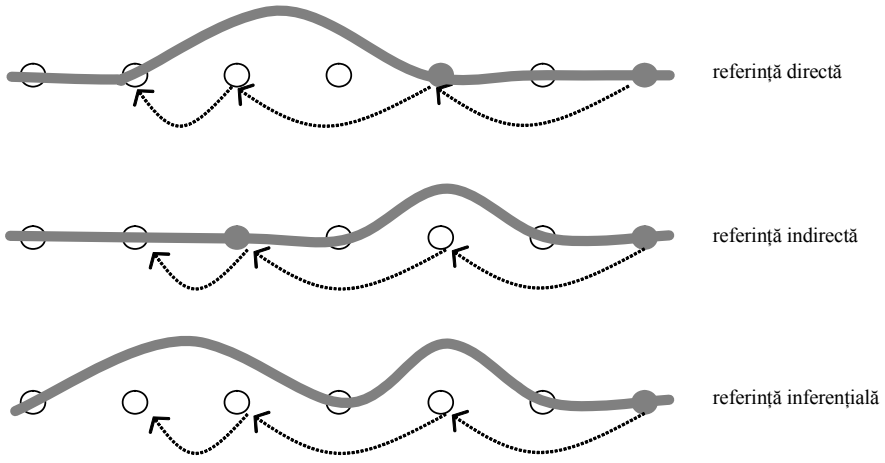


Figura 7: Referinte evocative si post-evocative

O categorie particulara de referinte post-evocative sunt **referintele pragmatice** (ce pot fi numite si **pseudo-referinte**). În acest tip de referinte participa expresii referentiale care pot fi interpretate fara un antecedent pentru ca interpretarea lor se bazeaza pe cunostinte exterioare textului, ce vin din cunoasterea comuna asupra lumii, deci din pragmatica. Desi exista cel putin înca o expresie referentiala în text ce realizeaza aceeasi entitate de discurs, expresiile co-referentiale pot sa nu aiba, în mod necesar, o reprezentare unica, fara ca prin acesta întelegerea textului sa sufere.

Recunoasterea antecedentului se datoreaza, în toate cazurile, unor procese de *pattern-matching* îmbogatite cu euristici, în care intervin structura de caracteristici morfo/sintactico/semantice ce definesc anaforul si structurile de caracteristici ce definesc entitatile de discurs deja introduse [Cristea, Dima, 2001, Cristea *et al*., 2002a].

5. Relatia dintre structura de discurs si cursivitate

5. 1 Linii de argumentatie

Expresiile nervura ale unitatilor ce compun un discurs arata tot atâtea moduri diferite în care poate fi citit acel discurs. Fiecare în parte da o rezumare a discursului prin prisma unitatii de discurs curente. Atunci când interesul este orientat catre un anumit episod al povestirii, putem sari peste pasaje întregi pentru a ne concentra asupra manierei în care elementul de interes se leaga cu ansamblul discursului. În acelasi fel, putem avea în vedere o alta pista si atunci lectura focalizeaza un alt fir de interes. Acest nou fir poate sa aiba elemente în comun cu primul dar poate, de asemenea, sa incorporeze si altele noi. Fiecare fir în parte poate pune în evidenta anumite particularitati, legate însa strâns de linia principala a discursului. Toate aceste sub-discursuri sunt coerente si nu exista referinte anaforice pentru a caror interpretare sa avem nevoie de fragmente aflate în afara rezumatului însusi. Acest lucru înseamna ca traseele referentiale ale rezumatului contin suficiente elemente care sa duca la recuperarea întelesului anaforilor.

Sa luam urmatorul text:

Ex. 8

1. *Piton primise-n taina porunca de la Hera sa-l pîndeasca pe Apolo,*
2. *cînd va trece prin munte,*
3. *si sa-i rapuna viata.*
4. *Hera-l ura pe fiul cel nou nascut al Letei,*
5. *pentru ca sotul sau, prea puternicul Zeus, tinea mai mult la dînsul decit la fiii ei: Hefaistos si Ares.*
6. *Cînd a ajuns Apolo în muntele Parnas,*
7. *dihania uriasa s-a avîntat spre dînsul,*
8. *dornica sa-l ucida.*
9. *Dar zeul si-a întins arcul.*
10. *A tras prima sageata.*
11. *Erau doar patru zile de cînd vazuse lumea,*
12. *si întiia lui sageata a si nimerit monstrul.*

Alexandru Mitru - *Legendele Olimpului*, Editura Tineretului, 1966

Structura de discurs a acestui text este cea din Figura 8. Tabela 1 da expresiile nervura si domeniile de referentialitate evocativa ale nodurilor terminale. În coloana *dea*(*u*) au fost, totodata, marcate în aldine domenii de referentialitate imediata maximale vis-r-vis de relatia de incluziune (cele mai lungi trasee *dea*). Astfel *dea*(1) $\subseteq$ *dea*(2) $\not\subset$ *dea*(3) $\subseteq$ *dea*(4) $\subseteq$ *dea*(5) $\not\subset$ *dea*(6) s.a.m.d. Vom numi aceste secvente care întrerup lanturi de incluziuni **linii de argumentatie** (*la*), în cazul nostru: 1 2, 1 3 4 5, 1 3 6 7, 1 3 7 8 si 1 3 7 9 10 11 12. Daca *la*(u_1) precede imediat *la*(u_2), atunci în *la*(u_2) se regasesc domeniile tuturor unitatilor dintre u_1+1 si u_2. În particular, în *la*(u_2) se regasesc unitatile ce preced imediat unitatea *u*, pentru orice *u* între u_1+1 si u_2, în domeniul lor de accesibilitate imediata (adica acel domeniu care confera discursului maximul de coerenta). Cu alte cuvinte, pe *la*(u_2) putem aplica definitiile CT de calculare a tranzitiilor pentru orice *u* între u_1+1 si u_2.

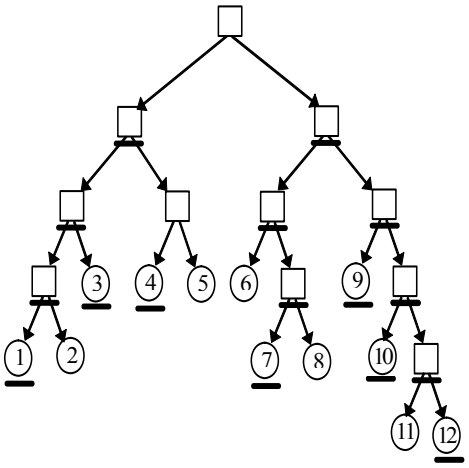


Tabela 1: Nervurile si domeniile unitatilor din Ex. 8

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 3 4
5	1 3 4 5 7 9 10 12	**1 3 4 5**
6	1 3 6 7 9 10 12	1 3 6
7	1 3 (6) 7 9 10 12	**1 3 6 7**
8	1 3 7 8 9 10 12	**1 3 7 8**
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 3 7 9 10 11 12**

Figura 8**: Structura de discurs a Ex. 8**

5.2. O generalizare a CT

Urmând recomandarile teoriei centrelor, sa presupunem ca marcam tranzitiile ce apar între unitati de discurs cu scoruri care sa dea un grad al usurintei de prelucrare:

CONTINUARE	(CON)	4
RETINERE	(RET)	3
SCHIMBARE LINA	(SSH)	2
SCHIMBARE ABRUPTA	(ASH)	1
LIPSA Cb	(-)	0

În felul acesta, tranzitiile 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. Sa notam un scor în spiritul CT al unui segment s cu S_{CCT}^{s} (CCT de la *Classical Centering Theory*). El ne va da o masura a usurintei de interpretare a segmentului: cu cât un segment *s*, în totalitatea lui, e mai fluent, cu atât scorul lui va fi mai mare si cu cât el este mai abrupt, mai dificil de prelucrat, cu atât scorul lui va fi mai scazut. În fine, sa adunam aceste scoruri pentru toate segmentele discursului, într-un scor al sumei segmentelor S_{CCT}:

$$S_{CCT}= \sum_{s} S_{CCT}^{s}$$

Sa ne imaginam acuma ca fortam nota si calculam aceste scoruri si dincolo de granitele de segment, deci inclusiv în punctele de frontiera dintre segmente. Sa notam acest scor global cu S_{CCT}^{G}. În scorul global S_{CCT}^{G} contribuie cu scoruri de tranzitii toate unitatile cuprinse între a doua unitate si ultima. În mod normal tranzitiile în punctele de trecere între segmente ar trebui sa fie foarte abrupte, cotate deci slab ori zero, si deci scorul global atasat textului n-ar trebui sa fie modificat semnificativ. Daca apare totusi o diferenta, ea trebuie sa fie datorata unor tranzitii accidentale peste granita de segment. În orice caz trebuie sa avem $S_{CCT}^{G} \geq S_{CCT}$.

Sa procedam acum în mod analog, ca suport folosind de data aceasta liniile de argumentatie iar nu secventele liniare de unitati ale segmentelor în sensul clasic. Datorita comportamentului lor similar segmentelor, putem numi liniile de argumentatie **segmente în sens ierarhic**. Sa notam S_{HCT}^{s} (HCT de la *Hierarchical Centering Theory*) suma scorurilor unitatilor apartinând unei linii de argumentatie (segment ierarhic) *s*. Ca sa dam o masura a fluentei discursului în acceptiunea ierarhica, similara scorului global S_{CCT}^{G}, în calculul scorului global al discursului în sens ierarhic nu va trebui sa repetam contributiile unitatilor ce apar în mai mult decât o singura linie de argumentatie. Daca notam $S_{HCT}^{s'}$ scorul unui segment ierarhic *s'* în care am pastrat numai unitatile noi fata de segmentul anterior, atunci scorul global ierarhic al discursului este:

$$SHCTG= \sum_{s'} SHCTs'$$

Cea de a doua conjectura a VT (a coerentei): Scorul global în sensul ierarhic al unui discurs este mai bun sau cel putin egal decât scorul global în sensul clasic: $S_{HCT}^{G} \geq S_{CCT}^{G}$.

Pentru un anumit detaliu de granularitate în definirea segmentelor în sens clasic, unui segment în sens clasic îi corespunde o secventa de nervura, deci o portiune a unei linii de argumentatie. În spiritul acestei observatii, cea de a doua conjectura enunta prezumtia ca tranzitiile la distanta lunga calculate în lungul nervurilor sunt sistematic mai line decât tranzitiile accidentale la granitele dintre segmente. Sa notam ca aceasta presupozitie este

conforma unor observatii facute de autori precum Passonneau [1995] si Walker [1998], furnizând totodata o explicatie pentru rezultatele lor.

În cele ce urmeaza prezentam o analiza comparativa clasic-ierarhic care probeaza ipoteza coerentei, pe discursul din Ex. 8.

Tabela 2: Analiza Ex. 1 în maniera CCT

n	u_n	$C_f(u_n)$	$C_b(u_n)$	Traz.	Scor
1	*Piton primise-n taină poruncă de la Hera să-l pîndească pe Apolo,*	[Piton], [Hera], [Apolo],	[**Piton**]	-	-
2	*cînd ∅ va trece prin munte,*	*∅* = [Apolo], [munte]	[**Apolo**]	SSH	2
3	*şi ∅ să-i răpună viaţa.*	*∅* = [Piton], *i*=[Apolo], [viaţa]	[**Apolo**]	RET	3
4	*Hera-l ura pe fiul cel nou născut al Letei,*	[Hera], [Leta], *fiul cel nou-născut al Letei*=[Apolo]	-	-	0
5	*pentru că soţul său, prea puternicul Zeus, ţinea mai mult la dînsul decit la fiii ei: Hefaistos şi Ares.*	[Zeus], *său*=[Hera], *dînsul*=[Apolo], [Hefaistos], [Ares]	[**Hera**]	ASH	1
6	*Cînd a ajuns Apolo în muntele Parnas,*	[Apolo], [munte]	[**Apolo**]	SSH	2
7	*dihania uriaşă s-a avîntat spre dînsul,*	*dihania uriaşă*=[Piton], *dînsul*=[Apolo]	[**Apolo**]	RET	3
8	*∅ (era) dornică să-l ucidă.*	*∅* = [Piton], l=[Apolo]	[**Piton**]	SSH	2
9	*Dar zeul şi-a întins arcul.*	*zeul* = [Apolo], [arcul]	[**Apolo**]	SSH	2
10	*∅ A tras prima săgeată.*	*∅* = [Apolo], [săgeata]	[**Apolo**]	CON	4
11	*Erau doar patru zile de cînd ∅ văzuse lumea,*	[4 zile], *∅* = [Apolo], [lumea]	[**Apolo**]	RET	3
12	*şi întiia lui săgeată a şi nimerit monstrul.*	*lui*=[Apolo], [săgeata], *monstrul*=[Piton]	[**Apolo**]	CON	4
				Total	26

În constructia tabelului de mai sus am presupus ca toate referintele anaforice au fost corect rezolvate. Unitatile carora le corespund tranzitiile listate în tabela sunt cele ale caror numere apar în caractere aldine în prima coloana, adica 2-12, în numar total de 11. Scorul total de 23 corespunde unei scor mediu pe tranzitie de 26/11=2,36, ceea ce înseamna ca textul, conform aprecierii CT, se comporta, în medie, intermediar între o schimbare lina (SSH) si o retinere (RET), mai apropiat de o schimbare lina.

Daca luam în calcul liniile de argumentatie indicate de nervuri, pot fi puse în evidenta 5 sub-discursuri, în lungul carora vom calcula, de asemenea, tranzitiile. În tabelele 3÷7 de mai jos unitatile pentru care consideram tranzitiile sunt, de asemenea, indicate în caractere aldine în prima coloana. Sa remarcam ca citirea textelor date de liniile de

argumentatie produce, în toate cazurile, discursuri perfect coerente. În ansamblu, doar câte o tranzitie este calculata pentru fiecare unitate, la fel ca si în interpretarea clasica.

Tabela 3: Analiza HCT a primei linii de argumentatie, secventa de unitati 1-2

n	u_n	$C_f(u_n)$	$C_b(u_n)$	Traz.	Scor
1	*Piton primise-n taină poruncă de la Hera să-l pîndească pe Apolo,*	[**Piton**], [**Hera**], [**Apolo**],	[**Piton**]	-	-
2	*cînd ∅ va trece prin munte,*	*∅*= [**Apolo**], [**munte**]	[**Apolo**]	SSH	2
				Total	**2**

Tabela 4: Analiza HCT a celei de a doua linii de argumentatie, secventa de unitati 1-3-4-5

n	u_n	$C_f(u_n)$	$C_b(u_n)$	Traz.	Scor
1	*Piton primise-n taină poruncă de la Hera să-l pîndească pe Apolo,*	[**Piton**], [**Hera**], [**Apolo**],	[**Piton**]	-	-
3	*şi ∅ să-i răpună viaţa.*	*∅*= [**Piton**], *i*=[**Apolo**], [**viaţa**]	[**Piton**]	CON	4
4	*Hera-l ura pe fiul cel nou născut al Letei,*	[**Hera**], [**Leta**], *fiul cel nou-născut al Letei*=[**Apolo**]	-	-	0
5	*pentru că soţul său, prea puternicul Zeus, ţinea mai mult la dînsul decît la fiii ei: Hefaistos şi Ares.*	[**Zeus**], *său*=[**Hera**], *dînsul*=[**Apolo**], [**Hefaistos**], [**Ares**]	[**Hera**]	ASH	1
				Total	**5**

Se constata ca tranzitia RET a unitatii 3 catre 2 din analiza CCT s-a transformat într-o tranzitie CON, pe nervura, dinspre 3 catre 1.

Tabela 5: Analiza HCT a celei de a treia linii de argumentatie, secventa de unitati 1-3-6-7

n	u_n	$C_f(u_n)$	$C_b(u_n)$	Traz.	Scor
1	*Piton primise-n taină poruncă de la Hera să-l pîndească pe Apolo,*	[**Piton**], [**Hera**], [**Apolo**],	[**Piton**]	-	-
3	*şi ∅ să-i răpună viaţa.*	*∅*= [**Piton**], *i*=[**Apolo**], [**viaţa**]	[**Piton**]	-	-
6	*Cînd a ajuns Apolo în muntele Parnas,*	[**Apolo**], [**munte**]	[**Apolo**]	SSH	2
7	*dihania uriaşă s-a avîntat spre dînsul,*	*dihania uriaşă*=[**Piton**], *dînsul*=[**Apolo**]	[**Apolo**]	RET	3
				Total	**5**

Tabela 6: Analiza HCT a celei de a patra linii de argumentatie, secventa de unitati 1-3-7-8

n	u_n	$C_f(u_n)$	$C_b(u_n)$	Traz.	Scor
1	*Piton primise-n taină poruncă de la Hera să-l pîndească pe Apolo,*	[**Piton**], [**Hera**], [**Apolo**],	[**Piton**]	-	-
3	*şi ∅ să-i răpună viaţa.*	*∅*= [**Piton**], *i*=[**Apolo**], [**viaţa**]	[**Piton**]	-	-
7	*dihania uriaşă s-a avîntat spre dînsul,*	*dihania uriaşă*=[**Piton**], *dînsul*=[**Apolo**]	[**Piton**]	-	-
8	*∅ (era) dornică să-l ucidă.*	*∅*= [**Piton**], *l*=[**Apolo**]	[**Piton**]	CON	4
				Total	**4**

Se constata ca tranzitia SSH a unitatii 8 catre 7 din analiza CCT s-a transformat într-o tranzitie CON, pe nervura, tot între 8 si 7 (C_b-ul unitatii 7 s-a schimbat din [**Apolo**] în [**Piton**], pentru ca, pe nervura lui 8, precedenta unitate a lui 7 este acum 3, iar nu 6 ca în secventa liniara).

Tabela 7: Analiza HCT a ultimei linii de argumentatie, secventa de unitati 1-3-7-9-10-11-12

n	u_n	$C_f(u_n)$	$C_b(u_n)$	Traz.	Scor
1	*Piton primise-n taină poruncă de la Hera să-l pîndească pe Apolo,*	[**Piton**], [**Hera**], [**Apolo**],	[**Piton**]	-	-
3	*şi ∅ să-i răpună viaţa.*	*∅* = [**Piton**], *i*=[**Apolo**], [**viaţa**]	[**Piton**]	-	-
7	*dihania uriaşă s-a avîntat spre dînsul,*	*dihania uriaşă*=[**Piton**], *dînsul*=[**Apolo**]	[**Piton**]	-	-
9	*Dar zeul şi-a întins arcul.*	*zeul* = [**Apolo**], [**arcul**]	[**Apolo**]	SSH	2
10	*∅ A tras prima săgeată.*	*∅* = [**Apolo**], [**săgeata**]	[**Apolo**]	CON	4
11	*Erau doar patru zile de cînd ∅ văzuse lumea,*	[**4 zile**], *∅* = [**Apolo**], [**lumea**]	[**Apolo**]	RET	3
12	*şi întiia lui săgeată a şi nimerit monstrul.*	*lui*=[**Apolo**], [**săgeata**], *monstrul*=[**Piton**]	[**Apolo**]	CON	4
				Total	**13**

Însumând scorurile tranzitiilor pentru toate liniile de argumentatie se obtine scorul total: 29, ceea ce corespunde unei tranzitii medii a discursului, calculata conform HCT de 29/11=2,63, asadar o tranzitie mai apropiata de retinere, mai buna decât scorul mediu calculat conform CCT.

6. Validarea conjecturilor VT

Validarea conjecturilor VT s-a realizat pe corpusuri adnotate la structura si la legaturi co-referentiale. Astfel în [Cristea *et al*., 1998] se raporteaza o investigatie efectuata pe texte în limbile engleza, franceza si româna ce au însumat un total de 176 de unitati de discurs. Plecând de o adnotare în maniera RST a structurii de discurs, un program a calculat expresiile nervurilor unitatilor. Pentru verificarea conjecturii coeziunii, utilizând adnotarea legaturilor referentiale s-a calculat apoi procentajul referintelor directe, indirecte si pragmatice. În medie 99,1% dintre referinte se încadreaza acestor trei categorii (87,1% directe, 8,5% indirecte si 3,5% pragmatice). Pentru verificarea conjecturii coerentei, suplimentar marcajelor de structura si lanturi co-referentiale s-au marcat manual, pentru fiecare unitate, C_b-ul, în varianta clasica si în varianta ierarhica, si s-au calculat tranzitiile în cele doua variante. Scorul S_{HCT} a fost mai bun decât scorul S_{CCT} în toate cazurile (scorurile medii pe tranzitie au fost de 2,03 în varianta ierarhica fata de 1,89 în cea clasica).

În [Cristea *et al*., 2000] se raporteaza experimente care au urmarit sa compare potentialul modelelor ierarhice, precum cele bazate pe VT, de a regasi un antecedent într-o plaja de cautare data fata de modelele lineare (modele ce presupun o parcurgere lineara a textului dinspre unitatea anaforului spre începutul textului). Pentru aceasta s-au utilizat 30 de texte englezesti (însumând aproximativ 1560 de unitati de discurs), adnotate la structura

RST si lanturi co-referentiale. Presupunând o plaja de cautare de doar 2 unitati, cautarea pe nervura a adus cu aproximativ 16% mai multi antecedenti decât cautarea liniara. Dupa cum era de asteptat, pe masura ce lungimea textului cautat creste cele doua tipuri de modele se apropie în ceea pe priveste potentialul de a regasi legaturi co-referentiale. O cautare ierarhica înapoi într-o plaja de 5 unitati rezolva potential doar 70% dintre anafore, pentru ca o performanta potentiala de 90% sa poata fi atinsa doar daca se organizeaza o cautare într-o lungime de 12 unitati pe nervura. O alta investigare a urmarit compararea efortului necesar regasirii unui anumit antecedent în cele doua tipuri de abordari (liniara si ierarhica), unde prin efortul necesar gasirii unui antecedent se întelege numarul de unitati de discurs ce separa, în domeniu, unitatea anaforului de unitatea celei mai recente ancorari 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 continut 1200 de expresii referentiale, spatiul de cautare pentru legaturi co-referentiale s-a redus cu aproximativ 800 de unitati.

Un alt tip de investigatie empirica [Ide, Cristea, 2000] a urmarit frecventa referintelor evocative în comparatie cu cele post-evocative si depistarea unor corelatii între tipul de referinte si puterea de evocare a anaforilor. Studiul a comparat prezicerile avansate de VT relativ la domeniul de referentialitate evocativa cu cele ale modelului stiva al AST, corelând exceptiile (referinte ce nu se supun prevederilor celor doua teorii) cu puterea de evocare a anaforilor (pentru VT exceptiile marcheaza, evident, referinte din categoria celor inferentiale). Într-o ordine descendenta a puterii de evocare (v. si [Gundel *et al.*, 1993]) tipurile de anafori care dau nastere la exceptii sunt: referinte pragmatice > nume proprii > substantive comune > pronume. Pronumele constituie mijloace de referire foarte fragile. Un emitent al unui mesaj utilizeaza un pronume când e sigur ca structura permite recuperarea cu usurinta a entitatii referita de pronume. Practic, exceptând câteva cazuri în care un pronume putea fi înteles fara un antecedent (*our* în *our streets*, de exemplu), este imposibila utilizarea unui pronume pentru a referi o entitate aflata în afara *dea*. La extrema cealalta se plaseaza referintele pragmatice ce-si recupereaza antecedentul din cunostinte exterioare discursului si numele proprii. Interesant este ca aceasta sortare descrescatoare a tipurilor de anafori data de puterea de evocare se aliniaza numarului de exceptii raportate în cazul VT (56,3% – pragmatice, 22,7% – nume proprii, 16,0% – substantive comune si 5,0% – pronume) si nu are nici o semnificatie în cazul AST (0,0% – pragmatice, 26,1% – nume proprii, 39,1% – substantive comune si 34,8% – pronume). Ea probeaza corectitudinea conjecturii coeziunii.

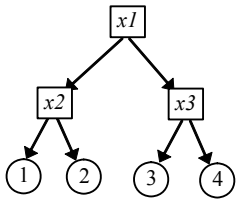
7. O proprietate de granularitate

Atunci când arborele de structura al discursului se modifica trecându-se de la o granularitate mai fina la una mai grosiera, constrângerea de accesibilitate, conjecturata de VT, se pastreaza.

Demonstratie

Sa presupunem un arbore de discurs *D*, notat la structura si pe care s-au calculat expresiile *head* si nervura ale nodurilor. O operatie de marire a granularitatii poate fi efectuata daca o întindere de text, initial repartizata în mai multe unitati, si **careia îi corespunde un nod în structura initiala**, este "compactata" într-o singura unitate de discurs mai mare ce va lua locul nodului radacina din structura initiala. Pentru a vedea în ce masura o astfel de operatie poate afecta accesibilitatea vom investiga rezultatul aplicarii ei asupra expresiilor *head* si nervura.

Definitia expresiei *head*, punctul 1, obliga ca expresia *head* a ceea ce înainte de compactare era un nod interior, fie el *n*, sa fi fost data de concatenarea unui sir de etichete de noduri nucleare aflate în secventa de text subîntinsa de *n*. Sa notam acest nod, dupa compactare, cu o eticheta compusa din secventa nodurilor terminale pe care le acopera. De exemplu, pentru arborele:



daca subarborele cu radacina *x2* ar fi compactat, atunci eticheta sa ar trebui sa fie notata 1-2, iar daca întregul arbore aflat sub *x1* ar fi compactat, atunci eticheta sa ar trebui sa fie notata 1-2-3-4 (e imposibil sa avem un nod notat 2-3).

Acest lucru înseamna ca, aplicând o compactare asupra unui arbore, în expresiile *head* ale nodurilor sale secvente de noduri vor fi acum înlocuite cu etichete compuse care contin cel putin aceleasi noduri, eventual mai multe, decât în expresiile originale. De exemplu, presupunând ca în arborele de mai sus, nodurile nucleare sunt *x2*, *x3*, 1 si 3, atunci, daca înainte de compactare am fi avut *head*(*x1*)=1 3, o expresie compusa din doua etichete, dupa compactarea întregului arbore vom avea *head*(*x1*)=1-2-3-4, adica o eticheta compusa, dar care include etichetele nodurilor ce apareau în expresia *head* originala. Vom numi astfel de expresii – expresii contrase si le vom nota cu *contr*(*e*), unde *e* este expresia corespunzatoare înainte de compactare (avem deci *contr*(1 3) = 1-2-3-4). Sa remarcam ca secventele de etichete din expresiile contrase sunt formate întotdeauna din etichete de noduri adiacente, ceea ce permite comutarea functiilor *seq* si *contr*: *seq*(*contr*(e_1), *contr*(e_2)) = *contr*(*seq*(e_1, e_2)).

Vom demonstra mai întâi ca expresiile nervura ale nodurilor din arborele compactat sunt obtinute din expresiile nervura originale prin înlocuirea expresiilor *head* originale cu expresiile contrase. Investigând definitiile expresiilor nervura, se poate constata ca nici o alta modificare nu apare în expresiile nervura cu exceptia expresiilor contrase. Într-adevar, cazul 1 se transcrie: expresia nervura a radacinii arborelui compactat reprezinta expresia *head* contrasa a arborelui original, adica *contr*(*h*), cu *h* – expresia *head* a radacinii arborelui originar.

Sa presupunem acum ca ne aflam într-un nod *n* ale carui expresii *head* si nervura pe arborele originar, necompactat sunt, respectiv *h* si *v*, iar *contr*(*h*) este expresia *head* pe arborele compactat. Consideram mai întâi cazul când *n* este fiu al nodului radacina, a carui expresie *head* este $contr(h_0)$, unde h_0 reprezinta expresia *head* pe arborele necompactat. Daca *n* este nuclear, atunci conform cazului 2 (sectiunea 3), avem doua subcazuri:

a) *n* nu are un frate nenuclear în stânga: atunci nervura sa este chiar nervura parintelui, adica $contr(h_0)$;

b) *n* are un frate nenuclear în stânga de *head* $contr(h_1)$: nervura nodului *n* va fi $seq(mark(contr(h_1)), contr(h_0)) = seq(contr(mark(h_1)), contr(h_0)) = contr(seq(mark(h_1), h_0)) = contr(v)$;

Daca *n* este un nod nenuclear, atunci conform cazului 3, avem:

a) *n* este în stânga: nervura sa este $seq(contr(h_0), contr(h)) = contr(seq(h_0, h)) = contr(v)$;

b) *n* este în dreapta: nervura sa este $seq(simpl(contr(h_0)), contr(h)) = seq(contr(simpl(h_0)), contr(h)) = contr(seq(simpl(h_0), h)) = contr(v)$.

Folosind inductia, se probeaza în mod analog ca expresia nervura a nodului *n* este o expresie contrasa si pentru cazul în care *n* este un nod interior, nu neaparat imediat sub radacina, fiu al unui nod de nervura $contr(v_0)$.

Cum expresia accesibilitatii este definita ca un prefix al expresiei nervura din care au fost îndepartate marcajele, iar nervurile sunt expresii contrase, deci eventual continând mai multe etichete de noduri, înseamna ca orice referinta care pe arborele originar satisface prima conjectura, cu alte cuvinte are loc între ultima unitate a unei expresii nervura si alta ce o precede, dupa compactare va satisface de asemenea conjectura, pentru ca nici o unitate nu a disparut din domeniu.

8. Discutii, aplicatii ale teoriei

Plecând de la o reprezentare a structurii de discurs similara celei din RST si în care esentiala este distinctia dintre nucleu si satelit, VT defineste nervura unui nod al arborelui ca secventa de unitati ale discursului ce sunt suficiente pentru a rezuma/interpreta întinderea de text acoperita de nod în contextul întregului discurs. Presupunerea principala pe care se bazeaza notiunea de nervura este ca *referintele inter-unitati sunt posibile cu*

precadere între unitati ce se afla într-o relatie structurala, chiar daca acestea sunt dispuse la distanta una de alta în text. Mai departe, referintele se realizeaza cu precadere spre unitati nucleare si doar în putine cazuri catre sateliti, reflectând intuitia ca nucleele gazduiesc ideile principale ale discursului. Acest lucru se regaseste în calculul expresiei nervurii pe arbori (binari) polarizati-stânga (pe orice nivel exista un nucleu în stânga), în care orice referinta se realizeaza dinspre un nucleu sau un satelit catre un nucleu aflat în stânga (desi, nu orice nucleu). Facând uz de echivalarea modelului stiva al lui Grosz si Sidner [1986] cu structura de arbore utilizata de RST [Mann, Thompson, 1988], similaritate demonstrata de Moser si Moore [1996] si Marcu [1999], predictiile VT asupra accesibilitatii referentiale sunt consistente cu cele ale modelului stiva. În cazurile în care însa arborele de discurs nu e polarizat-stânga (exista cel putin un satelit care precede nucleul sau, deci care apare ca frate stâng pe un nivel al structurii) VT ofera o interpretare mai naturala a accesibilitatii decât modelul stiva, corectând totodata slabiciunile acestuia. Într-adevar, într-o secventa *A*-satelit, *B*-nucleu, deci în care *B* domina *A* în termenii AST, *B* ar trebui sa apara în stiva pozitionat sub *A*, desi el este procesat în secventa dupa *A*. Totodata, VT formalizeaza intuitia ca într-o secventa de unitati *A*, *B*, *C*, unde *A* si *C* sunt sateliti ai lui *B*, *C* nu poate accesa *A* din cauza interpunerii unui nucleu, ce capteaza întreaga atentie.

Referentialitatea în lungul nervurilor este una naturala, usor de procesat si care, în general, nu necesita mijloace de evocare foarte puternice. Dimpotriva, iesirea din acest domeniu incumba utilizarea unor mijloace de evocare anaforica viguroase. Pe acest criteriu se face distinctia dintre referentialitate evocativa si ne-evocativa (sau inferentiala), referintele evocative fiind detaliate în directe si indirecte, iar între cele ne-evocative remarcându-se referintele pragmatice, ce nu necesita un antecedent pentru întelegere.

În privinta coerentei discursului, VT utilizeaza domeniile de referentialitate pentru a introduce notiunea de linie de argumentatie si a deduce din ea pe cea de segment în sens ierarhic ce generalizeaza segmentul în sens clasic (asa cum este el utilizat în AST si CT). Totodata VT avanseaza conjectura ca segmentul în sens ierarhic da o mai corecta interpretare a portiunilor de discurs ce se comporta din punctul de vedere al coeziunii si coerentei ca un tot unitar. Aplicând concluziile CT relative la coerenta discursului în lungul segmentelor în sens ierarhic CT poate fi generalizata pentru a o transforma într-o teorie globala a coerentei.

Au fost trecute în revista o seama de experimente care probeaza ca prezumtiile VT sunt corecte si independente de limba. Un aspect important îl constituie, de asemenea, faptul ca prezumtiile VT sunt stabile la schimbarea granularitatii în segmentarea discursului.

Aplicatiile VT se înscriu în trei directii importante: rezolutia anaforei, parsarea discursului si rezumarea automata. În [Cristea *et al*., 2002a] si [Cristea *et al*., 2002b] este descrisa o arhitectura care actioneaza ca un motor general si configurabil de rezolutie anaforica. Una dintre componentele oricarui model de rezolutie este o definitie a domeniului de referentialitate. Rezolutia anaforica se realizeaza, asadar, ghidata de structura de discurs.

În [Seretan, Cristea, 2002] se propune o abordare inversa, în care cunostinte asupra legarilor anaforice pot fi utilizate pentru corectarea structurii. Noi credem ca procesul de rezolutie anaforica si de construire a structurii de discurs sunt interdependente într-un asemenea grad încât în analiza de discurs ele trebuie sa aiba loc simultan. În interpretarea unui text exista o interconditionare reciproca între referinte si structura care trebuie sa conduca la obtinerea acelei reprezentari în care constrângerile, actionând ca forte, produc o stare de echilibru, ce trebuie sa fie un fel de stare de energie potentiala minima a sistemului. Oamenii dispun de un mecanism cognitiv care le permite sa ajunga în mod natural la cea mai plauzibila interpretare a unui text. Acest lucru este rasplatit de atingerea unei stari mentale „confortabile" ce trebuie sa-si aiba suportul în satisfacerea la maxim a unui sistem de constrângeri. În [Tablan *et al*., 1998] si [Cristea, 2000] se descrie un mecanism de parsare care modeleaza acest comportament uman. Prin combinarea unor scoruri contribuite de referinte cu scoruri contribuite de o analiza HCT se obtine cea mai fluida posibil structura de discurs (deci manifestând maxim de coerenta) si care prezinta maximul de referinte pe nervuri (fiind deci cea mai coeziva posibil).

Notiunea de *head* din VT este similara celei de multime de promovare (*promotion set*) pe care Marcu [Marcu, 2000] o utilizeaza pentru a obtine un rezumat ghidat de structura de discurs. Sa remarcam ca definitia nervurii presupune rezumarea ca o alternativa a întelegerii unei unitati de discurs în context. Credem ca valentele teoriei nervurilor în realizarea unei strategii de rezumare focalizata [Mani, 2001] pe o anumita entitate sau segment de discurs au fost doar tangential studiate pâna acum [Sofronie, 1999], [Postolache, 2001] si merita atentie în abordarile viitoare. Credem, de asemenea, ca fiind interesanta o directie de studiu care sa aprecieze maniera în care nervura poate constitui un cadru de sub-specificare a structurii [Schilder, 2001], plecând de la observatia ca structuri diferite (dau nu fundamental diferite) pot manifesta aceeasi expresie a nervurilor.

Bibliografie

Brennan, S.E.; Walker Fredman, M. and Pollard, C.J. 1987. A centering approach to pronouns. *Proc. 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 Computational Linguistics*, Madrid.

Cristea, D., Ide, N., and Romary, L. 1998. Veins Theory: A Model of Global Discourse Cohesion and Coherence, *Proceedings of the 17th 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 Computational 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, vol. 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 Conversational 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. *Computational Linguistics*, 12(2), p. 203-225.

Grosz, B.J. and Sidner, C. 1986. Attention, intentions, and the structure of discourse. *Computational 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, Ruqaiya. 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 Computational Linguistics (ACL-88)*, State University of New York, Buffalo, p. 163-169.

Ide, N. and Cristea, D. 2000. A Hierarchical Account of Referential Accessibility. *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, ACL'2000*, 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 rappelled et on résume les histories, *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 Functional Theory of Text Organization. *Text*, 8(3), p. 243–281.

Marcu, D., 1999. A formal and computational 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, vol. 63, p. 81-97.*

Moser, M. and Moore, J.D. 1996. Toward a synthesis of two accounts of discourse structure. *Computational 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.

Richadeau, F. 1969. La lisibilité. Langage-Typographie-Signes-Lecture, Paris.

Schank, R. and Abelson, R. 1977. Scripts, plans, goals and understanding, 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.

Sereţan, V. and Cristea, D., 2002. The Use of Referential 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.) *Computational 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. Functional Centering. *Proceedings of the 34th Annual Meeting of the Association for Computational 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., Iida, M., Cote, S. 1994. Japanese Discourse and the Process of Centering. *Computational Linguistics*, 20(2), p. 193-232.

Walker, M.A. 1996. Limited attention and discourse structure. *Computational 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 cautare documentara multilingv

Amalia TODIRASCU
INRIA Lorraine, LORIA, Campus scientifique BP 239,
54506 Vandoeuvre-lcs-Nancy Cedex, France,
todirasc@loria.fr

Abstract

Aceasta lucrare prezinta un sistem de cautare documentara bilingv francez-roman pentru un domeniu limitat, cel al securitatii computerelor. Cautarea si indexarea se fac pe baza unei ontologii comune domeniului. Identificarea instantelor conceptelor în texte sau în întrebarile utilizator se realizeaza cu ajutorul unor tehnici robuste de analiza limbajului natural, combinate cu o ontologie specifica domeniului.

Introducere

Sistemele de cautare de informatii indexeaza o baza de documente, în general pe baza unor liste de cuvinte cheie extrase din documentele respective. Ele primesc întrebarile utilizatorului, încearca o mapare a întrebarii cu indexul construit care regaseste documentele. Raspunsul sistemului contine un numar de documente care sunt relevante în raport cu întrebarea utilizatorului. Fiecare sistem defineste un criteriu de relevanta specific. Aceste sisteme sunt evaluate pe baza a doi parametri: rapel (numarul de documente regasite/numarul total de documente relevante) si precizie (numarul de documente relevante regasite/numarul de documente regasite). În cazul unui sistem de cautare multilingv, raspunsul la o întrebare poate contine mai multe documente relevante, chiar daca sunt scrise în alte limbi decît cea în care a fost formulata cererea.

Sistemele de cautare de informatii clasice ofera utilizatorului solutii imprecise sau vide. Aceste rezultate se datoreaza utilizarii ca indecsi doar a cuvintelor cheie, extrase pentru fiecare document în parte. Majoritatea sistemelor de cautare ignora fenomene caracteristice limbajului natural: fenomenul de ambiguitate (un cuvînt poate avea mai multe sensuri) sau de polimorfism (un concept poate fi exprimat în mai multe moduri). În plus, un sistem care îsi propune sa faca cautare într-o baza de date multilingva trebuie sa fie capabil sa gaseasca informatia ceruta în orice document disponibil, indiferent de limba în care a fost scris. Rezolvarea problemelor specifice limbajului natural (ambigitate, traducere automata) necesita resurse lingvistice importante pentru fiecare limba care este tratata de

catre sistem, daca aplicam tehnicile clasice de analiza limbajului natural. Tehnicile clasice de analiza sintactica nu sunt adaptate sistemelor de cautare documentara, datorita dimensiunilor prea mari ale bazei documentare.

Tehnicile robuste de analiza sintactica, inspirate de domeniul extragerii de informatii (GATE [6], FASTUS) sunt dedicate rezolvarii unor probleme dedicate, precise (identificarea numelor proprii, ale grupurilor nominale simple). Printre acestea, automatele cu stari finite [5], colocatii [9] sau liste de pattern-uri sintactice reprezentînd structura sintactica a grupului nominal simplu sunt resursele lingvistice necesare pentru aceste componente. Aceste tehnici au avantajul de a fi robuste, de a putea trata o cantitate importanta de informatii în timp real, precum si de a fi portabile de la un domeniu si/sau o limba la alta.

O alternativa 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 ca pe de o parte, ontologiile generice nu sunt disponibile decât în numar prea restrâns (WordNet \cite{vossen} si Corelex [3] sunt doar doua exemple de resurse libere). Pe de alta parte, textele nespecifice pun probleme analizoarelor existente, datorita faptului ca necesita resurse adaptate domeniului: dictionare, gramatici locale.

Într-o aplicatie de cautare de informatii pe un domeniu restrâns, asteptarile utilizatorului sunt altele decât cele pentru texte ne-specifice. Ne asteptam la o precizie mai buna a sistemelor. Aceasta impune folosirea de tehnici adaptate acestor sisteme, bazate pe existenta unui model redus al domeniului. În acest context, voi prezenta o metodologie de extragere a conceptelor candidat din corpus. Acestea sunt folosite de catre un expert uman pentru a îmbogati o ontologie existenta. De asemenea voi prezenta o metoda de indexare a documentelor pe baza unei ontologii, metoda care modifica metoda clasica de indexare semantica latenta.

2. Ontologii

Notiunea de ontologie este dificil de definit, mai multe puncte de vedere coexista. Pentru a simplifica, o ontologie este un model restrâns al unui domeniu specific, format din multimea claselor de obiecte ce populeaza acest domeniu si relatia lor cu celelalte clase.

Ontologiile reflecta un anumit grad de subiectivitate din partea expertului ce a definit-o. Fiecare expert poate avea o anumita viziune a claselor de obiecte ce trebuie incluse în descrierea ontologiei.

O problema a acestor ontologii este legata de portabilitate. O aplicatie definita pentru un anumit domeniu dat va trebui adaptata unui alt domeniu prin construirea unei ontologii corespunzatoare. Construirea lor manuala este dificila si trebuie tinut cont de posibilele redundante, incoerente, ce pot fi introduse în baza de cunostinte de catre expertul uman care o construieste. În ultimii ani, s-au facut eforturi deosebite pentru a putea reutiliza ontologiile existente: dezvoltarea unor formate de interschimb dedicate

(Knowledge Interchange Format - KIF), a standardelor (Ontology Interface Layer - OIL) [8], dezvoltate în cadrul proiectului Semantic Web (http://www.semweb.org).

Pentru a evita problemele legate de formatul în care a fost reprezentata ontologia, au fost propuse mai multe metode de extragere a ontologiilor din corpusuri. Acestea disting mai multe etape:

- identificarea termilor (posibilele instante ale conceptelor exprimate în limbaj natural);
- identificarea relatiilor între termi;
- identificarea relatiilor între termi si concepte.

Majoritatea acestor etape necesita validarea rezultatelor de catre un expert uman, sau asignarea unor interpretari (identificarea unei relatii între doua multimi de termi). Metodele statistice interpreteaza contexele existente si regrupeaza termii cu acelasi context în clase diferite [1], [7]. Relatiile între termi sunt interpretate pe baza informatiilor de subcategorizare asociate verbelor. Dezavantajul metodelor statistice este acela ca necesita corpusuri adnotate de talie importanta pentru a putea învata.

Metodele bazate pe inferente logice propun proceduri semi-automate pentru a verifica validitatea cunoasterii existente. Conceptele noi, deduse de catre regulile de inferenta, sunt adaugate ierarhiei domeniului daca sunt coerente cu cunoasterea existenta. Relatiile pot fi identificate printre cunoasterea sintactica (subcategorizare [4]). Problemele acestei metodologii sunt supragenerarea de concepte si costul verificarii incoerentelor si inconsistentelor cunoasterii sunt principalele neajunsuri ale metodei. Mai multe formalisme au fost dezvoltate în acest scop, si printre acestea logicile terminologice joaca un rol important.

2.1. Logici Terminologice

Logicile terminologice (LT) sunt formalisme de reprezentare a cunoştinţelor care sunt derivate din reţelele semantice, dar sintaxa şi semantica lor sunt bine definite. Ele combină proprietăţi ale sistemelor orientate-obiect, ale sistemelor bazate pe frame-uri şi ale logicilor modale.

LT propun o organizare ierarhica a cunoasterii, pe doua nivele: unul conceptual (T-Box), care descrie clasele abstracte continând obiectele relevante pentru modelarea domeniului si un nivel asertional (A-Box), continând instantele claselor. Clasele de obiecte (concepte) sunt descrise de relatii (numite roluri) cu alte concepte, si cu atributele lor (rolurile cu valori atomice).

2.1.1. Sintaxa ?i semantica logicilor terminologice

Operatorii LT sunt inspirati de logica de prim ordin:

Operator	Operator Logic	Semantica

D = SOME R C	$\exists$x (xRC)	Există cel puțin o instanță a lui C în relație cu R
D = ALL R C	$\forall$x(xRC)	restricționează co-domeniul relației R
D = AND C1 C2	C1$\wedge$C2	Conjuncția de descrieri conceptuale
D = OR C1 C2	C1$\vee$C2	Disjuncția de descrieri conceptuale
C1$\subseteq$C2	C1$\subseteq$C2	Axiom: C1 conține condiții necesare pentru C2
D= NOT C	$\neg$C	complementul conceptului C
D = (n.R.C	(y1...yn (1 (i(n, R(x, yi)(C(yi))	Există cel puțin n obiecte de tip C în relația R cu D

Figura 1. Operatori în LT

Folosind toti acesti operatori, sau doar o parte a acestora, mai multe expresivitati sunt posibile: definirea conceptelor si a rolurilor ALC (folosind SOME, ALL, AND, OR, NOT ca operatori, axiomele conceptuale), posibilitatea utilizarii rolurilor tranzitive (R+), a rolurilor inversabile (I), a ierahiilor de roluri (H), a atributelor (f) sau a restrictiilor numerice.

Unele comenzi LT sunt explicate mai jos. CN este un nume de concept, C este o descriere conceptuala (orice combinatie de operatori AND, SOME, NOT, ALL). Comenzile LT sunt inspirate de formalismul KRSS ([2]):

1. (define-concept CN C) - defineste un nou concept ca o descriere conceptuala;
2. (instance IN C) - defineste o instanta a unui concept dat;
3. (implies C1 C2) - introduce o noua axioma conceptuala, definind conditiile C1 necesare pentru descrierea conceptuala C2;

LT sunt fragmente decidabile ale logicii de prim ordin. Acestea propun algoritmi decidabili pentru verificarea coerentei si consistentei cunostintelor. LT propune mecanisme logice pentru a identifica subsumarea, regasirea instantelor, drumurile care unesc mai multe concepte. Clasificarea este o ordonare partiala a ierarhiei de concepte, în raport cu relatia de subsumare.

Câteva exemple de comenzi:

(concept-subsumes? C1 C2) testeaza daca C1 subsumeaza C2

(concept-parents C) regaseste stramosii directi ai conceptului C

(concept-children C) regaseste fii directi ai lui C

(classify-tbox) calculeaza toate relatiile de subsumare între toate conceptele definite în T-Box

(concept-instances C) regaseste toate instantele conceptul C

2.2. Logici terminologice pentru sisteme de extragere si de regasire a informatiilor

Rolul cunostintelor din domeniu într-un sistem de extragere a informatiilor este acela de a valida reprezentarea semantica a entitatilor care sunt potential elevante, identificate în text prin tehnici de procesare a limbajului natural. Aceste entitati pot fi folosite pentru a adauga noi concepte la ontologia existenta. Cea mai mare parte a sistemelor de extragere a informatiilor foloseste tehnici NLP robuste pentru identificarea candidatilor si entitatile candidat nu sunt validate de catre o interpretare semantica. Sistemele de extragere a informatiilor pot folosi cunoastere implicita, cum ar fi relatiile de hiponimie/hiperonimie.

Logicile terminologice prezinta avantajul de a lucra cu date semi-structurate sau incomplete. Nu este necesara definirea explicita unor valori ca instante ale unor concepte. Valorile implicite nu sunt utilizate de catre logicile terminologice. Unele valori ale rolurilor sunt lasate nespecificate ca în urmatorul exemplu:

(define-concept computer (and physicalobject (some hasOperatingSystem OSystem) (some hasType Type)))

(define-primitive-concept Type)

(define-primitive-concept OSystem)

(instance sun1 (and computer (some hasType SparcStation)))

În acest exemplu, vom ilustra faptul ca definitiile implicite sunt acceptate de catre logicile terminologice (SparcStation nu sunt definite explicit de catre o instanta sau un subconcept al conceptului Type). Nu este definita explicit nici o instanta a rolului **hasOperatingSystem**.

Aceste proprietati nu sunt interesante pentru aplicatia noastra, dar erorile sunt posibile, iar cunoasterea domeniului este incompleta.

Relatiile de hiperonimie sau hiponimie sunt tratate cu ajutorul relatiilor de subsumare între conceptele domeniului. De exemplu, daca 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 instanta 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 aplicatia noastra avem nevoie de o logica terminologica care sa propuna rationament la nivel de instanta, sa permita lucrul în contextul unei lumi deschise, precum si proceduri optimizate de calcul a relatiilor de subsumare sau de clasificare. Printre putinele sisteme care implementeaza rationament la nivel de instanta am ales RACER ([10]), fiind unul dintre cele mai performante.

În sectiunea urmatoare voi prezenta metoda de extragere a termilor din texte folosind sistemul DLIR [16]. Textele vor fi traduse într-o reprezentare conceptuala unica, permitând regasirea informatiilor în mai multe limbi.

3. Arhitectura

Sistemul DLIR contine mai multe module: un modul de analiza sintactica robusta, un modul de întretinere a ontologiei domeniului, un modul de indexare a documentelor bazat pe celelalte doua module. In cele ce urmeaza voi prezenta aceste module în detaliu.

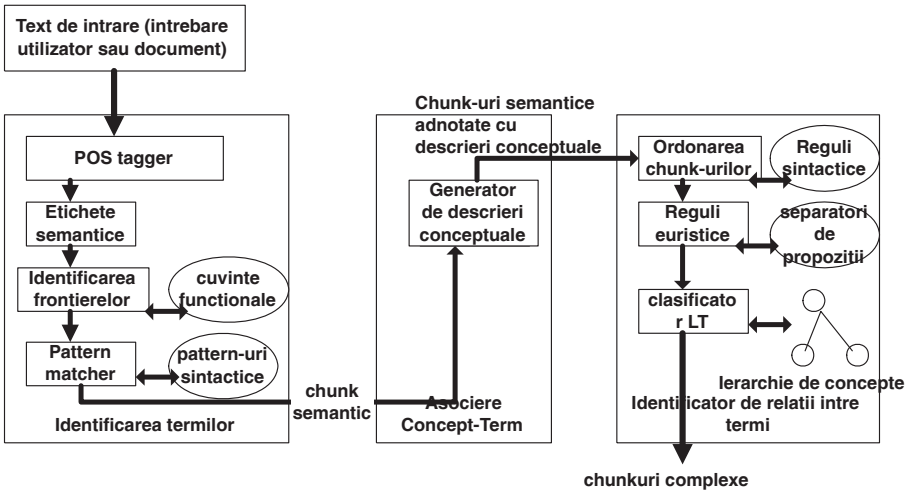


Figura 1: Instanţe ale conceptelor care apar în întrebare

3.1. Analiza sintactica robusta

Acest modul este dedicat identificarii termilor posibili, utilizând tehnici de analiza robuste, o serie de resurse specifice domeniului (o lista de corespondente cuvinte-concepte). Termii sunt combinati, conform unor reguli euristice pentru a crea concepte complexe. Acestea sunt validate ulterior, apelând modulul de acces la ontologia domeniului. Eventualele concepte valide sunt adaugate la ontologia existenta. Acest module contine mai multe submodule implementate în Java, în Perl si CLIPS (modulul care

aplica regulile de combinare a termilor). Notiunea de chunk semantic a fost propusa pentru a identifica termii candidat [16]. Metoda a fost testata pentru limba franceza, dar cum resursele folosite pentru identificarea termilor sunt relativ independente de limba pentru care a fost construita aplicatia, este posibila extinderea ei si pentru limba româna, dupa cum voi arata mai jos.

3.2. Identificarea chunk-urilor semantice

Scopul principal al acestui modul este acela de a identifica secventele de cuvinte care corespund celor mai semnificative concepte ale domeniului (*chunk-uri semantice*).

Un *chunk semantic* contine un pattern sintactic simplu (grup substantival simplu, grup verbal) si este delimitat de doi separatori de clauze.

Separatorii sunt cuvinte functionale, verbe auxiliare, sau anumite sintagme prepozitionale.

Exemplu."la victime d'une intrusion inattendue"

[victima unei intruziuni neasteptate]

În acest exemplu, "victima" si "unei intruziuni neasteptate" sunt chunk-uri semantice, care contin informatia relevanta.

Modulul contine mai multe submodule: un POS tagger, un tagger semantic, un identificator de frontiere si un pattern matcher. Identificarea chunk-urilor semantice este bazata pe informatia lexicala, propusa de POS tagger.

3.2.1. Part-Of-Speech tagging

Modulul care este dedicat identificarii partilor de vorbire asociate cuvintelor (folosind WinBrill, antrenat pentru franceza pe baza unui set de date propuse de Institut National pour la Langue Française [11]) identifica cuvintele care au un continut (substantive, adjective, verbe) si cuvintele functionale (prepozitii, conjunctii etc.).

Taggerul Brill foloseste un set de reguli contextuale si lexicale (bazate pe identificarea prefixelor si a sufixelor), învatate pe baza textelor adnotate, pentru a identifica partea de vorbire pentru cuvintele necunoscute.

Pentru limba româna, se foloseste QTAG adaptat pentru limba româna [17], datorita performantelor foarte bune (98\% rezultate corecte).

3.2.2. Tagger-ul semantic

Tagger-ul semantic contine un pattern matcher, care consulta un dictionar de talie redusa. Acesta contine o lista cu cele mai frecvente cuvinte si un set de sintagme asociate descrierilor conceptuale corespunzatoare.

Setul de descrieri conceptuale a fost stabilit de catre un expert pe baza unei liste de cuvinte si segmente repetate obtinute dintr-un corpus reprezentativ (200,000 cuvinte). Un

segment repetat este o succesiune de cuvinte care intervin într-un text cel putin de doua ori [14].

Acest modul asociaza fiecarui cuvânt conceptul sau descrierea conceptuala din dictionar. Un astfel de dictionar este creat pentru fiecare limba care este tratata de catre sistem.

3.2.3. Modulul pentru identificarea frontierelor

Acest modul identifica separatorii (cuvinte functionale sau constructii sintactice mai complexe) care delimiteaza chunk-urile semantice. Acest modul foloseste rezultatul POS tagger-ului (care identifica cuvintele functionale), precum si un set de sintagme (constituienti sintactici care contin auxiliare, prepozitii compuse). Setul de fraze este construit ca rezultat al studiilor corpusurilor de test pentru franceza si româna (200,000 cuvinte pentru fiecare limba). Separatorii grupurilor nominale si prepozitionale (determinanti, prepozitii) sunt cei mai buni candidati pentru identificare separatorilor de chunk-uri semantice; acestia reprezinta anumite relatii potentiale între concepte.

3.2.4. Pattern matcher

Scopul acestui modul este de a identifica nucleul chunk-urilor semantice, nucleu care este reprezentat de un grup nominal simplu sau un grup verbal.

Exemple. Un grup nominal simplu (în franceza) este identificat aplicând urmatoarele reguli:

N -⟩ NP

N ADJ -⟩ NP

Det N -⟩ NP

Det N ADJ -⟩ NP

3.2.5. DLgen

Acest modul interpreteaza informatia propusa de POS tagger si genereaza în mod automat o definitie de concept. Un expert trebuie sa verifice rezultatele acestui modul. Câteva exemple de reguli propuse pentru generarea descrierilor DL simple (valabile pentru ambele limbi):

- S1/N S2/ADJ este asociat definitiei (define-concept S1_S2 (AND S1 (SOME hasAtr "S2")))
- S1/N S2/NNP este asociat definitiei (define-concept S1 (SOME hasName "S2"))
- S1/ADJ S2/N este asociat definitiei (define-concept S2_S1 (AND S2 (SOME hasAtr "S1")))
- Verbele sunt traduse ca nume de roluri: S1/VB este asociat rolului **hasS1**.

Unele pattern-uri identifica negatiile, chiar daca este imposibil sa enumeram toate posibilitatile si sa detectam corect domeniul negatiei:

- sans/ADV S1/N este asociat definitiei (define-concept not_S1 (NOT S1
- nici_unul/ADV S1/N este asociat definitiei (define-concept not_S1 (NOT S1))

Rezultatele propuse de **DLgen** sunt 61% corecte. Iesirea este validata de un expert folosind clasificatorul LT pentru a verifica definitiile conceptuale obtinute în mod automat.

3.3. Relatii între termi

Acest modul foloseste inferentele LT, ca si regulile de sintaxa, pentru a combina descrierile conceptuale asociate fiecarui chunk semantic. Folosim un criteriu de ordonare al chunk-urilor, precum si reguli de combinare a conceptelor pentru a crea concepte complexe. Descrierile rezultante sunt validate de clasificatorul LT.

3.3.1. Ordonarea chunk-urilor

Modulul interpreteaza ordinea chunk-urilor si pozitia chunk-urilor în propozitie.

Clasificam chunk-urile în doua categorii: **chunk-uri principale** si **chunk-uri secundare**. Chunk-urile principale corespund notiunii de nucleu propuse de catre teoriile lingvistice clasice.

Chunk-urile secundare joaca rolul unui modificator, care adauga informatii suplimentare sensului nucleului. Chunk-urile secundare pot lipsi, dar restul propozitiilor este corect. Aceste exemple de reguli definesc chunk-uri diverse:

- chunk-urile care urmeaza dupa 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 commenc\'{e} Main] [Second r utiliser

les faux comptes Second]'

'atacurile au început prin utilizarea unor conturi false'

Cele doua chunk-uri principale detectate în exemplul de mai sus sunt primul chunk al propozitiei si verbul principal. Chunk-ul secundar este adnotat astfel pentru ca urmeaza dupa prepozitia **r**.

3.3.2. Reguli euristice

Regulile sunt stabilite de catre expert pe baza unui studiu asupra corpusului reprezentativ pentru fiecare limba. Corpusul a fost adnotat cu categoria lexicala propusa de

POS tagger si adnotat manual cu descrierile conceptuale. Setul de reguli heuristice este stabilit pe baza unei liste de pattern-uri de forma <*Chunk*1>?x/FW<*Chunk*1>.

Exemplu de reguli euristice sintactice: daca o prepozitie este un separator între doua chunk-uri semantice si prepozitia asociaza un substantiv cu un modificator, atunci putem combina descrierile conceptuale ale celor doua chunk-uri într-o descriere semantica mai complexa, rolul care leaga conceptele fiind cel de modificator:

if (<MainChunk1> <Border> <SecChunk2>)

and (Noun in MainChunk1)

and (Modifier in SecChunk2)

then (and sem(MainChunk1) (some hasModifier sem(SecChunk2)))

Fiecare pattern este asociat unui cuvânt tinta care identifica conditiile pentru aplicarea regulilor. Prepozitiile, verbele la modul participiu, sunt câteva exemple de cuvinte asociate regulilor euristice. Un numar de 43 reguli (pentru franceza) si un numar de 21 de reguli pentru româna au fost descrise în CLIPS. Iesirea acestor reguli va fi o serie de *chunk-uri complexe*, ce trebuiesc validate de catre expert, cu ajutorul ontologiei domeniului, care este independenta de limba.

3.4. Indexare semantica}

O posibilitate de indexare a documentelor este aceea de folosi direct concepte drept index si nu cuvinte cheie. O metoda eficienta de indexare o reprezinta indexarea semantica latenta. Aceasta metoda construieste o matrice document-cuvinte cheie si foloseste tehnici de descompunere a matricilor folosind metoda valorilor proprii. În acest fel se elimina coloanele si liniile care sunt vide. Propunem utilizarea conceptelor care fac parte din ontologie în locul cuvintelor cheie. Este posibil ca într-un sistem de cautare a informatiilor multilingv sa avem diferente între ontologiile dintr-o limba într-alta. Avantajul este ca putem folosi drept index concepte care sunt comune ambelor ontologii. Pentru aplicatia noastra am folosit o ontologie construita manual care contine 54 de concepte si 34 de relatii.

Numarul de concepte este mai redus decât numarul de termi, exploatând în special relatiile între termi.

Elementele matricii contin o pondere *weight*(*C*,*i*) calculata astfel:

$$weight(C,i) = \frac{f(C,i)}{\sum_{j=1}^{n} f(C,j)}$$

pentru fiecare concept, codificând frecventa instantelor conceptului în document si frecventa instantelor în toate documentele indexate de sistem.

f(C,i) - frecventa conceptului în documentul i;

Conceptele sunt legate prin rolurile dintre acestea. Frecventa unui concept care este situat în ierarhie foarte sus poate fi compus din suma frecventelor instantelor sale. Instantele conceptelor în LT sunt instantele tuturor subconceptelor si ale instantelor sale directe.

Indexarea documentelor se face aplicând metodele de extragere a termilor prezentate în sectiunea precedenta, înainte de a exploata sistemul. Se folosesc conceptele ontologiei care a fost construita manual. O serie de concepte mai generale ar putea fi obtinute scufundând ontologia specifica domeniului cu WordNet ([16]).

Evaluarea acestui sistem a fost realizata pentru un set restrâns de întrebari (50) numai pentru limba franceza. Rezultatele au fost comparate cu cele furnizate de un sistem care foloseste cuvinte-cheie pentru indexare. Pentru 74\% din întrebari raspunsurile sistemului (rapel si precizie) au fost comparabile cu cele obtinute prin metoda de indexare bazata pe cuvinte-cheie. În celelalte cazuri, raspunsurile au fost mai slabe decât indexarea pe baza de cuvinte cheie. Ontologia folosita este departe de a fi completa, ceea ce a dus la neidentificarea unor termi.

4. Concluzii si perspective

Articolul prezinta o modalitate de a folosi ontologia unui domeniu pentru cautare de informatii bilingva: franceza si româna.

Sistemul integreaza tehnici de analiza sintactica robusta pentru extragerea celor mai relevante chunk-uri semantice. Metoda foloseste o ontologie a domeniului construita manual. Pentru evaluarea pertinenta a metodelor de indexare pe baza de concepte, ontologia va fi actualizata si extinsa cu ajutorul rationamentelor propuse de logicile terminologice, ca si folosirea cunostintelor sintactice, folosite pentru extragerea unei reprezentari semantice pentru texte si întrebari. Expertul uman trebuie sa intervina pentru a decide daca conceptele identificate în texte pot fi adaugate ontologiei domeniului.

Referinte bibliografice

[1] Assadi, H., Bourigault, D., 2000, Analyse syntaxique et statistique pour la construction d'ontologies r partir des textes. In J.Charlet, M.Zacklad, G.Kassel, D.Bourigault (eds.) - Ingénierie des connaissances Evolutions récentes et nouveaux d\'{e}fis, Eyrolles Publishing House, pp. 243-256.

[2] Baader, F., Hollunder, B., 1991. A Terminological Knowledge Representation Systems with Complete Inference Algorithms, Proceedings of the Workshop on Processing Declarative Knowledge.

[3] Buitelaar, P., 1998. CORELEX: Systematic Polysemy and Under-specification, Ph.D. thesis, Brandeis University, Department of Computer Science

[4] Capponi, N., Toussaint, Y., 2000, Interprétation de classes de termes par généralisation de structures prédicat-argument. In J.Charlet, M.Zacklad, G.Kassel, D.Bourigault (eds.), Ingénierie des connaissances - Evolutions récentes et nouveaux défis, Eyrolles Publishing House, pp. 337-356.

[5] Chanod J.P., 1999. Natural Language Processing and Digital Libraries. In M.T.Pazienza (ed.), Information Extraction, Springer-Verlag, LNAI 1714, pp.17-31.

[6] Cunningham, H., Wilks, Y., Gaizauskas, R.J., 1996. New Methods, Current Trends and Software Infrastructure for NLP. In Proceedings of the conference on New Methods in Natural Language Processing (NeMLaP-2), Bilkent University, Turkey, 1996, pp.1-12.

[7] Daille, B., 1996, Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. In J.Klavans, P.Resnik (eds.) - The Balancing Act - Combining Symbolic and Statistical Approaches to Language, MIT Press, pp. 49-66.

[8] Fensel D. et al., 2000, OIL in a nutshell. In R. Dieng et al. (eds.), Knowledge Acquisition, Modeling, and Management, Proceedings of the European Knowledge Acquisition Conference (EKAW-2000), Lecture Notes in Artificial Intelligence, LNAI, Springer-Verlag.

[9] Heid, U., 2000, A linguistic bootstrapping approach to the extraction of term candidates from German text, Terminology, pp 161-180.

[10] Haarslev V., Muller R, 2001, Description of the RACER System and its Applications, Proceedings of the International Workshop on Description Logics (DL-2001), Stanford, USA, 1.-3. August 2001, pp. 132-141

[11] Lecomte, J., Le Catégoriseur BRILL14-JL5/WINBRILL-0.3, InaLF, InaLF/ CNRS report, December 1998.

[12] Riloff,E., Lorenzen, J., 1999, Extraction-based Text Categorization Generating Domain-Specific Role Relationships Automatically. In ed. T.Strzalkowski, Natural Language Information Retrieval, Kluwer Academic Publishers, pp. 167-196.

[13] Riloff, E., Shepherd, J., 1997, A Corpus-Based Approach for Building Semantic Lexicons. In 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] Todirascu, 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. In *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] P. 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 învatarii unor concepte dintr-o limba straina

Stefan TRAUSAN-MATU
Universitatea "Politehnica" Bucuresti, Facultatea de Automatica si Calculatoare,
Centrul de Cercetari Avansate în Învatare Automata,
Prelucrarea Limbajului Natural si Modelare Conceptuala al Academiei Române
email: trausan@cs.pub.ro, trausan@racai.ro
URL: www.racai.ro/~trausan

1. WWW, o prezenta din ce în ce mai comuna

În mai putin de zece ani, reteaua globala de documente World Wide Web (WWW sau, pe scurt web), a devenit omniprezenta si este posibil ca într-un timp nu prea lung sa înlocuiasca o mare parte din carti, televizorul, cinematograful, ziarele si revistele (toate acestea fiind deja disponibile pe web) si, în plus sa furnizeze chiar posibilitatea imersiunii în realitati virtuale. Un singur exemplu cred ca este suficient: anul trecut rezultatele bacalaureatului au fost publicate pe web.

WWW a atins deja dimensiuni comparabile cu imensa Biblioteca a Congresului SUA. Extinderea sa este datorata usurintei cu care poate fi parcurs de catre oricine are un calculator, pe de o parte, si de simplitatea cu care oricine poate publica ceva pe el. Pe de alta parte, costul accesului la resursele web este de cele mai multe ori infim.

WWW este un hipertext extins la scara întregului glob prin rețeaua mondială de calculatoare Internet. Pe fiecare calculator pot fi plasate unul sau mai multe documente care constituie noduri (pagini) în hipertext. Oriunde într-o astfel de pagină poate exista o legătură la o alta pagină, de pe acelaşi sau de pe alt calculator, în acest mod putând fi unite informaţii aflate în locuri diferite. O nouă pagină pentru web poate fi creată uşor chiar de utilizatori nu neapărat profesionişti în informatică, în acest scop existând mai multe editoare de texte specializate.

Termenul de hipertext se pare ca provine de la termenul de spatiu hiperbolic sau hiperspatiu, aparut în 1704 si folosit de matematicianul F. Klein pentru geometria cu mai multe dimensiuni [Rad91]. Din aceasta perspectiva, un hipertext este un text cu mai multe dimensiuni explicite (fata de doar o dimensiune, în cazul textului liniar). De fapt, orice text are implicit mai multe dimensiuni, deoarece, chiar daca forma de prezentare a unui text este liniara, pe hârtie, în el exista o structura implicita, data de discurs. De asemenea, exista conexiuni implicite, subiective între parti ale textului, concepte legate între ele, Hipertextul

este o organizare a unui text în care toate aceste legaturi sunt explicitate si pot fi exploatate în parcurgerea facuta pe un calculator.

În jurul anului 1962, Douglas Engelbart a dezvoltat primul sistem hipertext, prezentat atunci drept o arhitectura conceptuala destinata cresterii potentialului intelectului uman ("Conceptual Framework for Augmenting Human Intellect") [Eng95]. Sistemul era destinat manipularii de concepte structurate într-o retea în care arcele sunt relatiile între concepte.

Primul sistem declarat ca fiind hipertext a fost creat de Theodor Nelson în 1967 sub numele de "Xanadu". Nelson îsi propunea atunci sa dezvolte un sistem, masiv paralel, destinat muncii creative si studiului. El a plecat în îndeplinirea acestei idei de la dorinta de a gasi cea mai buna abstractie care sa unifice literatura si arta cinematografica.

Sistemele hipertext (hipermedia) permit accesul personalizat la volume imense de informatii. În acelasi timp, însa, ele sufera de problema afluxului de informatie cu care este bombardat un utilizator. O solutie este dezvoltarea de instrumente, aplicatii, medii informatice pentru facilitarea accesului la cunostintele dorite pe web. Aceste instrumente trebuie sa facilteze întelegerea, abstractizarea textelor, extragerea informatiilor utile. Acesta este unul din motivele pentru care le-am denumit instrumente hermenofore. Trebuie remarcat faptul ca ideea de a considera hipertextele ca intrumente de sprijinire a activitatilor cognitive a stat chiar la baza conceperii acestora, dupa cum am precizat în paragrafele anterioare.

În continuare, dupa o trecere în revista a problematicii ontologiilor, în sectiunea urmatoare se va introduce conceptul de mediu hermenofor, se va justifica necesitatea acestuia si se vor prezenta caracteristicile acestora. Lucrarea va fi încheiata cu o exemplificare de sistem care are câteva trasaturi ale unui mediu hermenofor si de o sectiune de concluzii.

2. Ontologii

Termenul de “ontologie” a fost, pâna nu de mult, folosit exclusiv în filosofie, pentru a denumi teoria asupra existentei, mai corect spus, asupra ceea ce considera ca exista cel care întocmeste teoria. Construirea multor sisteme filosofice pleaca de la o ontologie, adica de la definirea categoriilor fundamentale de entitati din realitate si a relatiilor dintre ele. Chiar daca ontologia nu este întotdeauna explicita, orice demers conceptual construieste o ontologie, chiar implicit, inconstient.

În ultimii ani, termenul de ontologie este folosit si în stiinta calculatoarelor. Cea mai frecventa extindere a folosirii acestui concept este în cadrul sistemelor de inteligenta artificiala bazate pe cunostinte. Majoritatea programelor de calculator cu inteligenta artificiala prelucreaza structuri de simboluri, care sunt menite a reprezenta conceptele, cunostintele referitoare la domeniul considerat. Aceste structuri simbolice sunt grupate într-o asa numita baza de cunostinte care constituie, de fapt, un model al domeniului respectiv.

În ultimii ani se considera ca aceasta baza de cunostinte trebuie vazuta ca o ontologie, o conceptualizare, o teorie asupra ceea ce exista în domeniul avut în vedere. O ontologie este, din aceasta perspectiva, o "specificare a unei conceptualizari … Termenul este împrumutat din filosofie, unde însemna o considerare sistematica a existentei. În inteligenta artificiala se refera la precizarea a ceea ce se considera ca <<exista>>" [Gru96].

Între concepte pot exista diverse relatii. Cea mai importanta relatie este probabil cea hiperonimica [WN], taxonomica, între un concept si unul sau mai multe concepte mai generale, din care deriveaza, care îl subsumeaza, din a caror combinatie a fost generat. Prin aceasta relatie se pot "mosteni" proprietati de la conceptul (conceptele) mai general(e) la cel mai particular, daca aceste proprietati nu sunt redefinite la conceptul din urma. Alte relatii sunt cea meronimica [WN] ("parte-întreg"), între un concept si partile sale sau cea antonimica, între doua concepte (adjective) opuse.

O ontologie include, asadar:

- categoriile, conceptele fundamentale,
- proprietatile conceptelor,
- relatiile si distinctiile între concepte.

O ontologie este rezultatul unei experientieri, a unor experiente traite, în care sunt evidentiate niste constante, niste regularitati, care ne îndreptatesc sa afirmam ca vor fi regasite în viitor. În urma investigatiei facuta pentru a gasi esenta regularitatilor se delimiteaza entitati mentale denumite concepte sau categorii, care pot fi diferentiate de alte categorii. Aceste entitati pot intra în combinatie cu altele formând noi concepte.

Un aspect deosebit de important în ceea ce priveste rolul ontologiile este faptul ca ele exprima o comuniune, (co)existenta unei diversitati de concepte, cu diferente si relatii între ele. O presupozitie este ca exista doar un numar limitat de concepte sau categorii, ceea ce înseamna ca se poate face un fel de cuantificare, de discretizare a realitatii. Acestea constituie un punct de sprijin pentru achizitia de noi concepte sau pentru rationamentele facute de om sau de calculator.

Partajarea unei ontologii este esentiala în sistemele bazate pe agenti (programe) inteligenti pentru, de exemplu, comertul electronic, pentru a le asigura autonomia, flexibilitatea si agilitatea. Ontologiile sunt liantul care integreaza sisteme de baze de date, sisteme de obiecte, sisteme bazate pe cunostinte, în diverse aplicatii integratoare si bazate pe colaborare. Ele reduc ambiguitatile semantice în partajarea si reutilizarea cunostintelor. "Scopul suprem este dezvoltarea de ontologii reutilizabile care pot fi aplicate pentru mai multe discipline". [OORG]

"O ontologie are drept prim scop facilitarea comunicarii între calculatoare, independent de tehnologiile unui anumit sistem individual, arhitectura de prelucrare a informatiilor si domeniul aplicatiei. Ingredientii cheie care constituie o ontologie sunt un vocabular de termeni de baza si o specificare precisa a ceea ce înseamna acesti termeni." [OORG] O ontologie este însa mai mult decât un vocabular. Ea este punctul de plecare

pentru dezvoltarea de structuri de cunostinte, nu numai taxonomii sau clasificari de concepte ci si relatii complexe. [OORG]

Din punct de vedere al programelor de calculator care folosesc ontologiile, exista doua tipuri de ontologii. Primul tip este cel al ontologiilor destinate sistemelor bazate pe cunostinte, de exemplu, al unui sistem de diagnostic medical. Aceste ontologii sunt caracterizate de un numar relativ redus de concepte, dar legate între ele printr-un numar mare si variat de relatii. Conceptele sunt grupate în scheme conceptuale complexe sau scenarii. Pentru fiecare concept pot exista una sau mai multe particularizari.

Spre deosebire de primul tip de ontologii, ontologiile lexicalizate includ un numar foarte mare de concepte, legate printr-un numar redus de tipuri de relatii (de exemplu, hiperonimica, meronimica etc.). Conceptele sunt reprezentate, de exemplu în WordNet [WN], prin multimi de cuvinte sinonime. Astfel de ontologii sunt folosite în sistemele de prelucrare a limbajului uman.

Corespondenta ontologiei WordNet (care este conceputa pentru limba engleza-americana) pentru limbile europene este EuroWordNet. Aceasta din urma aduce avantajul ca, fiind dezvoltata pentru mai multe limbi (engleza, franceza, germana, italiana, olandeza etc.), permite si dezvoltarea de aplicatii multilingve. În prezent, în cadrul Centrului de Cercetari Avansate în Învatare Automata, Prelucrarea Limbajului Natural si Modelare Conceptuala al Academiei Române este în desfasurare, în colaborare cu mai multe tari din regiunea balcanica proiectul BalkanNet pentru integrarea în EuroWordNet a limbilor din zona, inclusiv a limbii române.

3. Medii hermenofore

Denumim mediu hermenofor o colectie integrata de instrumente (pe care le vom numi hermenofore) si aplicatii informatice directionate catre facilitarea unor activitati de tip hermeneutic ale unui utilizator care exploreaza resurse aflate pe web. Termenul "hermenofor" [Tra01] poate fi parafrazat prin "generator de hermeneutica", pentru a sugera faptul ca un mediu hermenofor faciliteaza activitati hermeneutice, care acorda un rol important experientierii si sunt orientate spre descoperirea unor întelesuri, a unor structuri profunde, greu detectabile.

Elaborarea de medii hermenofore este absolut necesara în contextul actual al exploziei numarului si volumului de resurse si a interconexiunilor între acestea pe web. Sistemele hipertext (hipermedia) aduc noi dimensiuni cum ar fi interactivitatea, posibilitatile cu totul remarcabile de vizualizare, accesul personalizat la volume imense de informatii. În acelasi timp, însa, ele introduc si unele probleme datorate afluxului de informatie, care poate duce la depasirea capacitatilor cognitive ale utilizatorului, la dezorientare si chiar la alienare. Este un fapt ca utilizatorul, chiar profesionist în informatica, poate fi dezorientat în "labirintul" de pagini de web si resurse de tot felul (baze de date, documente, imagini, ontologii, lexicoane etc.) interconectate.

O solutie la problemele enumerate mai sus este dezvoltarea de instrumente, aplicatii, medii informatice pentru facilitarea accesului la cunostintele dorite pe web. Se poate spune, din aceasta perspectiva, ca browserele de web, "motoarele de cautare", agentii (asistentii) software sunt rudimente de medii hermenofore. Justificarea necesitatii considerarii perspectivei hermenofore este lipsa abilitatilor hermeneutice ale acestor aplicatii. Un exemplu tipic este faptul ca "motoarele de cautare pe web" (de exemplu Google [Goo]) furnizeaza mii sau chiar zeci de mii de documente ca raspuns la o cerere. Alt exemplu este limita actuala a programelor de calculator în întelegerea textelor cu scopul traducerii, sumarizarii sau extragerii cunostintelor. Aceste probleme sunt datorate, în primul rând, problemelor generate de ambiguitatea limbajului natural, a aspectelor legate de semantica, de pragmatica, de interpretare, de considerarea contextului, a metaforelor, a cunostintelor de "bun simt". Toate aceste probleme sunt recunoscute ca fiind "nodul gordian" al aplicatiilor de inteligenta artificiala. Dupa cum remarca Teryy Winograd, programele de inteligenta artificiala nu pot depasi conditia unui birocrat, care nu poate sa actioneze când nu are "reguli", care nu se implica [Win87]. Putem spune ca, de fapt, problema este ca acestor aplicatii le lipsesc abilitatile hermeneutice. Ideea noastra este de a oferi un cadru în care puterea oferita de tehnologia informatiei sa fie integrata cu capabilitatile specific umane.

Hermeneutica este, dupa opinia lui P. Ricoeur, o abordare complementara celei structuraliste în analiza limbajului, a întelesului si simbolismului cultural. "Hermeneutica bazeaza întelegerea textelor pe intentiile si istoria autorilor si relevanta acestor fapte pentru cititori. În contrast, filosofia analitica identifica de obicei întelesul cu referenti externi pentru texte iar structuralismul gasind întelesul în aranjarea cuvintelor. Hermeneutica priveste textele ca mijloace pentru a transmite experienta, crezurile si judecatile de la un subiect sau comunitate catre altii. Astfel, determinarea întelesurilor este o problema de judecata practica si rationament de <<bun simt>> si nu privitor la o teorie a priori sau o demonstratie stiintifica." [MHD].

Hermeneutica este studiul interpretarii, initial ea referindu-se doar la interpretarea textelor [MHD]. În prezent s-a extins acceptiunea termenului hermeneutica, vorbindu-se de o pozitie hermeneutica în filosofie, care include pe Heidegger, Gadamer, Habermas si Ricoeur, deosebita de formalisti (filosofia analitica, neo-pozitivism sau pozitivismul logic), reprezentati prin Descartes, Leibniz si Russell [Wes97]. Distinctia între cele doua abordari pleaca de la problema capturarii întelesului. Pe când formalistii pretind ca pot reprezenta întelesul, semantica, doar prin identificarea unui denotat în lumea reala corespunzator unei expresii formale, adeptii hermeneuticii neaga aceasta posibilitate, pentru ei întelesul implicând si considerarea experientei, a credintelor subiectului. Se poate spune ca, dintr-un punct de vedere se ajunge la aceeasi disputa dintre Husserl si Heidegger sau dintre Dennett si Chalmers.

Mediile hermenofore furnizeaza informatiile dorite dintr-o perspectiva particulara, pentru un anumit utilizator, considerând un anumit domeniu si într-un anumit moment dat. Un mediu hermenofor trebuie conceput deci în scopul personalizarii interfatarii la resursele

web-ului, pentru a facilita întelegerea. Daca prezentarile facute într-un mediu hermenofor sunt structurate ca hipermedia, una din preocuparile principale ce trebuie avute în vedere este faptul ca utilizatorul trebuie sa experientieze parcurgerea unei secvente de pagini de web, secventa care trebuie sa respecte niste reguli de pragmatica.

În plus fata de furnizarea unei interfete adaptabile, o alta caracteristica a unui mediu hermenofor trebuie sa fie facilitarea initiativei utilizatorului. El trebuie sa poata experimenta, sa poata investiga resursele web-ului. Instrumentele hermenofore sunt destinate sprijinirii activitatii hermeneutice umane adica a unei atitudini directionate catre întelegerea unor cunostinte sau structuri ascunse în texte (hipertextelor, hipermedia). Un rol important în procesul întelegerii îl au modalitatile de a genera experientieri, adica experiente de traire, fapte de viata (conform teoriei ca întelegerea necesita un proces empatic [Wri95], [Mar97]). Unul dintre cele mai uzitate mijloace de acest gen este folosirea metaforelor [LaJ80], [Tra00]. În acest sens se înscrie preocuparea de a dezvolta instrumente (hermenofore) pentru detectarea, adnotarea si prelucrarea metaforelor.

O caracteristica pe care o consideram esentiala la un mediu hermenofor, în contextul precizat mai sus, este si posibilitatea de vizualizare multipla, din perspective diferite, a aceluiasi document. Enumeram aici, drept exemplu, în afara perspectivei continutului "brut" al unui document, alte perspective, date de concordante, adnotari (cu parti de vorbire, de exemplu), extrase, rezumate, arbori de analiza semantica, structuri care reprezinta continutul semantic. Remarcam, în acest context, rolul extraordinar de important al adnotarilor documentelor în limbajul extrem de versatil care este XML [XML].

Vom considera ca instrumentele hermonofore au ca scop revelarea si valorizarea unor cunostinte sau a unor structuri încorporate în volumele imense de hipetexte si hipermedia de pe web. Datorita faptului ca abordarea hermeneutica pune pe prim plan rolul experientierii umane, un instrument hermenofor trebuie neaparat considerat în relatie cu utilizatorul care îl foloseste. De aceea, el trebuie sa aiba asociat modelul utilizatorului, care sa contina cel putin urmatoarele informatii despre utilizator:

- ontologia sa,
- scopurile urmarite,
- profilul psihologic,
- istoricul actiunilor efectuate,
- preferintele sale (explicite sau implicite, derivate din observarea comportamentului sau).

Pe de alta parte, instrumentele hermenofore trebuie sa considere si aspectele legate de particularitatile autorilor documentelor :

- ontologiile considerate (de exemplu, ontologiile impuse de paradigmele sau de practicile domeniilor considerate),
- scopurile presupuse,

- elemente de istoric,
- aspecte psihologice general umane.

Instrumentele hermenofore pot fi împartite în mai multe clase, în functie de actiunile efectuate:

- cautare a documentelor relevante,
- categorizare a documentelor conform unei taxonomii predefinite,
- relevare de regularitati (de exemplu, colocatii) sau structuri în documente,
- segmentarea textelor,
- extragere de informatii sau cunostinte din documente,
- sumarizare,
- relevare de structuri pe web [WSD97],
- instrumente de adnotare (la nivel sintactic, semantic sau pragmatic) a documentelor.

Spre deosebire de instrumentele de minerit al textelor ("text mining"), instrumentele hermenofore pun, în plus, accentul pe aspectele legate de istoricul interactiunii, de experienta utilizatorului.

În sectiunea urmatoare se va prezenta sistemul GenWeb de instruire asistata a învatarii terminologiei financiare într-o limba straina [TMC02], [ABK02] care a fost dezvoltat ca un modul într-un proiect mai mare, denumit „Larflast" si finantat de Comunitatea Europeana. GenWeb a implementat instrumente hermenofore care identifica si utilizeaza metafore pentru a facilita întelegerea unui anumit concept [Tra00]. În acest scop, el cauta metafore în texte considerate relevante. Metaforele sunt identificate printre perechile de cuvinte care corespund la concepte din ontologia domeniului considerat (finante) si din ontologia metaforelor, aceasta din urma reflectând aspecte psihologice general umane [LaJ80]. Trecerea de la un concept la o multime de cuvinte (sinonime sau înrudite) se face pe baza ontologiei WordNet, derivata din investigatii psiholingvistice [WN]. Metaforele sunt adnotate în XML [XML], unul din atributele folosite în adnotare fiind scopul urmarit de autor [Tra00].

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 învata poate experientia. Tot pe post de instrumente hermenofore, în GenWeb este disponibila vizualizarea de concordante în context.

5. Sistem de instruire asistata cu calculatorul în întelegerea unor termeni financiari

Exista mai multe puncte de vedere asupra modului cum are loc un proces de învatare. Suntem de partea abordarii constructiviste [BlM96, TrA97, TNA98, Wil96] în conceperea proceselor educationale. Aceasta abordare considera ca fiecare dintre noi ne construim propria realitate, propriul bagaj de cunostinte, plecând de la experientele pe care le-am avut [ErK97]. Dupa cum remarca [BlM96], "Nucleul studiului este activitatea hermeneutica a constructiei de interpretari." Învatarea poate fi si ea vazuta constructivist ca un proces hermeneutic, de întelegere, de transpunere în domeniul studiat, de experimentare, de traire.

Plecând de la ideile învatarii constructiviste se ajunge la urmatoarele principii [ErK97]:

- învatarea este un proces activ în care studentii experimenteaza, cauta sa înteleaga singuri ceea ce învata, profesorul fiind mai mult un îndrumator;
- învatarea trebuie sa fie un proces auto-reglat de catre studenti;
- învatarea constructiva este un proces situational în sensul ca studentul trebuie introdus într-un mediu de învatare care îi permite sa experimenteze, în care se pot face simulari;
- învatarea trebuie sa fie sociala, trebuie sa existe o permanenta colaborare a studentului cu colegii lui.

Dintr-o alta perspectiva, învatarea poate fi considerata ca un proces de inducere de modele mentale adecvate [JoL83]. Întelegerea poate fi vazuta astfel ca momentul în care realitatea supusa comprehensiunii este pusa în corespondenta cu un model mental complet si valid. Empatia, identificarea eu-lui cu starea de lucruri considerata poate fi, în acest caz, tocmai sentimentul de "traire" în lumea modelului mental.

O practica deja raspândita este de a dezvolta sisteme inteligente de asistare cu calculatorul a instruirii ("Intelligent Tutoring Systems") care încearca sa monitorizeze procesul de învatare prin verificarea asimilarii conceptelor din ontologia domeniului considerat [Tra95]. Se considera ca un model adecvat al cunostintelor elevului poate fi construit prin raportare la aceasta ontologie. De fapt, aceasta metoda este folosita si în învatamântul traditional: noii termeni sunt introdusi prin genul proxim si diferenta specifica. În termenii ontologiilor, noii termeni sunt definiti prin superconceptele care-i subsumeaza si prin particularitatile care-i diferentiaza.

Orice profesor stie însa ca astfel de definitii sunt necesare dar nu sunt suficiente. Pentru a aprofunda termenii definiti sunt necesare exemple, imagini cu un grad mai mare sau mai mic de iconicitate, plecând de la poze si schite, diagrame si grafice, pâna la imagini sugerate, pâna la metafore. Acest fapt este prezent nu numai în învatamânt, el apare în orice proces de comunicare (învatamântul fiind, bineînteles, si el inclus).

În cele ce urmeaza nu ne vom referi la utilizarea imaginilor propriu-zise, care faciliteaza evident învatarea sau comunicarea. Vom considera un caz particular de imagini, mentale, sugerate, semne iconice lipsite de caracterul vizual dar care comunica o experientiere (de multe ori chiar mai puternic, printr-un efect care ar putea face sa ne gândim la perceptia subliminala). Este cazul metaforelor, care sunt folosite într-o proportie de cele mai multe ori nebanuit de mare în comunicarea inter-umana.

Pentru a ilustra puterea de expresie a metaforelor si, bineînteles, rolul lor în întelegerea unor termeni, am sa exemplific prin metafora "actiunile la bursa sunt niste creaturi foarte sensibile" (gasita într-un text pe situl de web al Bursei din New York - http://www.nyse.com). Nu este nevoie sa ne imaginam o anumita creatura concreta pentru a întelege ce sugereaza metafora exemplificata. Succesul unei metafore, puterea ei expresiva, capacitatea de comunicare sunt date de masura în care "rezonam" la mesajul transmis. Ori ce este mai percutant pentru un om decât faptul ca suntem creaturi extrem de sensibile? Prin urmare, succesul metaforei folosita într-un context foarte pragmatic, al discursului unui specialist în finante este determinat de inspiratia vorbitorului de a se referi la un fapt general uman. Nici o definitie de tip gen proxim-diferenta specifica nu poate comunica experienta referitoare la aspectul foarte fragil al actiunilor la bursa precum o face metafora de mai sus.

Rolul covârsitor al metaforelor în viata noastra a fost remarcat si de Lucian Blaga ("omul este un animal metaforic" [Bla85]) si a fost foarte bine evidentiat de Lakoff si Johnson într-o lucrare cu un puternic impact ("Metaforele cu care traim" - "Metaphors we live by" [LaJ80]. Cei doi autori americani considera ca "subcategorizarea si metaforele sunt doua extremitati ale unei continuum", ca metaforele "formeaza sisteme coerente în care ne conceptualizam experienta" [LaJ80]. Putem spune deci ca metaforele ofera alte mijloace expresive decât cele de categorizare oferite de ontologii. Ele nu tin de logica lui Ares, care categorizeaza, ci de logica lui Hermes, propusa de Noica [Noi86].

Dintr-o alta perspectiva, metaforele pot fi considerate instrumente empatice, care determina imersiunea cititorului (receptorului) în lumea experientelor autorului. Acest fapt era evidentiat si de Lakoff si Johnson: "Esenta metaforei este întelegerea si experientierea unui lucru prin altul" [LaJ80]. De exemplu, metafora amintita mai sus despre actiunile la bursa ne comunica o informatie pe care orice fiinta vie o întelege (sensibilitatea, perisabilitatea) dar care nu poate fi exprimata în categorizari.

Importanta metaforelor a fost revelata si de studiul preliminar facut în cadrul proiectului Larflast (care a avut drept scop elaborarea unui sistem de asistare cu calculatorul a învatarii terminologiei financiare într-o limba straina [Lar], [TMC02], [ABK02]) de o profesoara de limba engleza la o facultate economica din Sofia. Dânsa remarca ca o imoprtanta dificultate "întelegerea metaforelor. Limbajul economic si financiar este extrem de metaforic si, uneori, grupuri de metafore apar în imagini complexe. Deseori cuvinte uzuale sunt folosite în metafore elaborate, … cum ar fi <<a sustine o pierdere>>" [Vit99].

Proiectul Larflast a inclus mai multe module tipice pentru sisteme inteligente de instruire, cum ar fi o ontologie, un mecanism de inferenta, teste (grila) pentru diagnosticarea cunostintelor elevului si actualizarea modelului acestuia. Sistemul dezvoltat include cinci servere de web, unul la Bucuresti si altele la Leeds, Manchester, Montpellier si Sofia. Serverul de la Bucuresti, dupa ce este lansat, acceseaza serverul de la Sofia pentru a prelua modelul elevului (ce concepte stie si ce concepte nu) si apoi genereaza pagini de web personalizate.

Metaforele sunt identificate în texte considerate relevante care au fost obtinute în urma cautarii cu o masina de cautare uzuala (de exemplu, Google [Goo]. Textele gasite sunt grupate într-un corpus care este adnotat cu metaforele identificate. Acest corpus, împreuna cu ontologia domeniului si cu modelul studentului (construit pe baza raspunsurilor date de student la teste) sunt folosite pentru generarea personalizata de pagini de web. În figura urmatoare este ilustrata arhitectura sistemului GenWeb.

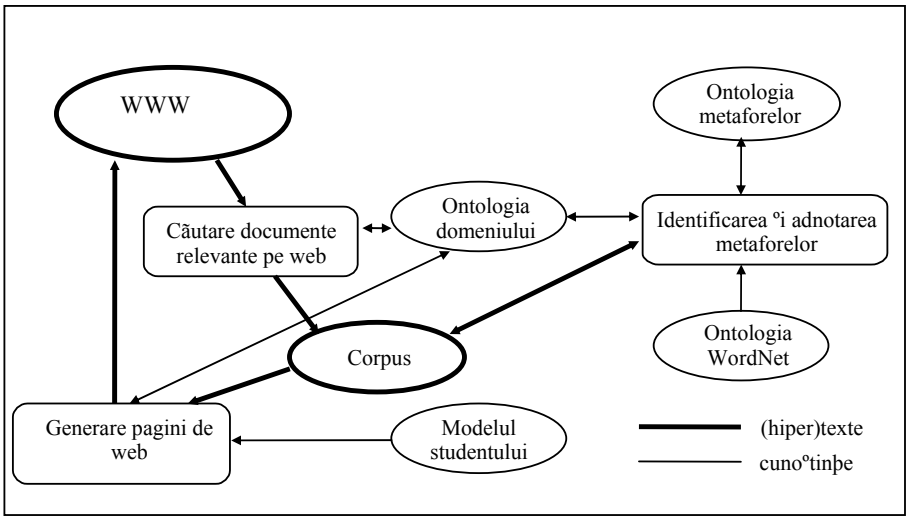


Figura 1

Pentru identificarea si adnotarea metaforelor a fost implementat un editor semantic specializat (fig.2) si un editor de concepte (fig.3).

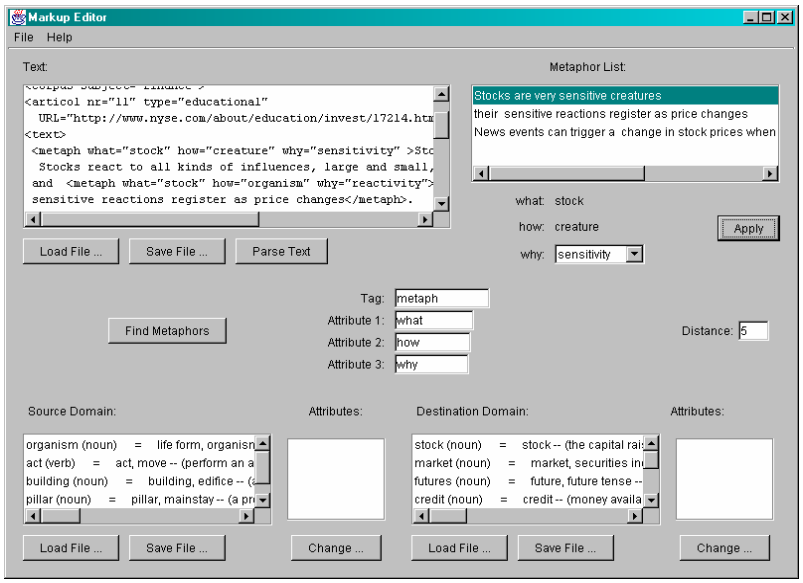


Figura 2

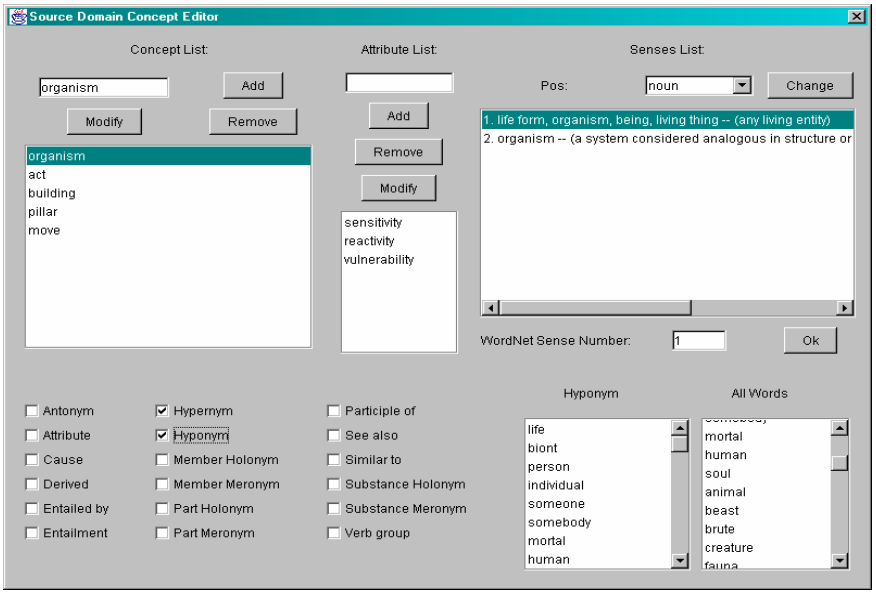


Figura 3

Modelul studentului este creat pe baza raspunsurilor la teste:

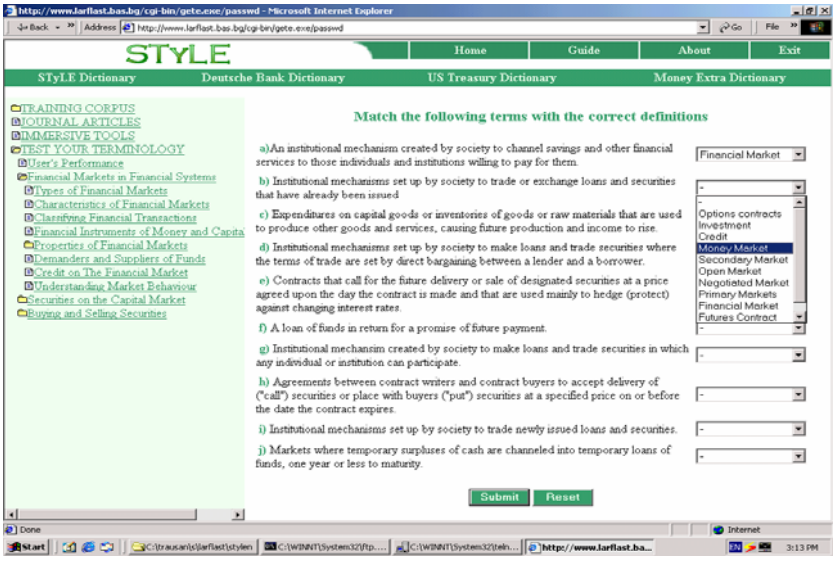


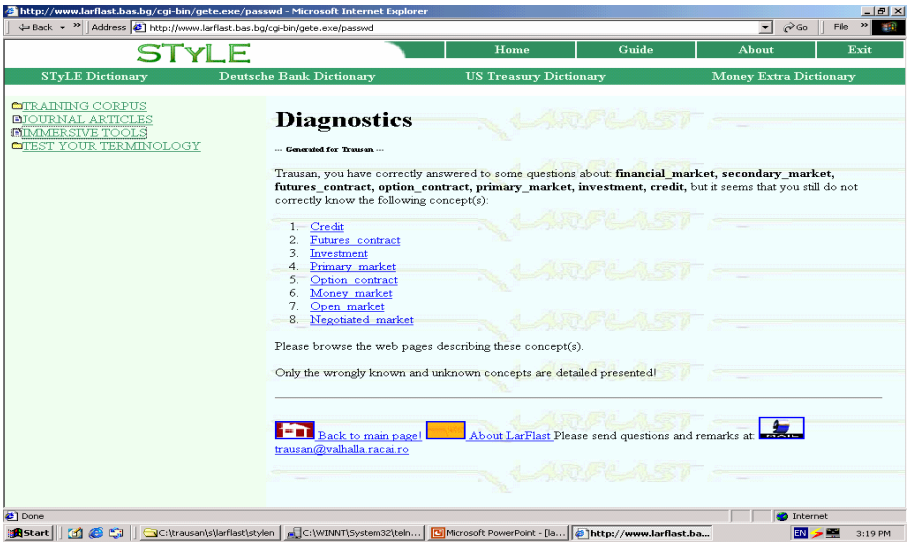
Figura 4

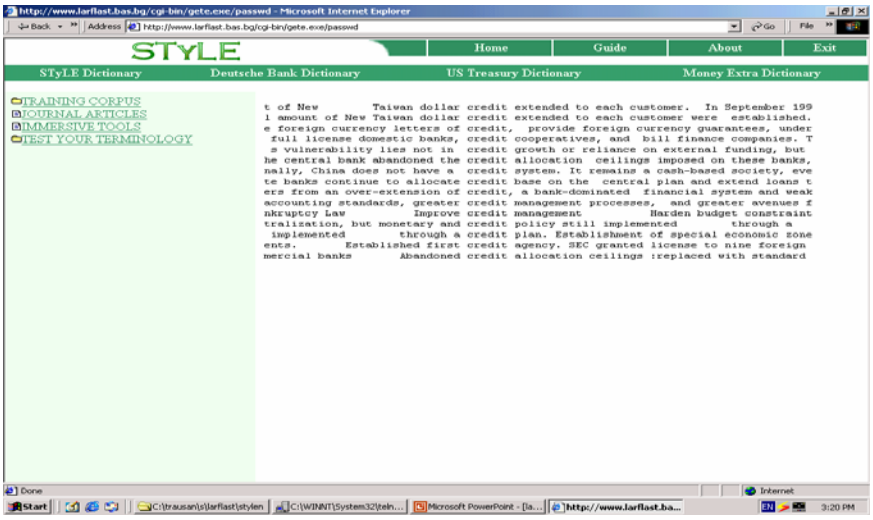
Paginile de web generate dinamic de modulul GenWeb, contributia românească la proiectul Larflast, se constituie în structuri care reflecta structura conceptuala (ontologia) a domeniului considerat. Parcurgerea acestora poate fi vazuta si în corespondenta cu facilitatile oferite de o Arta a memoriei [Cul94]. Din alta perspectiva, structurile trebuie concepute în ideea unei retorici specifice paginilor de web [Cli95], [THH95].

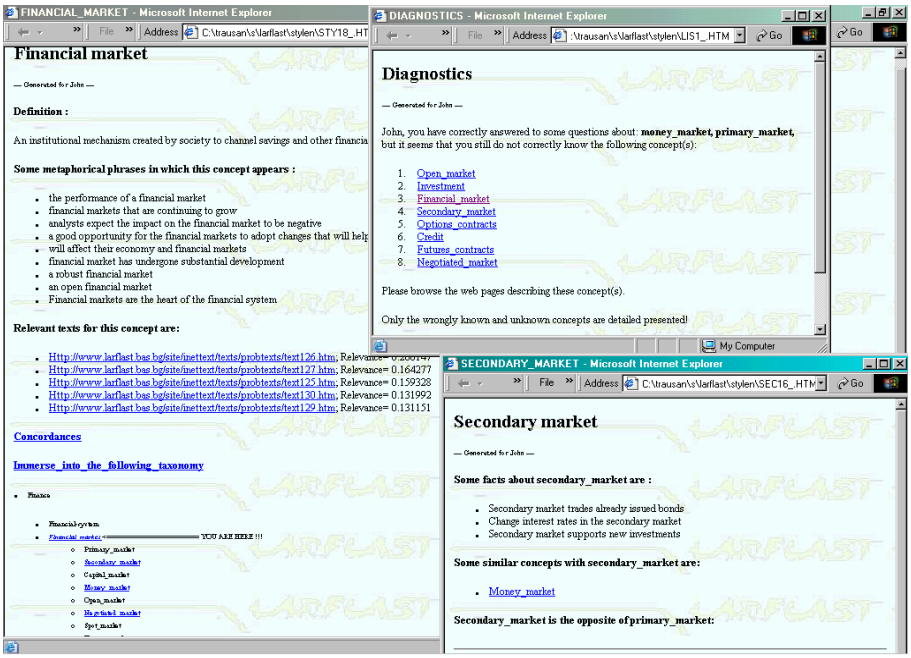
Sunt mai multe tipuri de pagini de web generate:

- pagini de diagnostic,
- pagini care definesc concepte, dau exemple de metafore si care includ structuri de paggini web care reflecta ontologia domeniului,
- pagini din structurile de mai sus,
- pagini cu concordante în context.

Aceste pagini sunt ilustrate în figurile urmatoare.


http://www.larflast.bas.bg/cgi-bin/gete.exe/passwd - Microsoft Internet Explorer
STyLE
Home
Guide
About
Exit
STyLE Dictionary
Deutsche Bank Dictionary
US Treasury Dictionary
Money Extra Dictionary
TRAINING CORPUS
JOURNAL ARTICLES
IMMERSIVE TOOLS
TEST YOUR TERMINOLOGY
Diagnostics
--- Generated for Trausan ---
Trausan, you have correctly answered to some questions about: financial_market, secondary_market, futures_contract, option_contract, primary_market, investment, credit, but it seems that you still do not correctly know the following concept(s):
1. Credit
2. Futures_contract
3. Investment
4. Primary_market
5. Option_contract
6. Money_market
7. Open_market
8. Negotiated_market
Please browse the web pages describing these concept(s).
Only the wrongly known and unknown concepts are detailed presented!
Back to main page!
About LarFlast
Please send questions and remarks at:
trausan@valhalla.racai.ro


http://www.larflast.bas.bg/cgi-bin/gete.exe/passwd - Microsoft Internet Explorer
STyLE
Home
Guide
About
Exit
STyLE Dictionary
Deutsche Bank Dictionary
US Treasury Dictionary
Money Extra Dictionary
TRAINING CORPUS
JOURNAL ARTICLES
IMMERSIVE TOOLS
TEST YOUR TERMINOLOGY
t of New Taiwan dollar credit extended to each customer. In September 199
l amount of New Taiwan dollar credit extended to each customer were established.
e foreign currency letters of credit, provide foreign currency guarantees, under
full license domestic banks, credit cooperatives, and bill finance companies. T
s vulnerability lies not in credit growth or reliance on external funding, but
he central bank abandoned the credit allocation ceilings imposed on these banks,
nally, China does not have a credit system. It remains a cash-based society, eve
te banks continue to allocate credit base on the central plan and extend loans t
ers from an over-extension of credit, a bank-dominated financial system and weak
accounting standards, greater credit management processes, and greater avenues f
nkruptcy Law Improve credit management Harden budget constraint
tralistion, but monetary and credit policy still implemented through a
implemented through a credit plan. Establishment of special economic zone
ents. Established first credit agency. SEC granted license to nine foreign
mercial banks Abandoned credit allocation ceilings :replaced with standard



Concluzii

În contextul dezvoltarii explozive a numarului de documente pe web este absolut necesara existenta unor medii care sa permita utilizatorilor explorarea în scopul extragerii cunostintelor din texte si structuri de documente web. Aceasta activitate trebuie sprijinita de ontologii, un rol foarte important avându-l integrarea ontologiilor de mari dimensiuni existente astazi pe web. În concluzie, un mediu hermenofor integreaza instrumente hermenofore cu ontologii într-o arhitectura în care utilizatorul trebuie sa poata experimenta, sa investigheze diverse transformari ale textelor. Se poate spune ca un mediu hermenofor înglobeaza sinergic instrumente de prelucrare a cunostintelor cu instrumente de prelucrare a textelor si cu tehnici specifice web.

Bibliografie

[ABK02] G. Angelova, S. Boytcheva, O. Kalaydjiev, St. Trausan-Matu, P. Nakov, A. Strupchanska, Adaptivity in a web-based CALL system, in F. van Harmelen (ed.): ECAI 2002. Proceedings of the 15th European Conference on Artificial Intelligence, IOS Press, Amsterdam, 2002, to appear

[Bla85] L. Blaga, Trilogia culturii, Ed. Minerva, 1985

[BlM96] Black, J.B., McClintock, An Interpretation Construction Approach to Constructivist Design, in B.G. Wilson (ed.), Constructivist Learning Environments: Case Studies in Instructional Design, Education Technology Publications, 1996.

[Cli95] Clibbon, K., Conceptually Adapted Hypertext For Learning, Proceedings of CHI'95, http://www.acm.org/sigchi/chi95/Electronic/documnts/kc_bdy.html

[CTr01] Constandache, G.G., St. Trausan-Matu, Ontologia si hermeneutica calculatoarelor, Editura Tehnica, 2001.

[Cul94] Culianu, I.P., Eros si magie în Renastere; 1484, Nemira, Bucuresti 1994.

[Eng95] Engelbart, D.G., Toward Augmenting the Human Intellect and Boosting our Collective IQ, CACM No.8, Vol.38, Aug. 95, pp. 30-33.

[ErK97] Ertl, B., Kraan, A.G., Internet-Based Learning Environments from a Constructivist point of view, Proceedings of RILW, Ilieni, 1997, p. 17-21.

[Goo] http://www.google.com

[Gru96] Gruber, T., What is an Ontology, http://www.kr.org/top/definitions.html

[JoL83] Johnson-Laird, P.N., Mental Models - Towards a Cognitive Science of Language, Inference, and Consciousness, Cambridge Univ. Press, 1983.

[LaJ80] Lakoff,G., Johnson, M., Metaphors We Live by, The University of Chicago Press, 1980.

[Lar] LarFLaST, http://www-it.fmi.uni-sofia.bg/larflast/

[Mar97] Marcus, S., Empatie si 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

[Sow99] J. Sowa, Knowledge Representation: Logical, Philosophical and Computational Foundations, Brooke Cole Publishing Co., Pacific Grove, CA, 1999, vezi si [CTr01].

[THH95] Thiring, M., Hannemann, J., Haake, J.M., Hypermedia and Cognition: Designing for Comprehension, Communications of the ACM, vol.38, no. 8, pp. 57-66, aug. 1995.

[TMC02] St. Trausan-Matu, D. Maraschi, S. Cerri, Ontology-Centered Personalized Presentation of Knowledge Extracted From the Web, in S.Cerri, G.Gouarderes (eds.), Intelligent Tutoring Systems 2002, Springer, Lecture Notes in Computer Science number 2363, to appear.

[Tra95] St. Trausan-Matu, Programe inteligente pentru asistarea invatarii, in Revista

Romana de Informatica si Automatica, vol.5, nr.4, 1995, pag. 7-16.

[Tra00] St. Trausan-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] St. Trausan-Matu, Interfatarea evoluata om-calculator, Ed. MatrixRom, 2001.

[Wes97] D.West, Hermeneutic Computer Science, CACM, Vol.40, No.4, pp. 115-116, 1997, si în [CTr01].

[Wil96] B.G. Wilson (ed.), Constructivist Learning Environments: Case Studies in Instructional Design, Education Technology Publications, 1996

[Win87] T. Winograd, 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., Explicatie si întelegere, Humanitas, 1995.

[WSD97] http://www.research.att.com/~suciu/workshop-papers.html

[XML] www.w3.org/xml

SECȚIUNEA III

TEHNOLOGII ALE LIMBAJULUI VORBIT

Experimente în vederea recunoaşterii vorbitorului

Corneliu BURILEANU,
Universitatea „Politehnica" din Bucureşti, Spl.Independeţei 303
cburileanu@messnet.pub.ro
Luigi BOJAN,
Graphco Technologies Inc., Newton, PA, USA

1. Introducere

Având în vedere funcţia realizată şi concomitent, sarcina de îndeplinit, tehnologia vorbirii se poate clasifica în mai multe domenii [1, 2]:

- Recunoaşterea automată a vorbirii. Se bazează pe analiza automată a semnalului vocal şi are în vedere informaţia transmisă de om maşinilor care "îl ascultă". Din această informaţie, maşina este programată să extragă acele caracteristici ce îi vor permite să deceleze cine vorbeşte, ce vorbeşte, în ce fel şi în ce condiţii.
- Sinteza automată a vorbirii. Se realizează răspunsul "prin voce" al maşinilor către operatorul uman.
- Codificare/decodificare (analiză şi sinteză) a vorbirii. Se referă la tehnici de compresie a informaţiei conţinută în semnalul vocal în vederea unor prelucrări ulterioare specifice sarcinii de îndeplinit.

Un domeniu interdisciplinar important, legat în mod esenţial de aplicaţiile de recunoaştere şi sinteză automată ale vorbirii este cel al dialogului om-maşină.

Termenul "comunicare om - maşină" pare forţat: maşina nu este o entitate socială, nu are nici scop nici cultură. Ea nu poate acţiona în lumea reală în sensul de a putea să răspundă corect la întrebări de genul: "ai putea să închizi uşa, te rog?". Ea nu este "conştientă" decât de propria sa "lume". Avem într-adevăr nevoie de a comunica cu maşinile? Au importanţă intenţiile lor, chiar dacă le-ar avea? Ce poate să-mi comunice sau să mă facă să ştiu o maşină?

Maşina îmi procură "uneltele" pentru a realiza o sarcină, ea mă face să proiectez noi obiecte (eventual, virtuale), ea mă aduce într-un univers artificial, îmi permite să utilizez un mediu de programare împreună cu alţi utilizatori umani, pentru a lucra într-o manieră cooperantă în acelaşi mediu informatic. Maşina se prezintă deci ca un *factor de interacţiune*. Ea trebuie să-mi furnizeze un spaţiu de 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 încredinţate.

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 conţine o varietate de informaţii utile: ce se vorbeşte, cine vorbeşte, în ce fel şi în ce condiţii. În cadrul recunoaşterii se pune problema identificării unui anumit tip de informaţii; de pildă, recunoşterea cuvintelor rostite înseamnă determinarea mesajului (ce se vorbeşte) indiferent (sau ajutându-se) de variabilităţile introduse de vorbitor (cine), maniera de a vorbi (în ce fel) şi zgomotul ambiental (în ce condiţii). 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 semnificaţie şi care, când e înţeleasă, 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 opţiuni fundamentale de abordare. Punctul de vedere adoptat poate viza prelucrarea unui semnal acustic ca oricare altul, poate ţine seama de mecanismul producerii vorbirii, poate simula recepţia senzorială, sau poate folsi modelul uman al percepţiei vorbirii.

Termenul de **recunoaştere a vorbitorului** desemnează orice aplicaţie 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 aplicaţii:

· **verificarea vorbitorului** îşi propune să determine dacă un fragment de semnal vocal aparţine sau nu unui anumit vorbitor [5, 6, 7, 8]. Există doi parametri care caracterizează performanţele sistemului: respingerea adevăratului vorbitor şi acceptarea unui impostor. Considerând un set de N vorbitori, informaţia (în biţi) obţinută este

$$I_{ver} = 1 \tag{1}$$

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

$$I_{ident} = log_2(N) \tag{2}$$

considerând probabilitatea de identificare *a priori* egală pentru toți vorbitorii.

Rezultă că, potențial, un sistem automat de verificarea vorbitorului are performanţe 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 condiţii normale de procesare [15]. Variaţia (lentă) în timp a formei tractului vocal este aproximată printr-o serie de secvenţe 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 esenţial de frecvenţele sale de rezonanţă, se ajunge la un model al cărui parametri se pot deduce prin rezolvarea unui sistem de ecuaţii liniare. Deşi aproximările avute în vedere par destul de restrictive, *analiza prin predicţie liniară* (LPC) dă rezultate deosebite pentru că semnalul vocal are o redundanţă deosebită; este motivul pentru care metoda ne permite să aproximăm un eşantion de semnal printr-o combinaţie 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, evoluţia 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 amănunţită a influenţelor 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 excitaţiei glotale, tractului vocal şi radiaţiei buzelor poate fi făcută într-o modalitate care ţine seama mai detaliat de fiecare efect în parte [18, 19].

În concluzie, presupunerile fundamentale care stau la baza parametrizării propuse sunt:

- efectele excitaţiei tractului vocal şi ale tractului propriu-zis pot fi separate;
- tractul vocal este invariant pe durate scurte de timp, ceea ce are drept rezultat obţinerea unui model descris de un sistem liniar ai cărui parametri variază lent în timp (constanţi "în timp scurt").

Fundamental pentru modul în care concepem abordarea analizei semnalului este asimilarea *analizei* cu *parametrizarea* semnalului şi, în consecinţă, cu *compresia* sa. Alegerea parametrilor a avut în vedre şi considerente pragmatice:

- complexitatea prelucrării;
- gradul de compresie,
- tipul de aplicaţie,
- în ce măsură parametrii sunt semnificativi şi robuşti.

O primă variantă a schemei bloc care descrie funcţionarea sistemului de recunoaşterea vorbitorului este prezentată în fig. 1. Blocul de *preprocesare* presupune filtrarea şi achiziţia semnalului în condiţii normale pentru orice sistem de recunoaştere. În această secţiune vom descrie obţinerea *cepstrului* pornind de la *analiza LPC*, iar în secţiunea următoare vom descrie principiile *cuantizării vectoriale* şi deci procedura de recunoaştere propriu-zisă.

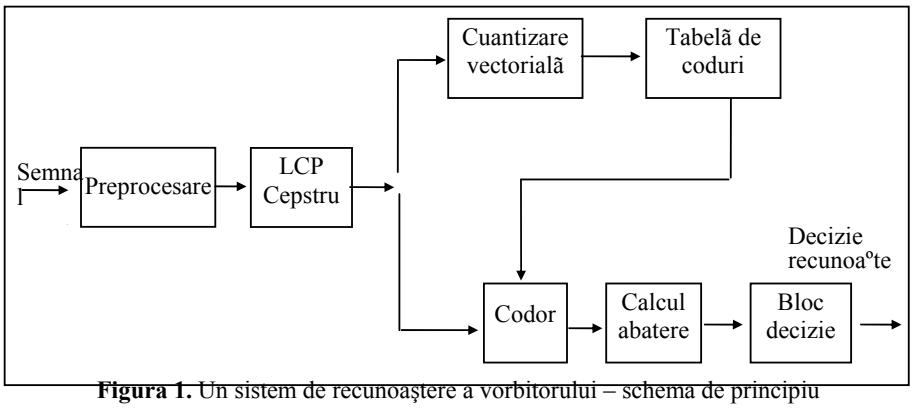


Figura 1. Un sistem de recunoaştere a vorbitorului – schema de principiu

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 "

$$s(n) \xrightarrow{H} \hat{s}(n) \tag{4}$$

în care $\hat{s}(n)$ va fi numit *cepstrul complex* asociat semnalului $s(n)$.

Prin definiție

$$\hat{S}(z) = \sum_n \hat{S}(n) \cdot z^{-n} \equiv ln\, S(z) \tag{5}$$

Astfel, cepstrul complex asociat semnalului devine

$$\hat{S}(n) = \hat{e}(n) + \hat{v}(n) \tag{6}$$

ceea ce permite separarea componentelor printr-o "filtrare temporală" aplicată cepstrelor

$$s(n) \xrightarrow{H} \hat{s}(n) \begin{cases} \xrightarrow{L} \hat{e}(n) \xrightarrow{H^{-1}} e(n) \\ \xrightarrow{L} \hat{v}(n) \xrightarrow{H^{-1}} v(n) \end{cases} \tag{7}$$

Obținerea parametrilor cepstrali se poate realiza ținând seama de câteva proprietăți ale cepstrului.

Fie $c(n)$ partea pară a cepstrului complex al semnalului

$$c(n) = [\hat{s}(n) + \hat{s}(-n)] / 2 \tag{8}$$

Secvența $c(n)$ se numeşte *cepstrul real* al semnalului $s(n)$

$\hat{s}(n)$ este o secvență cauzală – ca şi $s(n)$; rezultă

$$\hat{s}(n) = c(n) \cdot \begin{cases} 0 & pentru \quad n < 0 \\ 1 & pentru \quad n = 0 \\ 2 & pentru \quad n > 0 \end{cases} \tag{9}$$

Cum transformata "*z*" a unei secvențe cauzale e determinată complet prin partea reală a transformatei sale Fourier, rezultă

$$c(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} ln\left|S\left(e^{j\omega}\right)\right| e^{jn\omega} \, d\omega \tag{10}$$

Vom prefera calculul coeficienților cepstrali din coeficienți α_i ai analizei prin predicție liniară (LPC) conform relațiilor recursive:

$$\begin{aligned} c(1) &= -\alpha_1 \\ c(i) &= -\alpha_i - \sum_{n=1}^{i-1} \left(1 - \frac{n}{i}\right) \cdot \alpha_n \cdot c(i-n) \quad i > 0 \end{aligned} \tag{11}$$

Figura 2 prezintă evoluția coeficienților cepstrali pentru o voce feminină şi una masculină.

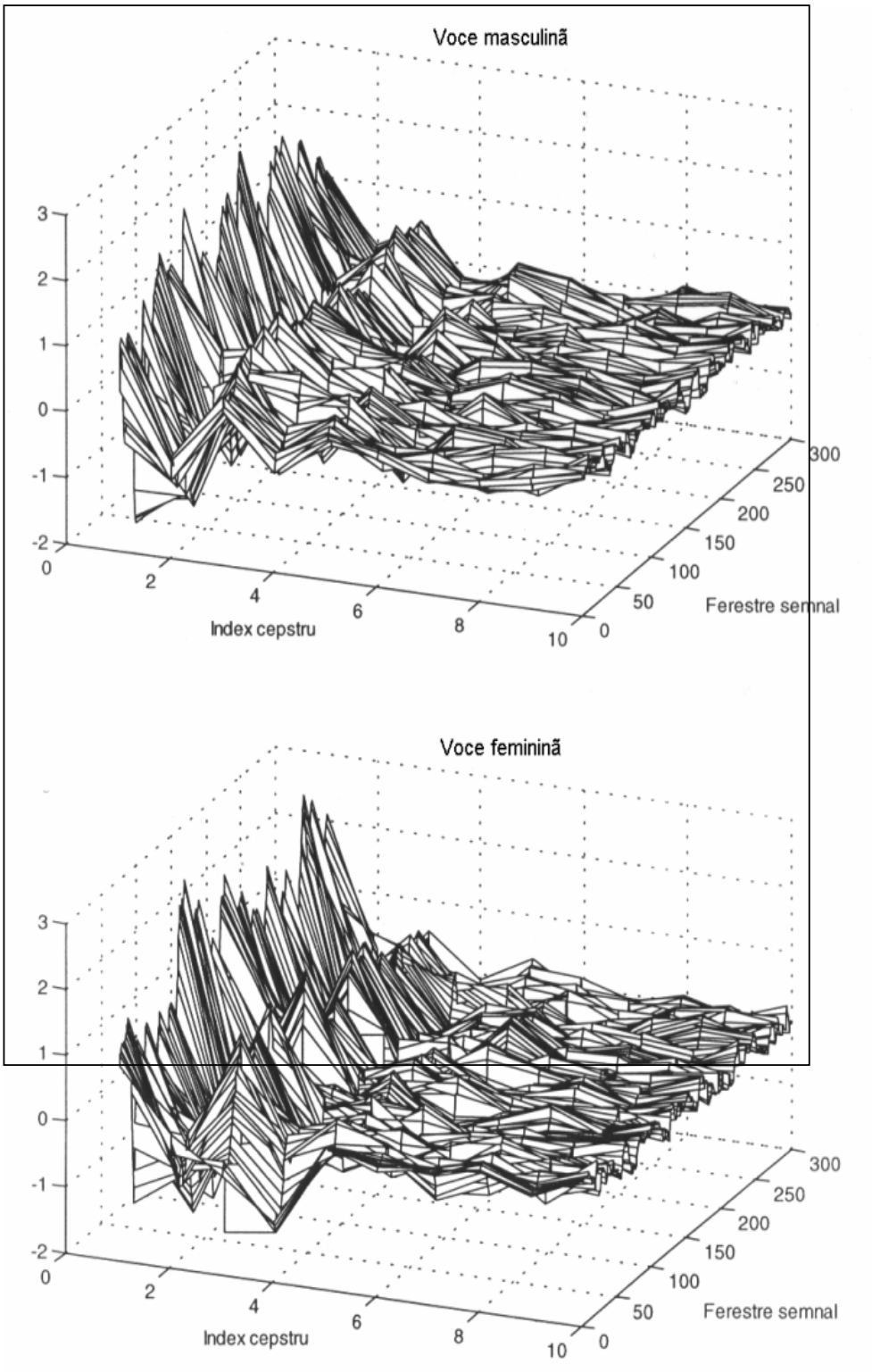
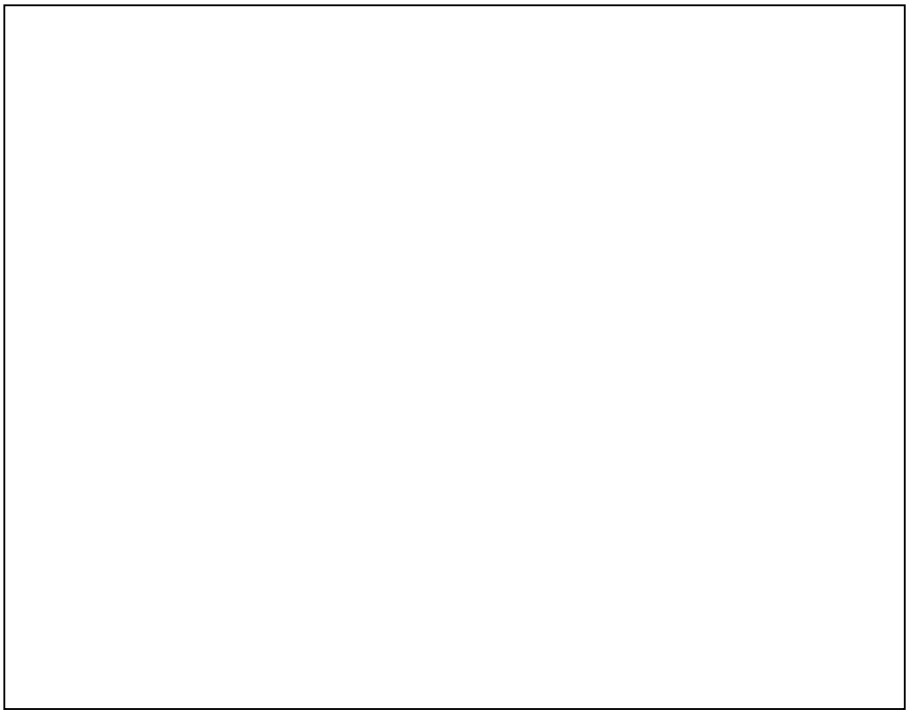
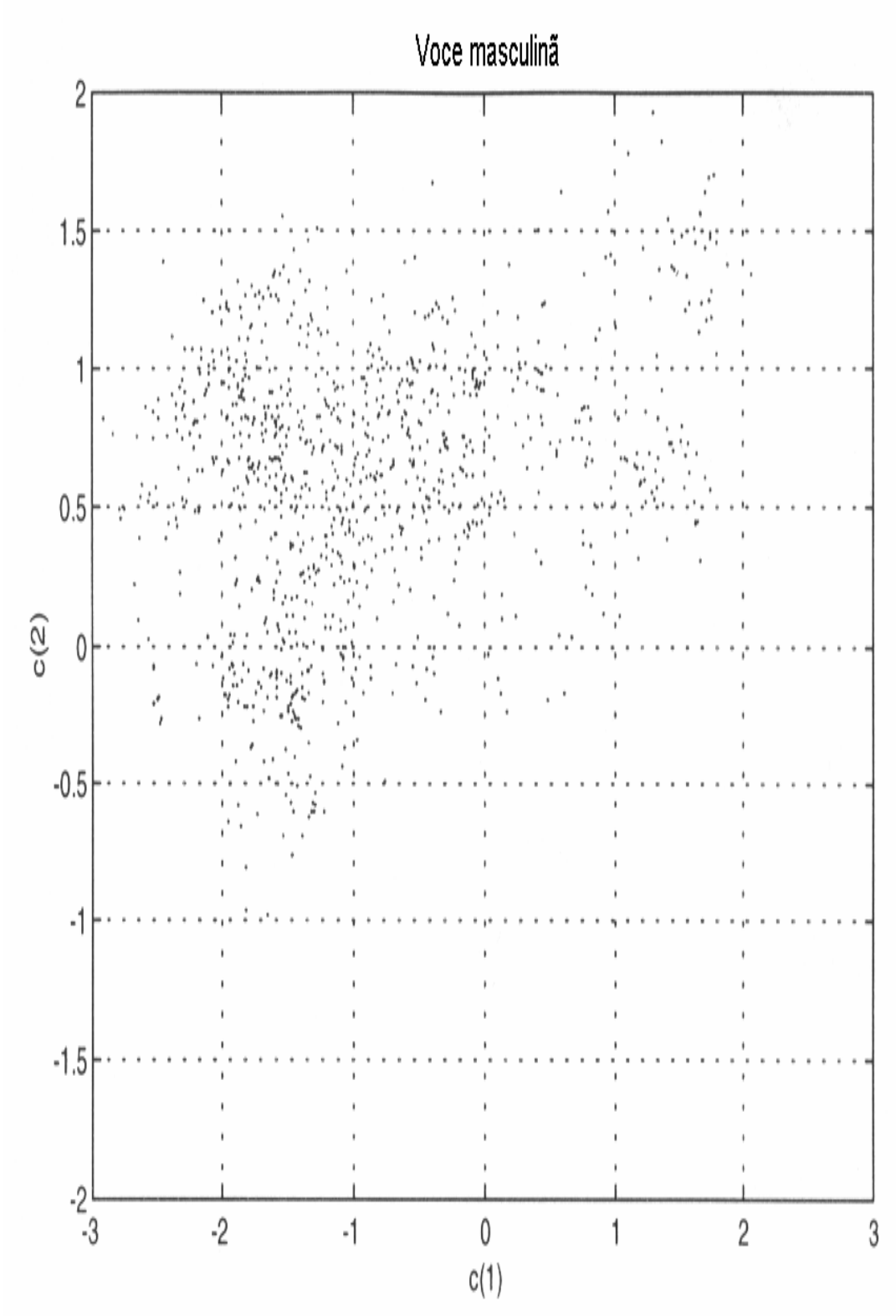

Voce masculină
3
2
1
0
-1
-2
0
2
4
6
8
10
Index cepstru
0
50
100
150
200
250
300
Ferestre semnal
Voce feminină
3
2
1
0
-1
-2
0
2
4
6
8
10
Index cepstru
0
50
100
150
200
250
300
Ferestre semnal

Figura 2. Evoluția în timp a coeficienților cepstrali ai semnalului vocal

Materialul vocal a fost achiziționat folosind un microfon de calitate (considerat fără zgomot) şi a fost eşantionat cu frecvența de 8 kHz. Intervalele de analiză au lungimea de 240 ms, cu o suprapunere de 160 ms. Analiza prin predicție liniară s-a efectuat cu ordinul de predicție $p = 10$, iar pentru estimarea coeficienților de predicție liniară s-a folosit algoritmul Levinson-Durbin. O primă observație este aceea că modulul amplitudinii coeficienților este descrescător cu ordinul acestora. Pentru coeficienții de ordinul 5-10, evoluția coeficienților cepstrali tinde să devină uniformă. Amplitudinea redusă a acestora anunță existența unor dificultăți de estimare în condiții de zgomot.

În scopul unei aprecieri calitative, fig. 3 prezintă distribuția coeficienților cepstrali în planul c(1)-c(2), pentru aceiaşi doi vorbitori (masculin şi feminin). Se poate observa distribuția diferită a principalilor coeficienți cepstrali pentru cei doi vorbitori. Se remarcă o concentrare a coeficienților în anumite zone ale planlui c(1)-c(2).




Voce masculinã
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
c(2)
-3
-2
-1
0
1
2
3
c(1)

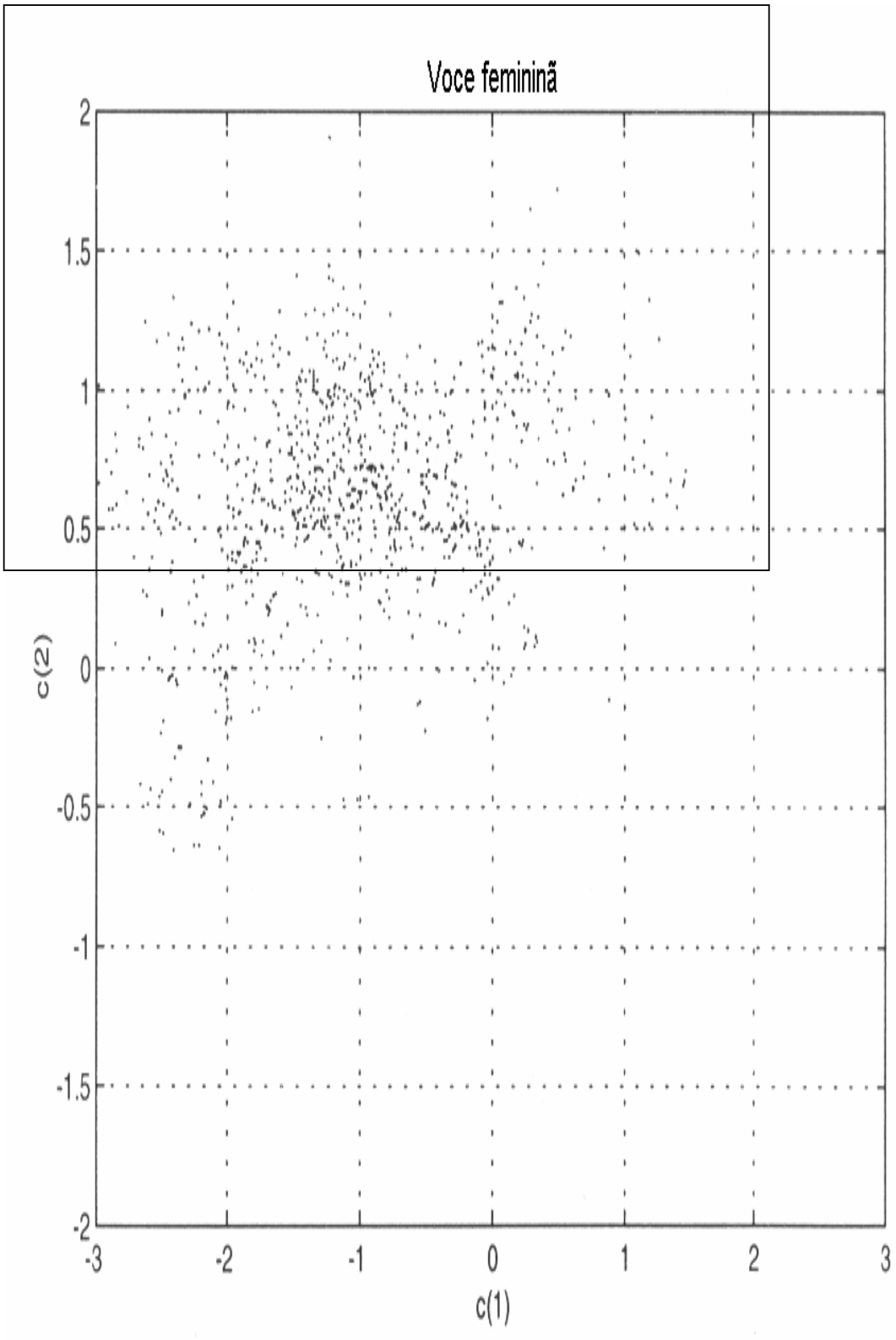

Voce femininã
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
c(2)
-3
-2
-1
0
1
2
3
c(1)

Figura 3. Reprezentarea coeficienților cepstrali în planul c(1) – c(2)

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 vocale şi parametrii cepstrali corespunzători

3. Cuantizarea vectorială

Din punctul de vedere al sistemelor de recunoaştere a vorbitorului, o persoană produce în timpul vorbirii o secvenţă de vectori de parametri. Aceştia caracterizează atât vorbitorul cât şi cuvintele pronunţate. Pentru un interval de timp suficient de lung, ne aşteptăm ca datele achiziţionate să acopere spaţiul vectorial într-un mod care depinde mai mult de caracteristicile vorbitorului şi mai puţin de ceea ce a pronunţat. Se face pesupunerea că, având la dispoziţie un volum suficient de date, se poate genera un model al vorbitorului care să fie utilizat într-un proces de recunoaştere [20, 21].

Principiul cuantizării vectoriale este aplicat în sensul compresiei unui volum mare de vectori acustico-fonetici, reprezentând material vocal pronunţat de către un vorbitor, într-un set restrâns de vectori denumit *tabelă de codu*ri (sau de *centroizi*). În etapa de antrenare, partiţionarea spaţiului acoperit de vectorii spectrali este făcută astfel încât media distanţelor minime dintre fiecare vector cepstral şi cel mai apropiat centroid să fie minimizată. În etapa de testare, un set de vectori provenind de la un vorbitor necunoscut, este codat utilizând tabela de vectori corespunzătoare vorbitorului vizat. Distorsiunea totală medie este utilizată în decizia de recunoaştere [22].

Fie $\{X_n\}$ ansamblul de N versiuni cunoscute ale vectorului X.

Fie $\{G_k\}$ o partiţie a acestui ansamblu în K clase; o clasă G_k cuprinde g_k elemente, astfel ca

$$\sum_{k=1}^{K} g_k = N \tag{12}$$

Notăm cu $X_P^{(k)}$ cuvântul prototip (“centroid”, “vector-cod”) al unei clase G_k

Distanţa medie între centroizi este

$$\overline{\sigma_P} = \frac{1}{K \cdot (K-1)} \sum_{i,j=1}^{K} D\left(X_P^{(i)}, X_P^{(j)}\right) \tag{13}$$

Distanţa medie între vectorii dintr-o aceeaşi clasă, parcurgând toate clasele este

$$\overline{\sigma_G} = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{g_k \cdot (g_k - 1)} \sum_{i,j=1}^{g_k} D\left(X_i^{(k)}, X_j^{(k)}\right) \tag{14}$$

Raportul $\frac{\overline{\sigma_P}}{\overline{\sigma_G}}$ reprezintă calitatea partiţiei

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:

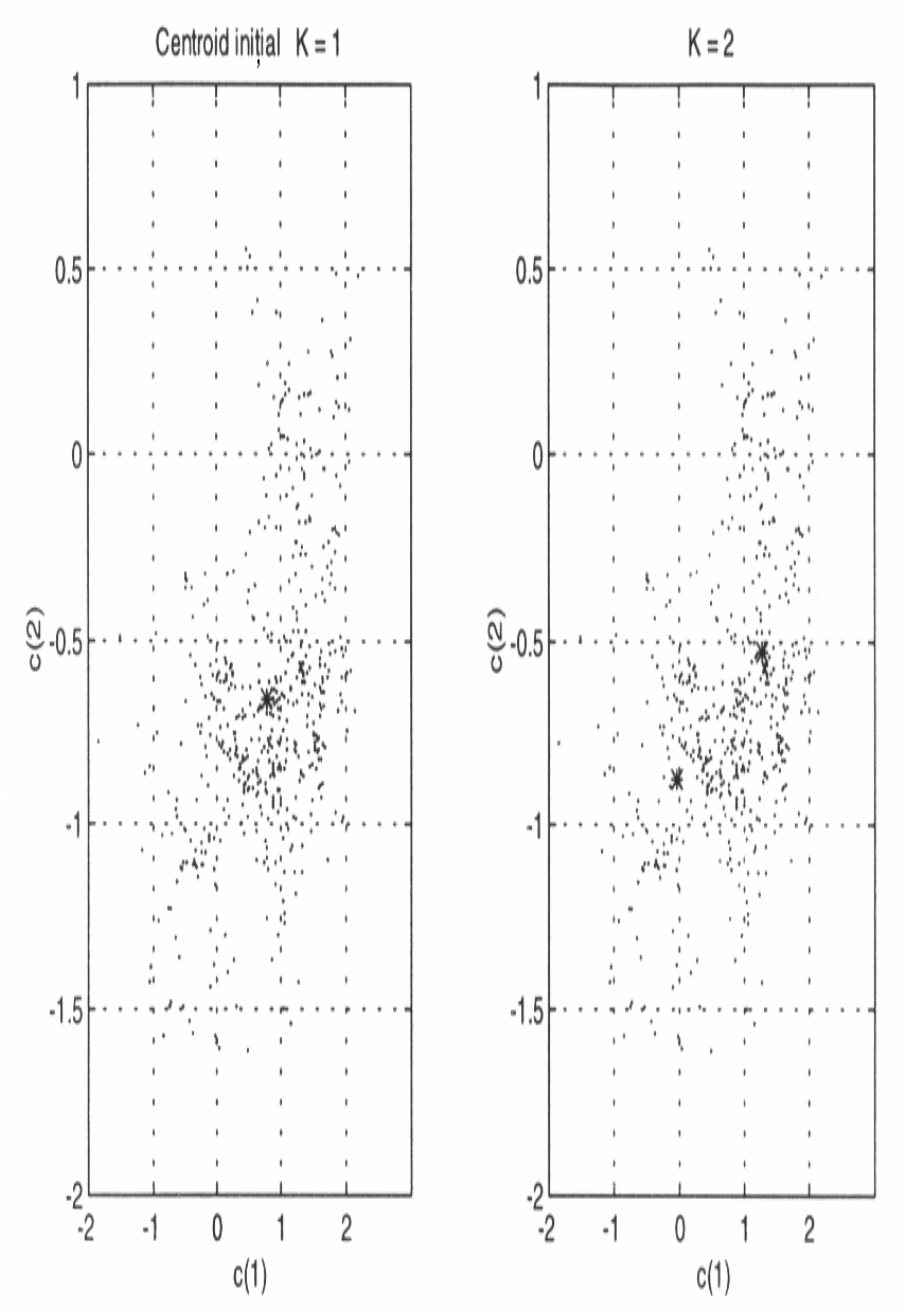
$$X_n \in G_k \quad dacă \quad D\left(X_n, X_P^{(k)}\right) < D\left(X_n, X_P^{(i)}\right), \quad \forall i \neq k \tag{15}$$

• se iterează găsirea centroizilor căutând în fiecare clasă k vectorul $X_n^{(k)}$ care are distanța față de vectorul cel mai depărtat al clasei minimă:

$$X_n^{(k)} \equiv X_P^{(k)} \quad dacă \quad \max_m D\left(X_n^{(k)}, X_m^{(k)}\right) \quad e\ minimă \tag{16}$$

• 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 crește.


Centroid inițial K = 1
K = 2
c(2)
c(1)
1
0.5
0
-0.5
-1
-1.5
-2
-2
-1
0
1
2

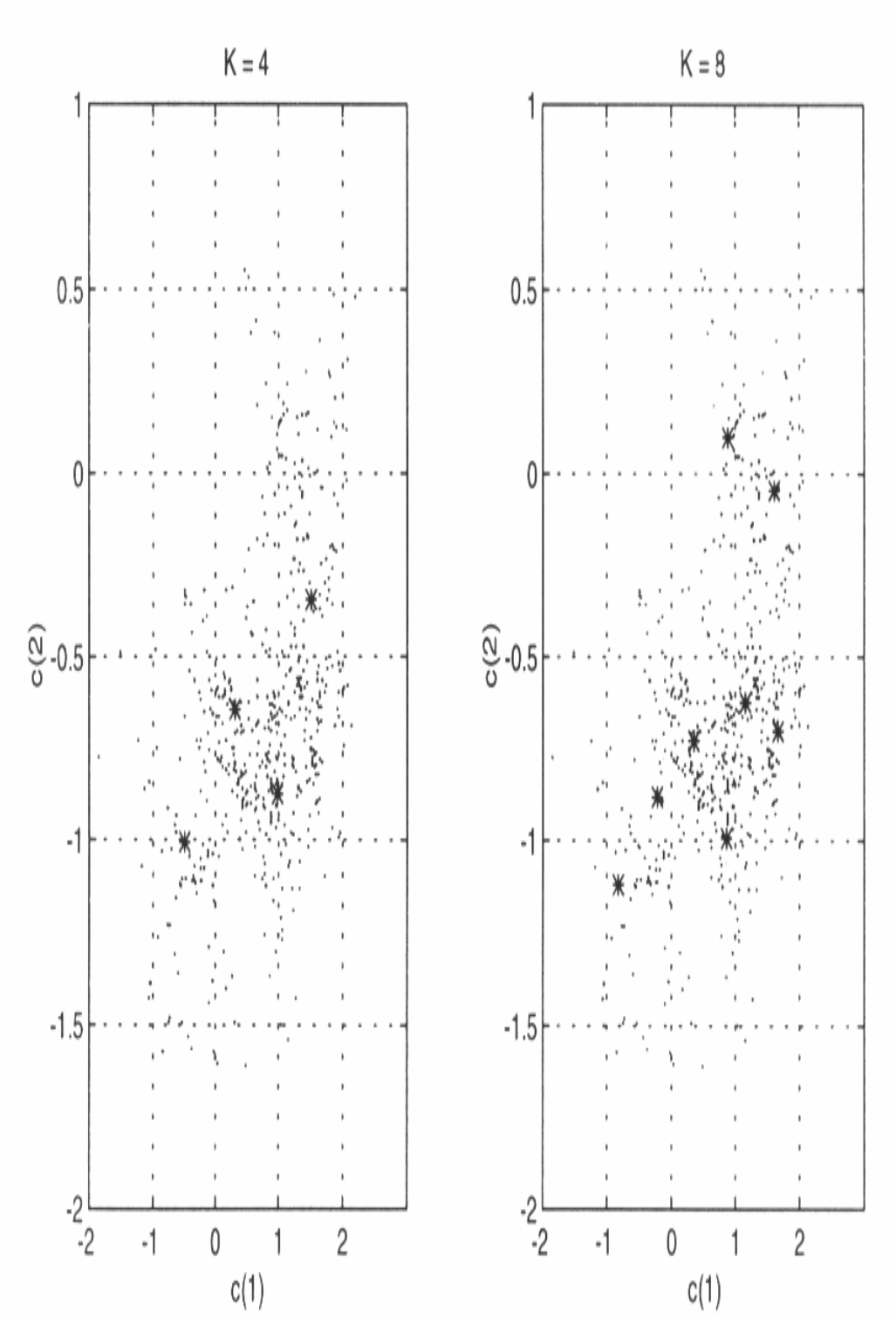

K = 4
K = 8
c(2)
c(1)
1
0.5
0
-0.5
-1
-1.5
-2
-2
-1
0
1
2

Figura 5. Evoluția algoritmului de cuantizare vectorială (8 centrozi)
"•" – vectori cepstrali; ∗ – centroizi

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 numarul 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.

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 același material vocal ca și 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 și 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 și 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 și 3 femei) cu vârsta ce variază de la 21 la 50 de ani. Fiecare vorbitor în parte a rostit 31 de fraze: 11 fraze echilibrate din punct de vedere fonetic au fost folosite pentru antrenare și 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 același 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 și mai apropiate de o situație reală de recunoaștere 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) și 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 (și implicit dimensiunea vectorilor cepstrali) este $P = 10$ iar dimensiunea tabelei de centroizi, $M = 64$. Ca distanță vectorială s-a folosit distanța euclidiană ponderată.

$$d(v_a, v_b) = \frac{1}{s_j^2} \sum_{j=i}^{p} (v_{aj} - v_{bj})^2 \tag{17}$$

unde s_j^2 este varianța componentei j calculată pe întreg setul vectorilor de antrenare. Ca metodă de cuantizare vectorială s-a folosit algoritmul LBG modificat.

Sunt evidente tendinţele 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 funcţie de aplicaţia concretă, pragul de decizie se poate stabili *a posteriori* la o altă valoare, adecvată scopului propus. Spre exemplu, dacă se doreşte limitarea acceptării impostorilor la 2%, respingerea adevăraţilor utilizatori va fi de 19.7%. Reciproc, pentru o eroare de respingere a utilizatorilor reali de 2%, acceptarea impostorilor va fi de 12.9%.

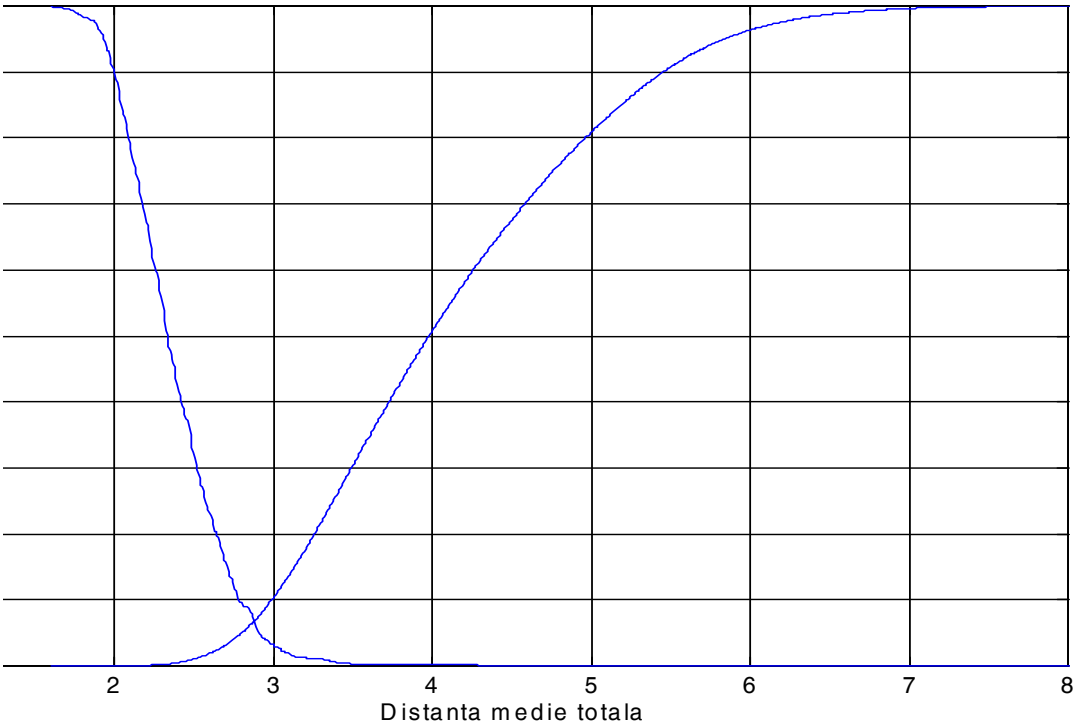


Figura 6. Eroarea de verificare a unui sistem de recunoaştere a vorbitorului utilizând cuantizarea vectorială

În experimentele de identificare a vorbitorului, fiecare frază de test provenind de la un vorbitor considerat necunoscut este comparată cu referintele 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 diferiți, 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 frecvenței fundamentale în recunoaşterea vorbitorului

Frecvența fundamentală poate fi utilizată ca parametru discriminator suplimentar în conjuncție cu algoritmi de cuantizare vectorială a vectorilor cepstrali.

Frecvența fundamentală F_0 sau *perioada fundamentală* T_0 (cunoscută şi sub numele de "pitch"), constituie un parametru important al vocii umane, care îşi găseşte utilizări practice în multe domenii ale procesării vorbirii. Încercări de utilizare a frecvenței fundamentale în procesul de recunoaştere a vorbitorului se cunosc încă de la începutul anilor '70, aceasta fiind pusă în corespondență directă cu prozodia. Majoritatea acestor experimente s-au desfăşurat utilizând sisteme de recunoaştere dependente de text şi metode de aliniere temporală. Sistemele de recunoaştere a vorbitorului independente de text bazate exclusiv pe frecvența fundamentală nu au dat rezultate satisfăcătoare.

Ideea prezentată în secțiunea de față este aceea de a folosi frecvența fundamentală ca *un parametru discriminator suplimentar*, în conjuncție cu algoritmi de cuantizare vectorială a vectorilor cepstrali [31]. Justificarea teoretică a acestei abordări rezidă în primul rând în modelul predicției liniare aplicat semnalului vocal, care presupune, aşa cum am arătat o separare clară între sursa de semnal şi tractul vocal. De asemenea, am arătat în secțiunea 2 că analiza cepstrală folosită pentru extragerea vectorilor cepstrali este un proces de deconvoluție, coeficienții cepstrali obținuți caracterizând în mod exclusiv tractul vocal. Ca atare, utilizarea ca date de intrare în acelaşi sistem atât a vectorilor cepstrali cât şi a frecvenței fundamentale nu reprezintă o abordare redundantă.

Cerințele de bază ale unui algoritm de extragere a frecvenței fundamentale sunt: acuratețea de estimare (evitarea armonicilor), robustețea deciziei sonor/nesonor, insenzitivitatea la zgomot, volumul de calcule minim. Se cunosc numeroşi algoritmi de estimare a frecvenței fundamentale (AMDF, Dubnowski, Rabiner, SIFT, etc.), fiecare prezentând avantaje şi dezavantaje. Trebuie arătat faptul că, din cauza, în principal, comportării nestaționare a semnalului vocal, niciunul din algoritmii cunoscuți nu este considerat perfect. Cu alte cuvinte, se acceptă ideea existenței erorilor atât în luarea decizei sonor/nesonor cât şi în obținerea valorilor propriu-zise ale frecvenței fundamentale. În experimentele prezentate mai jos s-a folosit algoritmul Rabiner, considerat ca fiind unul dintre cele mai robuste.

Ideea introdusă este aceea de a utiliza frecvența fundamentală în scopul unei clasificări grosiere a potențialilor candidați, atât pentru sarcina de verificare a vorbitorului, cât şi pentru cea de identificare. În cazul verificării, scopul propus este acela de a reduce eroarea de tip II, prin eliminarea vorbitorilor a căror frecvență fundamentală nu "corespunde" cu cea a vorbitorului de referință. În cazul sarcinii 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 putin 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

$$R_{P,\,medie} = \frac{media\ F_0\ pentru\ antrenare}{media\ F_0\ pentru\ test} \tag{18}$$

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).

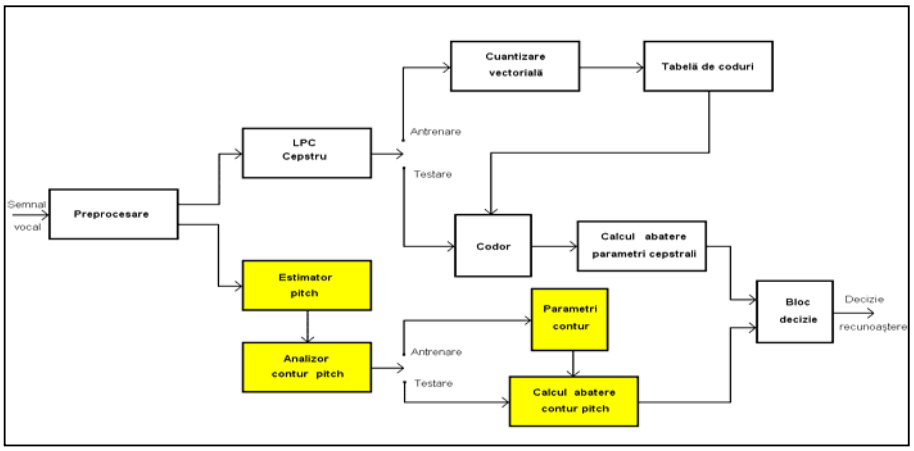


Figura 7. O variantă a sistemului de recunoaştere a vorbitorului – schema de principiu

Modul de discriminare a vorbitorilor este următorul: fixându-se un prag ε, dacă

$$R_{P,\,medie} \notin \left[1-\varepsilon, 1+\varepsilon\right] \tag{19}$$

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.

Utilizând elementul de discriminare descris mai sus s-au obţinut îmbunătăţiri importante atât în procesul de verificare a vorbitorului cât şi în cel de identificare. Rezultatele obţinute pentru 14 coeficienţi cepstrali şi 128 centroizi sunt prezentate în tabelul 1.

Tabelul 1

ε	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
Candidaþi identificare (%)	100	57.2	49.1	43.4	32.3	26.5

Cele mai bune rezultate s-au obţinut pentru $\varepsilon = 0.15$, caz în care eroarea de verificare obţinută este de aproape 2.5 ori mai mică decât în cazul folosirii doar a clasificării vectoriale. În cazul identificării, deşi îmbunătăţirile de acurateţe nu sunt impresionante, cel mai important rezultat îl reprezintă reducerea numărului candidaţilor, cu peste 65%. Pentru valori ale lui ε mai mici decât 0.10, se observă o degradare abruptă a performanţelor de verificare şi identificare, ceea ce indică faptul ca variaţia intra-vorbitor a frecvenţei fundamentale medii este mai mare decât acest prag.

6. Concluzii

Lucrarea de faţă 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 opţiunea de a aborda proiectarea ţinând 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 esenţiale ale procedurilor de recunoaşterea vorbitorului: achiziţia semnalului vocal, prelucrarea acustico-fonetică, recunoaşterea propriu-zisă.

Am subliniat importanţa parametrizării semnalului vocal. Analiza cepstrală care a fost aleasă pentru reprezentarea parametrică a semnalului vocal este legată de opţiunile fundamentale de analiză: separarea efectelor sursei de semnal şi ale tractului, separarea efectelor diverselor porţiuni din tractul vocal, analiza "în timp scurt"

Am utilizat cuantizarea vectorială ca metodă de recunoaºtere. Sunt prezentate o parte dintre rezultatele experimentelor realizate. Subliniem importanþa utilizării unor baze de date specifice şi, în consecinţă, am acordat spaţiu prezentării acestora.

O contribuţie pe care o considerăm interesantă la îmbunătăţirea performanţelor recunoaşterii vorbitorului o constituie utilizarea frecvenţei fundamentale ca parametru discriminator grosier. Sunt prezentate o serie de rezultate care probează în ce mod anumite performanţe sunt superioare abordării "clasice".

O parte dintre rezultatele obţinute sunt susceptibile de a fi generalizate pentru recunoaşterea vorbirii în limba română [32] (de pildă, coeficienţii cepstrali pentru foneme ale limbii române). De asemenea, utilizarea frecvenţei fundamentale apropie recunoaşterea vorbitorului de o anumită dependenţă de limba în care sunt rostite frazele de antrenare şi de test.

Referinţe bibliografice

[1] M.Dragănescu, C.Burileanu, coordonatori (1986). Analiza şi sinteza semnalului vocal – Editura Academiei Române, Bucureşti.

[2] M.Dragănescu, G.Ştefan, C.Burileanu (1991). Electronica funcţională – vol. I, 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, Vol.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 Automatic Speaker Verification – ESCA Workshop on Speaker Recognition, Identification and Verification, p. 95-103.

[6] M. Hanah s.a. (1994). The Role of the Reference Template in Speaker Verification – ESCA Workshop on Speaker Recognition, Identification and 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, Vol. 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 Speaker 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 Features 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 Features 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 Features 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, vol. 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, Romania, 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 Features – 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, Vol. 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 îndoială, *vorbirea* este una din principalele sale componente; ea este cea mai veche modalitate de comunicare între oameni şi este şi astăzi cea mai raspândită. 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 sa 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, apariția unor algoritmi îmbunătățiți pentru prelucrare de semnal, iar în domeniul comunicațiilor

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; totusi, 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 conversia text – voce, care constituie obiectul principal al lucrării de faţă, este astăzi un obiectiv de bază al domeniului prelucrării vorbirii şi subiect al unor cercetări intense. Un *sistem de sinteză pornind de la text* (TTS – *"Text-to-Speech"*) poate oferi o gamă variată de aplicaţii, de la accesul la poşta electronică şi diferite tipuri de baze de date, la pronunţarea unui text pentru persoane cu handicap vizual.

Este important de observat faptul că tehnologia de răspuns prin voce prezintă o serie de avantaje fundamentale pentru transmiterea informației:

- oricine poate înțelege un mesaj, fără antrenare sau concentrare deosebită;
- mesajul poate fi recepționat chiar dacă cel ce ascultă este implicat în alte activități, cum ar fi mersul, manipularea unor obiecte, sau citirea altor informații;
- rețeaua telefonică convențională poate fi utilizată pentru accesul rapid la distanță la o bază de informații;
- această formă de comunicare este mai economică decât cea tradițională prin mesaje scrise.

Toți aceşti factori precum şi numeroasele aplicații cerute de industrie au creat premisele unor cercetări aprofundate, obținându-se astfel în multe țări sisteme comerciale care pot produce vorbire sintetică pornind de la text, cu o inteligibilitate acceptabilă.

Într-adevăr, scopul principal al celor mai multe sisteme de sinteză existente este de a produce o vorbire *inteligibilă*. Din acest punct de vedere, sinteza pare a fi de mai multă vreme o tehnologie "stabilă", ieftină şi uşor de implementat; se spune chiar, uneori, că acest domeniu este în prezent suficient de bine dezvoltat, iar problemele rămase sunt minore din punct de vedere ştiințific. Dacă însă scopul este sinteza în timp real, pornind de la un vocabular nelimitat de cuvinte şi fără restricții asupra textului, iar vorbirea să fie nu numai inteligibilă, ci şi la fel de *naturală* ca cea umană, atunci se constată că performanțele actuale sunt departe de a fi satisfăcătoare. Rămân încă multe probleme importante de rezolvat: extinderea vocabularului oferit, înlăturarea restricțiilor impuse textului în privința unor caractere speciale, îmbunătățirea caracteristicilor de *prozodie,* posibilitatea de modificare a *ritmului* şi *stilului* vorbirii sintetizate, sau elaborarea unor sisteme de sinteză în mai multe limbi. Aceste sarcini se dovedesc a fi deosebit de dificile şi cer, evident, eforturi interdisciplinare susținute [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 a vorbirii care realizează conversia finală în semnal vocal, pornind de obicei de la o reprezentare parametrică a unor segmente acustice fundamentale".

3. Sinteza vorbirii pornind de la text

Pentru a înțelege mai bine dificultatea sarcinii unui sistem de sinteză pornind de la text, considerăm că este util să punem în evidență mai întâi modul (fiziologic) în care o persoană citeşte cu voce tare un text. Imaginea textului este sesizată de neuronii sistemului vizual, transmisă creierului sub forma unor stimuli electrici, aici fiind prelucrată pentru a putea permite comanda neuronilor responsabili de corecta activare a plămânilor, coardelor vocale şi organelor articulatorii. În acest fel se produce vorbirea, ea fiind permanent

monitorizată de creier (în special prin intermediul organelor auditive), în scopul ajustării configurației tractului vocal în timp real.

Desigur, cunoaştem încă prea puțin despre organizarea de ansamblu a sistemului nervos uman, care este capabil de această activitate complexă; putem propune totuşi următorul *model funcțional* prin care este prelucrată informația optică şi apoi este dată comanda de generare a vorbirii:

- Atunci când citim un text, efectuăm practic o sarcină de *recunoaştere de caractere*, ignorând, parțial inconştient, anumite erori de redactare a cuvintelor (caractere lipsă sau înlocuite cu altele) şi decodificând mai degrabă cuvântul ca un întreg; are loc un proces de inferență a informației dintr-un context posibil incomplet. De asemenea, recunoaştem cu uşurință caractere speciale sau abrevieri.
- Considerând *fonemele* ca fiind cele mai mici elemente sonore care permit diferențierea între ele a cuvintelor, este evident că secvența fonemică corespunzătoare unui cuvânt diferă de şirul de caractere grafice din care este compus cuvântul; creierul trebuie sa facă prin urmare o *transcriere fonetică pornind de la litere*, această operație practic instinctivă permițând pronunția unui număr nelimitat de cuvinte.
- În cele mai multe situații, suntem capabili să începem pronunția unei fraze mult înainte de terminarea ei; cu alte cuvinte, putem face o *stucturare sintactică*, descompunând fiecare propoziție în grupuri de cuvinte şi asociindu-le intonația corespunzătoare. Şi acest proces este practic inconştient, fiind bazat pe educație şi experiență.
- În sfârşit, putem discrimina cu uşurință cuvinte ce se scriu asemănător dar se pronunță diferit, după *înțelesul semantic*, fapt posibil datorită aceleiaşi capacități de deducție a creierului de care am vorbit mai sus.

Concluzia este simplă: pe baza experienței lingvistice căpătate în urma educației, o persoană familiară cu limbajul în care este scris un text depăşeste 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 la nivelul plămânilor, tensiunea glotală, ca şi configurațiile tractului vocal şi nazal, evoluează în timp. Toate acestea sunt 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.

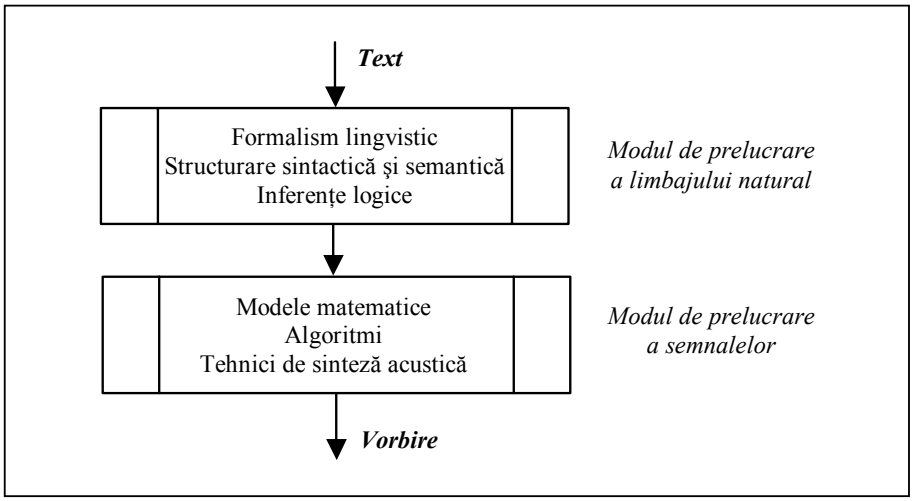


Figura 1. Diagramă funcțională pentru un sistem TTS

Ca şi pentru un cititor uman, schema cuprinde un *modul de prelucrare a limbajului natural*, capabil să producă o transcriere fonetică a textului citit, împreună cu informații despre intonație, accente, durate şi de asemenea un *modul de prelucrare a semnalelor*, care transformă informația simbolică primită în vorbire sintetică, pe baza unor tehnici de sinteză adecvate şi a unor structuri stocate în urma unei analize preliminare. Etapele de bază ale sintezei pornind de la text pot fi astfel descrise printr-un număr de transformări succesive ce trebuie aplicate asupra şirului de caractere ce reprezintă textul de intrare; scopul este de a se obține o vorbire **de calitate**, într-o limbă oarecare, fără constrângeri asupra textului introdus.

Trebuie menționat faptul că formalismul descris poate "sări" uneori peste anumiți pași, dacă se utilizează în mod adecvat cunoaşterea lingvistică şi matematică; acest lucru se întâmplă atunci când punem anumite restricții asupra textului ce trebuie pronunțat, sau impunem vorbirii sintetizate o inteligibilitate şi o naturalețe moderate. Cu alte cuvinte, proiectarea sistemului TTS se poate simplifica dacă se impun sistemului sarcini precise, corespunzătoare unor aplicații concrete.

Colectivul nostru de cercetare a început acum câțiva ani dezvoltarea unui sistem complet TTS în limba română, bazat pe *concatenare de difoneme*. Arhitectura acestui sistem este prezentată în Figura 2. Sistemul cuprinde o parte importantă de prelucrare lingvistică şi un modul de generare a semnalului de vorbire având la bază un algoritm de tip PSOLA [4]. După realizarea unei prime variante a sistemului, se depun în continuare eforturi pentru creşterea naturaleții vorbirii sintetizate, prin îmbunătățirea performanțelor la diferite nivele de prelucrare.

Modulul de prelucrare a limbajului într-un sistem TTS are ca sarcină transformarea textului de intrare într-o reprezentare fonetică şi prozodică, care trebuie să descrie cât mai fidel posibil pronunția sa. Acest lucru poate fi realizat parcurgând mai multe etape succesive, puse în evidență şi în figura anterioară. Vom discuta în cele ce urmează **modalitățile de rezolvare a părții de prelucrare inițială (preprocesare) a textului în cadrul sistemului nostru de sinteză în limba română.**

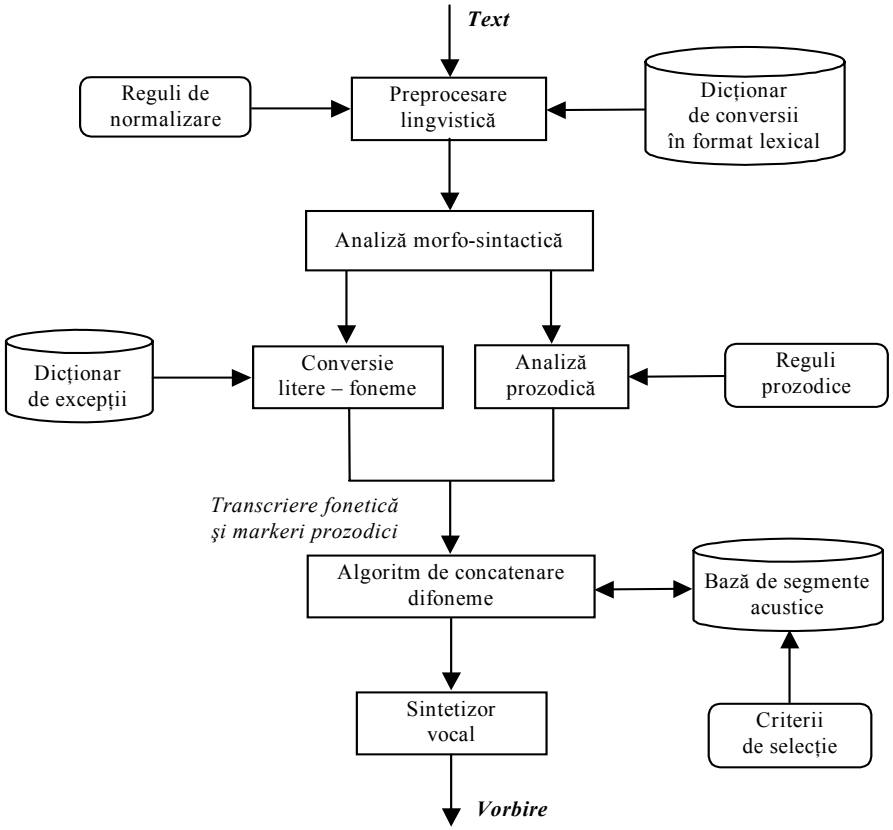


Figura 2. Arhitectura sistemului TTS în limba română

4. Preprocesarea textului de intrare în cadrul sistemului TTS în limba română

4.1 Probleme generale

Una dintre dificultățile majore ale sistemelor TTS constă în faptul că aceste sisteme trebuie să poată prelucra practic **orice text**, plecând de la propoziții simple izolate

şi mergând până la paragrafe complexe, care pot cuprinde un număr mare de propoziţii, cu posibile structuri negramaticale şi simboluri speciale. Ca atare, partea de preprocesare lingvistică a textului are un rol extrem de important, deoarece detectarea corectă şi interpretarea şirurilor de caractere de intrare influenţează acurateţea întregului sistem de sinteză şi contribuie la conversia unui text fără restricţii în vorbire sintetică.

Uzual, un text scris se prezintă sub forma unei secvenţe de caractere ASCII; el este alcătuit din cuvinte compuse cu ajutorul literelor alfabetului, dar şi din alte tipuri de caractere: spaţii albe, semne de punctuaţie, şiruri de numere, sau alte simboluri speciale (de exemplu operatori matematici). Textul poate conţine numerale (*12*, *12.450*, *1,245*), abrevieri (*prof.*, *dr*, *ing.*), sau acronime (*IBM*, *S.R.L.*, *TTS*). Aceste secvenţe sunt de obicei "anormale" din punct de vedere lingvistic faţă de majoritatea cuvintelor din text şi trebuie mai întâi transformate într-un format ce poate fi recunoscut de partea de analiză lingvistică. Această sarcină revine modulului de preprocesare, care trebuie de asemenea să realizeze o segmentare a textului de intrare (detectarea cuvintelor şi a sfârşitului frazelor) şi o prelucrare a semnelor de punctuaţie şi a simbolurilor speciale [5, 6, 7, 8].

La prima vedere, preprocesarea unui text pentru un sistem TTS poate părea banală; în realitate însă, lucrurile sunt destul de complicate. Spre exemplu, nu este totdeauna posibilă determinarea marginilor unei fraze pe baza semnelor de punctuaţie. Astfel, punctul (.) poate apare şi la sfârşitul unei fraze, dar şi în multe alte situaţii, ca de exemplu în abrevieri (*ing.*), acronime (*S.R.L.*), indicaţia că se omite un anumit fragment de text (...), sau numerale (*12.450 – douăsprezece mii patru sute cincizeci*), situaţii care trebuie diferenţiate prin procedee adecvate [9]. De asemenea, cratima creează dificultăţi în operaţia de segmentare; ea poate fi folosită pentru despărţirea în silabe, pentru scrierea cuvintelor compuse, pentru delimitarea unui nou paragraf, sau în enumerări.

O sarcină dificilă este şi conversia anumitor secvenţe de simboluri în cuvinte care să poată fi analizate lingvistic. Dacă unele abrevieri uzuale pot fi "expandate" imediat, cu ajutorul unui tabel de echivalenţe, există multe situaţii în care secvenţe de simboluri care nu se pot distinge pe baza ortografierii lor, cer tipuri diferite de conversii; de exemplu, numărul format din şapte cifre *6123456* poate reprezenta un număr întreg sau un număr de telefon şi va trebui citit diferit în cele două situaţii. În general, prezenţa şirurilor de numere în text ridică numeroase dificultăţi, deoarece ele pot apare în diferite contexte: ore, date, numere de telefon, expresii aritmetice etc.

Trebuie observat că aceste ambiguităţi create de natura multifuncţională a semnelor de punctuaţie sau de modul diferit de citire a aceloraşi secvenţe de simboluri, pot avea implicaţii majore asupra acurateţii întregului proces de prelucrare lingvistică şi în final asupra pronunţiei corecte a textului de către sistemul de sinteză.

Evident, numărul secvenţelor 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 puţine situaţii dificile decât comentariile economice, sportive, sau prezentările de spectacole. În ultimele situaţii menţionate, construcţiile 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, *I* – unu etc.

b2. Abrevieri scrise cu minuscule şi puncte (*tel.* – telefon, *a.c.* – anul 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: . ? ! : ; ... , - / ' " ( ) [ ] { }

b8. Simboluri speciale

b8.1. Simboluri matematice uzuale: + – ∗ (sau ×) : (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.* (**id**em), *etc.* (**etc**etera), *tel.* (**tel**efon), *a.c.* (**a**nul **c**urent), *a.m.* (**a**nte **m**eridian), *d.a.* (**d**după-**a**miaza), *P.S.* (**p**ost **s**criptum) – deci categorile **b2**, **b5.2** puse în evidență anterior.
- Dacă în abreviere **apare** litera finală a cuvântului, nu se pune punct după abreviere; exemple: *cca* (**c**ir**ca**), *dna* (**d**oam**na**), *dl* (**d**omnu**l**), *dnei* (**d**oam**nei**), *jr* (**j**unio**r**) – 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: *1.234* 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.
- 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 proporții (*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*.
- Dacă este înconjurat de numere izolate, se transcrie *pe*.
- Dacă este înconjurat de secvențe de cifre şi alte caractere matematice (*2×3/4×5*), 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 secvenţă de cifre, în funcţie 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 punctuaţie: *" ( ) [ ] { }*

Alte semne de punctuaţie ce pot fi utilizate în textele obişnuite sunt ghilimelele (sau semnele citării) şi parantezele rotunde; ele semnifică de obicei un citat, reprezintă porţiuni 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 consideraţii personale introduse în text. Apar de obicei în perechi şi vor fi declarate caractere extra-textuale, servind modulului de analiză prozodică pentru obţinerea 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 secvenţe de cifre şi se expandează de obicei prin utilizarea cuvintelor corespunzătoare semnificaţiei lor, cu ajutorul dicţionarului de conversii în format lexical.

8. Secvenţele de cifre

Secvenţe de cifre pot apare şi în texte obişnuite, dar mai ales în expresii matematice, împreună cu semne de punctuaţie sau simboluri matematice; evident, deoarece numărul lor posibil este practic infinit, ele trebuie expandate pe bază de regului de conversie, în funcţie de context.

Am propus anterior o serie de regului pentru cazurile cele mai frecvente (numere întregi sau zecimale, numerale ordinale, ore, date). O situaţie specială (pe care de asemenea am menţionat-o anterior), o reprezintă cazul în care o secvenţă de cifre, scrisă fără semne de punctuaţie, 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 prezenţa abrevierii *tel.*), această problemă rămâne în sarcina modului de analiză sintactică, care poate realiza o cercetare contextuală mai amplă.

9. Simbolurile matematice uzuale: + – * (sau ×) : (sau /) = < > % ~

Simbolurile matematice au o situaţie 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 dicţionarului de conversii în format lexical (de exemplu *plus*, *minus*, *înmulţit cu*, *împărţit la* etc.) Dacă totuşi în scrierea expresiei nu apar blancuri, contextul secvenţelor 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, cca, 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 acronimelor nu trebuie stocate în dicționar, deoarece pronunția lor nu necesită informații textuale suplimentare. De obicei, pronunția lor se reduce la citirea secvențială a caracterelor ce compun acronimul, individual (ca pentru *S.R.L.*), sau la citirea normală a cuvintelor, atunci când pronunția lor s-a generalizat în limbaj într-o formă compactă (*NATO*, *TAROM*); pentru citirea secvențială a acronimelor, este nesesar doar un set de reguli simple de transcriere a literelor rostite separat (de exemplu *S.R.L. – serele*).

Regula propusă pentru acronime este deci următoarea:

- Dacă secvența de caractere cuprinde cel puțin două majuscule şi nu este găsită în dicționarul de abrevieri, se caută în dicționarul de acronime:
  - dacă este găsită aici, secvența se expandează conform echivalenței din dicționar;
  - dacă nu este găsită în dicționarul de acronime şi nu cuprinde puncte, majusculele sunt (eventual) înlocuite cu minuscule şi secvența nu va suferi altă prelucrare (se va citi ca atare);
  - dacă nu este găsită în dicționar, dar cuprinde puncte, secvența este expandată secvențial, utilizând un set minim de reguli de transcriere a literelor rostite separat.

Pentru toate situațiile menționate, preprocesorul va semnala acronimul modulelor ulterioare, pentru o corectă analiză sintactică şi prozodică a textului.

5. Concluzii

Am discutat în această lucrare câteva aspecte fundamentale legate de sinteza automată a vorbirii, ca şi un număr important de reguli şi principiile generale pe baza cărora a fost proiectat preprocesorul de text pentru sistemul TTS în limba română. Nu am urmărit totuşi să descriem complet şi în detaliu funcţionarea şi implementarea acestuia; o serie de consideraţii suplimentare şi totodată modalitatea concretă de implementare (pentru o variantă preliminară) au fost prezentate de autor în [13] şi [14].

În varianta actuală, preprocesorul de text a fost îmbunătăţit pentru a rezolva unele situaţii dificile legate de abrevieri, numerale urmate de unităţi de măsură etc. De asemenea, un mecanism de automat de corecţie permite preprocesorului să fie "tolerant" cu anumite erori tipice de sintaxă, cum ar fi de exemplu fraze ce nu încep cu minuscule, sau un format "uşor" incorect pentru date 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 soluţiona **toate** ambiguităţile care se pot ivi, datorate în special numărului extrem de mare al abrevierilor, acronimelor – în general a secvenţelor neuzuale care pot apare într-un text scris. De asemenea, nu poate face faţă unor construcţii negramaticale (deşi, de exemplu, unele simboluri speciale neaşteptate sunt ignorate).

Desigur că un set mai mare de reguli şi un dicţionar de conversii în format lexical mai cuprinzător ar spori eficienţa preprocesorului, dar este posibil ca el să devină atât de complicat, încât să fie practic neoperaţional pentru un sistem TTS. Singura soluţie practică pentru tratarea cazurilor ambigui este folosirea unui set minim de reguli, păstrarea în dicţionar a celor mai uzuale situaţii (cu posibila adaptare a dicţionarului la **tipul** textului ce se citeşte) şi examinarea cazurilor rămase la un nivel superior, pe baza plauzibilităţii sintactice, semantice sau pragmatice a frazelor obţinute după preprocesare.

Referinţe 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 synthčse de la parole – în "Fondements et perspectives en traitement automatique de la parole" (H. Méloni – Coord.), Aupelf-Uref, pp. 109-120, 1996.

[3] D. Burileanu (1999). Contribuţii privind sinteza automată a vorbirii pornind de la text în limba română – Teză de doctorat. Universitatea "POLITEHNICA" Bucureºti, 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, vol. 2, pp. 913-916, 1993.

[7] M.Y. Liberman, K.W. Church (1992). Text Analysis and Word Pronunciation in Text-to-Speech Synthesis – în "Advances in Speech Signal Processing" (S. Furui, M. Sondhi – Coord.), Dekker, pp. 791-832, 1992.

[8] A. Lindstrom, M. Ljungqvist (1994). Text Processing within a Speech Synthesis System – Comunicare la "International Conference on Spoken language Processing", Yokohama, pp. 139-142, 1994.

[9] M. McAllister (1989). The Problems of Punctuation Ambiguity in Full Automatic Text-to-Speech Conversion – Comunicare la "EUROSPEECH'89", Paris, pp. 538-541, 1989.

[10] G. Beldescu (1984). Ortografia actuală a limbii române. Ed. Ştiinţifică şi Enciclopedică, Bucureşti, 1984.

[11] T. Dutoit (1997). An Introduction to Text-to-Speech Synthesis. Kluwer Academic Publishers, 1997.

[12] F. Şuteu, E. Şoşa (1993). Dicţionar Ortografic al Limbii Române. Ed. ATOS, Bucureşti, 1993.

[13] D. Burileanu (1999). Natural Language Processing for Speech Synthesis in Romanian Language –Comunicare la "The 12th International Conference on Control System and Computer Science", Bucureşti, vol. II, pp. 1-6, 1999.

[14] D. Burileanu, C. Dan, M. Sima, C. Burileanu (1999). A Parser-Based Text Preprocessor for Romanian Language TTS Synthesis – Comunicare la "EUROSPEECH'99", Budapesta, vol. 5, pp. 2063-2066, 1999.

Utilizarea tehnicilor nuantate (fuzzy) şi de dinamică neliniară pentru sinteza adaptivă a vorbirii

Horia-Nicolai L. TEODORESCU
Academia Română, Secţia Ştiinţa şi Tehnologia Informaţiei,
Calea Victoriei 125, Bucureşti
E-mail: hteodor@etc.tuiasi.ro

1. Introducere

În timp ce maşina realizează tipic transmisie de date, omul comunică. Diferenţa 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, varietăţii 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 dezideratelor mai sus menţionate. Realizarea unor maşini capabile să mimeze calităţile vocii umane şi să *dialogheze* cu oamenii, sau măcar să comunice într-o manieră similară în care omul o face, este un deziderat în numeroase domenii, de la dialogul om-calculator, la sistemele auto şi la sistemele de învăţare asistată de calculator [13-15]. Rezolvarea acestei probleme are implicaţii semnificative pentru acceptarea sintezei vocii într-o varietate de aplicaţii, 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 semanl vocal, chiar peste mai multe cuvinte sau propoziţii, implică ritm, accent, intonaţie, timbru, afect şi alte caracteristici ale vocii încă insuficient înţelese sau vag definite în literatură. Informaţia paralingvistică ce este conţinută de prozodie nu este nicăieri regăsită la nivelul "spus" prin cuvinte, dar – aşa cum am subliniat în [2] – această informaţie poate fi chiar mai importantă pentru ascultător decât informaţia lingvistică propriu-zisă. Incapacitatea sistemelor actuale de sinteză vocală de a reda prozodia naturală este evidenţiată chiar de marii producători de aplicaţii [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 sintezã automatã, atunci când acesta este recunoscut [2].

Sinteza varietală se referă la modificarile 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 de 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 vorbitor uman – se operează deci la un nivel relativ elementar, cu modificări de prozodie minimale.

Realizarea acestei adaptări este esențială în multe aplicații de sinteză a vocii, incluzând sinteza vocală pentru aplicații în medii industriale şi în mijloace de transport, sau sinteza vocală pentru proteze laringiene. Este remarcabil că această adaptare se poate realiza, la pretenții reduse, cu foarte puțin hard suplimentar şi/sau cu un soft minimal, aducând însă o îmbunătățire esențială în utilizare. În privința hardului, este necesar unul sau mai multe canale de culegere a semnalului de zgomot (semnal sonor ambiental).

Procesarea semnalului de zgomot, în vederea realizării controlului sistemului de sinteză automată, presupune determinarea puterii zgomotului ambiental într-o fereastră temporală şi determinarea componenței spectrale a semnalului ambiental. Primul parametru de caracterizare a zgomotului se obține ca medie aritmetică a pătratului semnalului *s*, într-o fereastră dată, de lărgime de *W* eşantioane şi caracterizată de momentul actual de timp, *n*:

$$P(t) = \sum_{k=0}^{W} s_{n-k}^2 \tag{1}$$

Caracterizarea spectrală se poate realiza sumar prin raportul *HL* dintre puterea la frecvențe "înalte" (frecvențele înalte corespunzând în mare benzii de frecvență ce include formanții nr. 2, 3, 4 şi 5 din spectrul vocal) şi puterea la frecvențele "joase" (până la aproximativ al doilea formant, deci până la frecvența de cca. 400 –500 Hz, ținând cont şi de vorbitorii feminini):

$$HL = \int_{0}^{500} S^2(\omega) \cdot d\omega \Big/ \int_{500}^{10000} S^2(\omega) \cdot d\omega \tag{2}$$

Deoarece parametrii respectivi sunt relaționați cu impactul pe care îl au asupra inteligibilității vorbirii, deci sunt dați de calități subiective, este natural să abordăm o definire probabilistă sau fuzzy a lor. Dată fiind simplitatea controlului nuanțat[93] (fuzzy), vom prefera a doua variantă. Un exemplu de definire[94] a funcțiilor de apartenență respective este prezentat în Figura 1. Este de presupus ca această definire să constituie doar un punct de plecare, îmbunătățirea calității sintezei realizându-se şi prin modificarea funcțiilor de apartenență.

[93] *Deşi nu este larg acceptat şi are o traducere mai dificilă în alte limbi, vom utiliza aici termenul "nuanțat", propus de Grigore C. Moisil, în locul englezescului "fuzzy".*

[94] *Pentru a nu încărca prezentarea, ecuațiile funcțiilor respective sunt date în Anexa 1.*

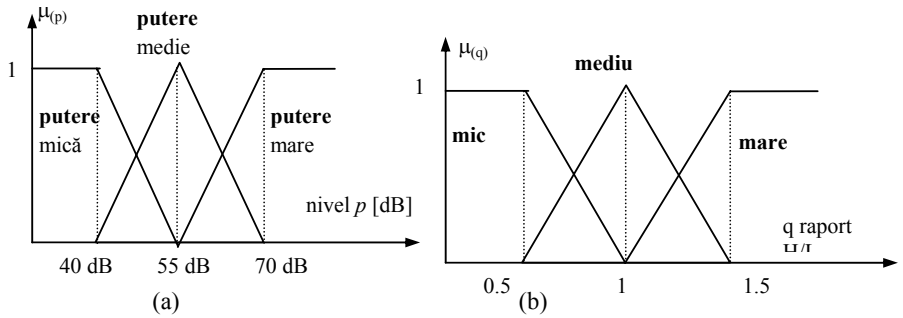


Figura 1. Funcțiile de apartenență ale premiselor regulilor folosite pentru determinarea modificărilor parametrilor de control ai sintetizorului

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 poate fi 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 nuanţate. Preferăm sistemele de tip Sugeno de ordin 0 (vezi Anexa 1), deoarece furnizează ca rezultat, direct, valori numerice, care vor fi interpretate ca şi coeficienţi de multiplicare ai 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 cu **0.3** ori.*

Toate regulile din Tabelul 1 şi toate celelalte tabele se interpretează într-un mod similar.

Rezultatul final se obţine prin agregarea rezultatelor parţiale, date de regulile respective. De exemplu, dacă valoarea intensităţii sonore este de 45 dB, iar raportul HL este de 0,7, prin aplicarea fuzzificării[95] se obţine gradul de adevăr al premisei (combinate) din regula respectivă, prin

$$\min\left(\mu_{putere=mica}(P_0), \mu_{LH=mic}(LH_0)\right)$$

unde $P_0 = 45$ iar $LH_0 = 0{,}7$. Folosind expresiile funcţiilor (v. Anexa 1), se obţin valorile $\mu_{putere=mica}(P_0) \approx 0{,}67$, $\mu_{LH=mic}(LH_0) = 0{,}6$, 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, $\alpha_{1,1}^{A}$. În total, sunt 9 reguli per tabel, deci există 9 valori de singletoni. Într-adevăr, în acelaşi timp, valorile de intrare corespund funcţiilor de apartenenţă „mediu" pentru „putere" şi LH, deci regulii:

*DACĂ Puterea (zgomotului) este **mică** şi parametrul LH este **mic***

[95] *Termenul echivalent românesc ar fi "nuanţare"*

ATUNCI *Amplitudinea creşte cu **0.0** ori.*

cu gradul de încredere în rezultat:

$$\min\left(\mu_{putere=medie}(P_0),\mu_{LH=medie}(LH_0)\right)$$

precum şi regulilor:

DACĂ *Puterea (zgomotului) este **mică** şi parametrul LH este **mediu***

ATUNCI *Amplitudinea creşte cu **0,1** ori.*

respectiv:

DACĂ *Puterea (zgomotului) este **medie** şi parametrul LH este **mic***

ATUNCI *Amplitudinea creşte cu **0,1** ori.*

cu gradele de încredere

$$\min\left(\mu_{putere=mica}(P_0),\mu_{LH=medie}(LH_0)\right)$$

şi respectiv

$$\min\left(\mu_{putere=medie}(P_0),\mu_{LH=mic}(LH_0)\right)$$

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, $P_0 = 57$ si $LH_0 = 0{,}7$.

Prin agregare (defuzzificare), considerată aici conform formulei uzuale:

$$y = \frac{\sum_{k=1}^{9} \alpha_k^A \mu_k^A(x_0)}{\sum_{k=1}^{9} \mu_k^A(x_0)} \tag{3}$$

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, α_k^X reprezintă abscisele singletonilor de ieşire din sistemele tip Sugeno respective, $\mu_k^X(\cdot)$ reprezintă gradele de încredere în concluzia 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”, controlul fiind desigur diferenţiat pentru cei patru parametri discutaţi.

Valorile astfel obținute sunt folosite, cum s-a precizat, ca factori de multiplicare ai parametrilor nominali[96]. De exemplu, dacă amplitudinea nominală este A_0, atunci, prin aplicarea controlului, amplitudinea efectivă a semnalului va fi:

$$A = A_0 \cdot \left(1 + \frac{\sum_{k=1}^{9} \alpha_k^A \mu_k^A(x_0)}{\sum_{k=1}^{9} \mu_k^A(x_0)} \right) \tag{4}$$

Sistemul de control este instantaneu, în sensul că nu ține cont decât de valorile recente (din fereastra prezentă, de lărgime *W*) ale zgomotului, nu şi de valorile anterioare. Controlul de amplitudine şi frecvență se poate exercita în afara sintetizorului propriu-zis, asupra unui amplificator şi a unui filtru plasate la ieşirea sintetizorului. Aceste două controale se pot prevedea de altfel şi în alte aplicații, precum sisteme de sonorizare mari (eventual distribuite, ca în cazul sonorizării unor spații mari, gen piețe sau stadioane), sau a unor sisteme de sonorizare locale (de exemplu, sisteme de interfonie). Controlul pauzelor dintre cuvinte şi un control fin al spectrului vocalelor necesită comanda directă a sintetizorului.

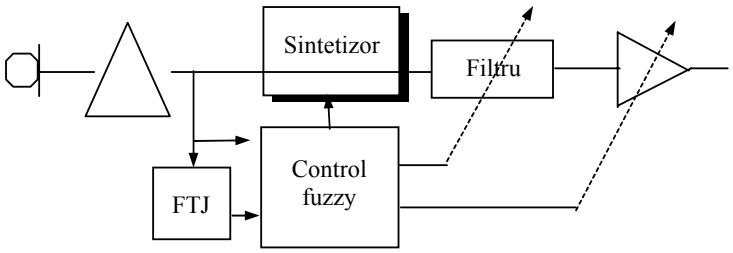


Figura 2. Schema bloc a unui sistem audio adaptiv la zgomotul ambiental

În cazul în care se utilizează doar primele două tipuri de adaptare, în amplitudine şi spectral, adaptarea se poate realiza şi cu mijloace hardware externe sintetizorului, putând, de altfel, fi utilizată în orice aplicație audio (de sonorizare etc.). Schema unui asemenea sistem de adaptare este cea prezentată în Figura 2, o variantă fiind inițial propusă în [4].

[96] *Nominali, în sensul că sunt valorile standard pentru sistemul de sinteză automată respectiv şi pentru sunetul respectiv produs în condițiile 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. In 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 caracterisitici 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).

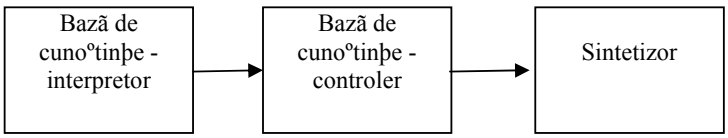


Figura 3. Schema de principiu a controlului contextual-interpretativ

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),*

ATUNCI afectul sintezei este mirare / neîncredere/…/ etc.

ori

DACĂ *oferta / răspunsul interlocutorului este neaşteptat negativ (conform bazei de cunoştinţe),*

ATUNCI 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 în spectrul 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 sintezã 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) a spectrelor formanţilor. "Claritatea" vocii se poate traduce prin segmentarea mai pronunţată pe cuvinte, precum şi vocale mai lungi (cu sau fară 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.) sau semnale lent variable, 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 iniţial în 1992 ([28] ş.a.), dar nepublicată în forma extinsă, credem că reprezintă o metodă promiţătoare de "personalizare" a vocii.

Considerăm un sistem dinamic neliniar, dependent de parametri; semnalul în timp generat de acesta este de forma $x(t) = x\left(t, \lambda_1, \lambda_2, ..., \lambda_q\right)$, unde λ_h reprezintă parametrii sistemului haotic şi permit modelarea specificităţii vorbitorului. Semnalul *x* poate fi folosit în modularea amplitudinii, frecvenţei fundamentale, sau spectrului semnalului vocal sintetizat. De exemplu, spectrul poate fi modificat folosind o lege de variaţie a frecvenţei centrale a formanţilor de forma:

$$f_j(t) = \left(1 + x_j(t)\right) \cdot f_{j0}(t) \tag{5}$$

unde $f_j(t)$ este frecvenţa formantului numărul *j* la momentul *t*, $x_j(t)$ este semnalul haotic respectiv $\left(x_j(t) << 1\right)$, iar $f_{j0}(t)$ este frecvenţa "nominală" a formantului respectiv.

Un exemplu simplu de sistem haotic ce poate fi folosit în acest scop este dat de ecuaţiile:

$$\begin{cases} r_{n+1} = \lambda_3 \cdot u_n^3 + \lambda_2 \cdot u_n^2 + \lambda_1 \cdot u_n + \lambda_0 \\ u_n = \lambda_4 \cdot r_n + \lambda_5 \end{cases} \tag{6}$$

unde setul de coeficienţi $\left(\lambda_0, \lambda_1, ..., \lambda_5\right) \in \mathbf{R}^6$ se alege în domeniul de valori ce corespunde unui comportament haotic al sistemului (vezi Anexa 2). Setul de coeficienţi $\left(\lambda_0, \lambda_1, ..., \lambda_5\right)$ se poate seta specific pentru fiecare sistem de sinteză automată, "personalizând" sistemul. Valorile de ieşire ale generatorului se scalează corespunzător şi se folosesc la modularea unuia dintre parametrii de sinteză. Pentru exemplul din secţiunea 3, amplitudinea semnalului sonor devine, prin utilizarea modulaţiei haotice:

$$A_n = A_0 \cdot \left(1 + \frac{\sum_{k=1}^{9} \alpha_k^A \mu_k^A(x_0)}{\sum_{k=1}^{9} \mu_k^A(x_0)} \right) \cdot \left(1 + \kappa \cdot r_n\right) \tag{7}$$

unde κ este un coeficient de scalare a seriei de timp r_n. Coeficientul κ se alege astfel încât contribuţia termenului $\kappa \cdot r_n$ să fie de ordinul procentelor ($\kappa \cdot r_n < 0{,}1 \ \forall n$).

Desigur, scara de timp a procesului de generare de eşantioane de semnal vocal diferă de scara de timp a proceselor haotice folosite în modulaţie, ceasul celui de al doilea proces fiind mult mai lent (de ordinul 1/100) decât al primului proces. Pentru evitarea tranziţiilor bruşte ale parametrului controlat, valorile generate pot fi interpolate şi se poate realiza o variaţie 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", dupa ceasul de generare a esantioanelor semnalului vocal):

$$r_k = r_{n-1} + \frac{r_n - r_{n-1}}{Q} \cdot k\ ,\ k = 0, 1, ..., Q \tag{8}$$

În scopul modulării haotice a mai multor parametri de sinteză (amplitudine, frecvenţa centrală a formanţilor, lărgimea formanţilor, elemente prozodice etc.), sunt necesare mai multe generatoare haotice, câte unul pentru fiecare parametru controlat. Alternativ, se poate folosi un sistem nuanţat (fuzzy) haotic, aceste sisteme generând simultan un număr mare de ieşiri necorelate sau slab corelate [28].

6. Concluzii şi discuţii

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. Adaptabilitate 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 aplicaţii de uz general, cât şi pentru aplicaţii medicale.

O aplicaţie de interes medical-educaţional este utilizarea unor sisteme de învăţare a unei limbi pentru copii de vârste mici (1 lună – 3 ani) care suferă de deficienţe 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 înveţe limba la această vârstă. Este, într-adevăr, demonstrat că învăţarea primelor elemente ale unei limbi la aceste vârste asigură o şansă mult mai mare de învăţare 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 soluţii pentru a face vocea sintetică purtătoare de informaţie emoţională. Astfel, în [31] se descrie o metodă de sinteză a "vocii emoţionale", capabilă să transmită trei emoţii (supărare-furie, bucurie, tristeţe) folosind elemente de prozodie şi segmente de tip vocală-consoană-vocală (specifice limbii japoneze). În [32], starea ("mood") şi personalitatea sunt văzute ca elemente esenţiale aparând în subsidiar în voce şi necesar a fi introduse şi în vocea sintetizată. Alţi autori [33] vorbesc de "nivelul de plăcere al audiţiei" (pleasantness) – dincolo de inteligibilitate – şi văd naturaleţea vocii sintetizate prin această prismă, a utilizării la 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 features efficiently.").

Comment [R1]: "

Credem cã, în viitor, o metodã comodã de a genera automat prozodia, pentru o voce artificialã datã ºi pentru o anumitã stare, ar putea fi constituitã de o procedurã inversã celei descrise în [34].

Incheiem cu un citat din [35]: "… in spite 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[97]: 1) improved linguistic analyses, 2) improved prosody modeling, and 3) improved speech synthesis models."

Mulţumiri. Această lucrare a fost realizată cu sprijinul material al Academiei Române – Institutul de Informatică Teoretică Iaşi – precum şi cu sprijinul material parţial al Societăţii "Tehnici şi Tehnologii" s.r.l. Iaşi. Autorul multumeşte colegilor Dragoş Burileanu, Bogdan Branzilă si Oana Geman pentru sugestii şi corecţii la o formă preliminară a lucrării.

Referinţe bibliografice

[1] Teodorescu H.N., Chelaru M., Sofron E., Adascalitei A.: Adaptive speech synthesis. In vol. Digitale Sprach-verarbeitung - Prinzipien und Anwendungen. VDE Verlag, Berlin (W), pp. 183-188, 1988

[2] Teodorescu H.N.: Interrelationship, Communication, Semiotics, and Artificial Consciousness. In: Kitamura, T. (Ed.): *What Should be Computed to Understand and Model Brain Functions?* FLSI Book Series, vol. 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-adaptabile. *Electronic Engineering* (UK), July 1987, p. 23

[5] Rodriguez, W., Teodorescu H.N., Grigoras Fl., Kandel A., Bunke H.: A Fuzzy information space approach to speech signal nonlinear analysis. *J. of Intelligent Systems* (Wiley), Dec. 1999

[6] Grigoras Fl., Teodorescu H.N., Apopei V.: Nonlinear Analysis and Synthesis of Speech. *Studies in Informatics and Control*, vol. 7, no. 1, March 1998, pp. 57-72

[7] Teodorescu H.N., Grigoras Fl., Apopei V.: Nonlinear processes in speech production. *Int. J. Chaos Theory and Applications*, vol. 2, no. 2 (1997), pp. 35-52

[97] *Aici, autorul citat face referire la L. R. Rabiner, "Applications of Voice Processing to Telecommunications,"* Proc. IEEE*, vol. 82, pp. 199–228, February 1994.*

[8] Teodorescu H.N., Grigoraş Fl.: Nonlinear Techniques in Speech Signal Analysis. Proc. International Conference on Intelligent Technologies in Human-Related Sciences, ITHURS'96. July 5-7, Leon, Spain. Vol. 2, pp. 293-298, 1996

[9] Grigoraş Fl., Teodorescu H.N., Apopei V.: Analysis of nonlinear and nonstationary processes in speech production, IEEE 1997 Workshop on Applications of Processing to Audio and Acoustics. Mohonk Mountain House New Paltz, New York, October 19-22, 1997 (IEEE Catalog # 97^{TH}8278)

[10] Burlui V., Teodorescu H.N., Morarașu C.S.: La fonction phonatoire chez l'edente total. Analyse en frequence. *Les Cahiers de Prothese* (France), No. 88, Decembre 1994, pp. 63-68 1994

[11] Teodorescu H.N. et al.: Fuzzy models in speech analysis and medical application, in Book of Summaries Int. Conf Modelling and Simulation, Istanbul, Turkey, July 1988, vol. 1, p. 162 (Summary)

[12] Teodorescu H.N., L. Buchholtzer, Chelaru M., Teodorescu L.: A laryngeal prosthesis based on perilaryngean reflexes, Proc. 9th Int. EMBS Conf. IEEE, Boston. Vol. 4, IEEE, pp. 2114-2115, 1987

[13] Anonymous Automotive Industry OEM/Supplier: Talking to computers vs. talking to humans 7/12/2000. http://www-nrd.nhtsa.dot.gov/departments/nrd-13/driver-distraction/Topics013040293.htm#A293

[14] Anne-Marie Derouault, The Future of Speech Recognition. Evolving speech recognition technology is driving transparent computing, making it easier for people to interact with computers. http://www.advisor.com/Articles.nsf/ ID/OA000107.DERO01

[15] House D., Bell L., Gustafson K. & Johansson L. Child-directed speech synthesis: evaluation of prosodic variation for an educational computer program. Proc of Eurospeech'99, pp. 1843-1846, 1999

[16] Heldner M., Strangert E. & Deschamps T.: Focus detection using overall intensity and high frequency emphasis. In: Andersson R, Abelin Ĺ, Allwood J & Lindblad P, eds. Proc of Fonetik 99; pp. 73-76, 1999.

[17] Heldner M., Strangert E. & Deschamps T.: A focus detector using overall 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. van Dommelen & T Fretheim, eds. Nordic Prosody: Proc of the VIIIth Conference, Trondheim 2000 . Frankfurt am Main: Peter Lang. 2001

[19] Karlsson I., Banziger T., Dankovicová J., Johnstone T., Lindberg J., Melin H., Nolan F. & Scherer K.: Within-speaker variability due to speaking manners. 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 Ĺ, Allwood J & Lindblad P, eds. Proc. of Fonetik 99, pp. 93-96, 1999.

[21] Karlsson I, Banziger T, Dankovicova J, Johnstone T, Lindberg J, Melin H, Nolan 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-Capkova 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 & Traunmüller 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*, Vol. 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. CHI'95 Proc., www.acm.org/sigchi/chi95/proceedings/shortppr/ml_bdy.htm

[27] Christophe d'Alessandro & Jean-Sylvain Liénard: 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. Vol. Proceedings of the 2nd Int. Conf. on Fuzzy Logic and Neural Networks. Vol. 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 Computational 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 Emotional 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 Features for Speech User Interfaces. www.cs.uta.fi/hci/spi/reports/Prosodic_Features_for_Speech_User_Interfaces.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 1. Funcții de apartenență

Reamintim ca o multime (clasică) $A \subset X$, 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 corespunzatoare astfel: o mulțime nuanțată, notată $\widetilde{A}$, 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:

$$\mu_A(x) = \begin{cases} 1 & pentru \quad x = a \in X \\ 0 & pentru \quad x \neq a \end{cases}$$

caz în care se numeşte *singleton*.

Un sistem de tip Sugeno, de ordin 0, este descris de reguli de forma:

DACA intrarea (premisa) # 1 SI premisa # 2 SI ... Si premisa # n ATUNCI concluzia

unde premisele sunt de forma: x_i *este* $\widetilde{A}_{ij}$, iar $\widetilde{A}_{ij}$ sunt valoari nuantate (fuzzy), de exemplu $\widetilde{A}_{i1}$ = "mare", $\widetilde{A}_{i2}$ = "mediu", atributelor lingvistice "mare", "mic" etc. fiindu-le atașate câte o funcție de apartenență. Specific sistemelor Sugeno este faptul ca în concluzie apar valori numerice şi nu valori nuanțate, concluzia fiind deci de forma "y = 0,3" (singleton).

Definițiile funcțiilor de apartenență pentru intensitatea sonoră din Figura 1.a sunt:

$$\mu_{Putere=mica}(p)=\begin{cases}1 & pentru \quad p \le 40\ dB\\ 1-\dfrac{p-40}{15} & pentru \quad 40 < p < 55\ dB\\ 0 & pentru \quad p \ge 55\ dB\end{cases}$$

$$\mu_{Putere=medie}(p)=\begin{cases}0 & pentru \quad p \le 40\ dB\\ \dfrac{p-40}{15} & pentru \quad 40 < p < 55\ dB\\ 1-\dfrac{p-55}{15} & pentru \quad 55 \le p < 70\ dB\\ 0 & pentru \quad p \ge 70\ dB\end{cases}$$

$$\mu_{Putere=mare}(p)=\begin{cases}0 & pentru \quad p \le 55\ dB\\ \dfrac{p-55}{15} & pentru \quad 55 < p < 70\ dB\\ 1 & pentru \quad p \ge 70\ dB\end{cases}$$

Definițiile funcțiilor de apartenență pentru raportul HL (Figura 1b) sunt:

$$\mu_{HL=mica}(q)=\begin{cases}1 & pentru \quad q \le 0.5\\ 1-\dfrac{q-0.5}{.5} & pentru \quad 0.5 < q < 1.\\ 0 & pentru \quad p \ge 1.\end{cases}$$

$$\mu_{HL=medie}(q)=\begin{cases}0 & pentru \quad q\le 0.5\\ \dfrac{q-0.5}{0.5} & pentru \quad 0.5<q<1.0\\ 1-\dfrac{q-1.}{0.5} & pentru \quad 1.\le q<1.5\\ 0 & pentru \quad p\ge 1.5\end{cases}$$

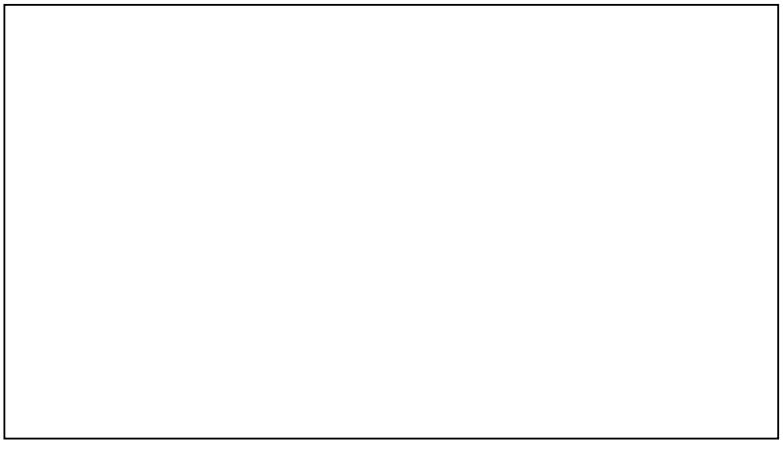
$$\mu_{HL=mare}(q)=\begin{cases}0 & pentru \quad q\le 1.\\ \dfrac{q-1.}{0.5} & pentru \quad 1.0<q<1.5\\ 1 & pentru \quad q\ge 1.5\end{cases}$$

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.



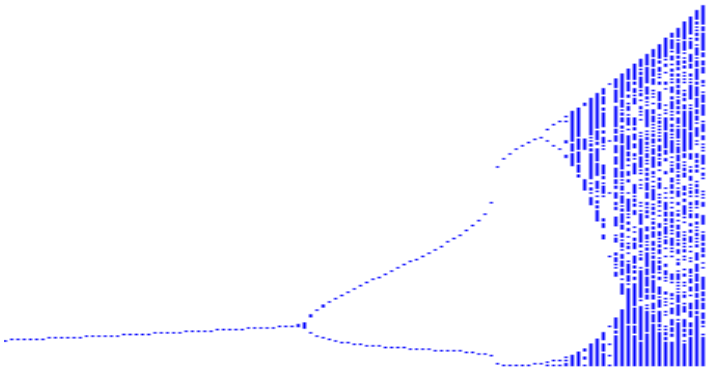


Figura A2-1. Diagrama de bifurcație a procesului

Legile folosite (conform codului, scris în limbajul C) sunt:

u[n]= (coeff_4)*r[n]+coeff_5 -0.005*(float)k;

x=u[n]; r[n+1]=poly(x, Q, coeff);

(Q este numărul de valori în vectorul coeficienților, Q=4)

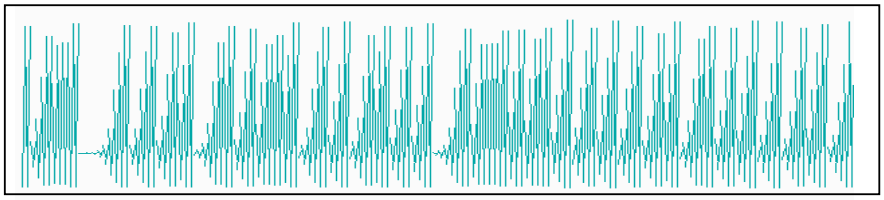


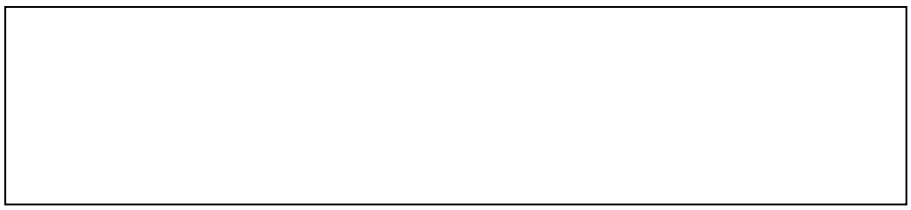
Figura A2-2

Semnalul în domeniul amplitudine-timp din Figura A2 a fost obținut pentru ecuațiile (cod C):

u[n]= coeff_4*r[n]+coeff_5-0.05*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 A3.



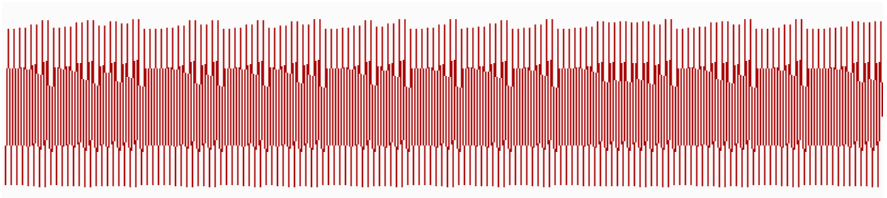


Figura A2-3

iar 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 A4:

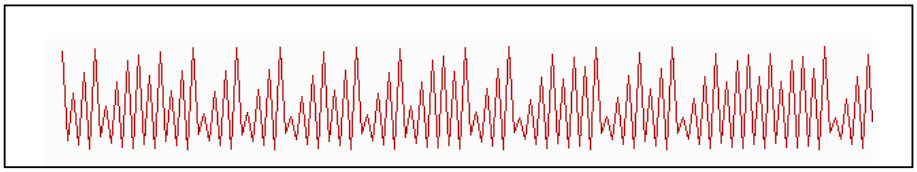


Figura A2-4

Regiunile spațiului parametrilor în care sistemul este stabil, după cum s-a spus deja, sunt relativ înguste. Pentru parametrii coeff_1-coeff_4 fixați şi coeficientul coeff_5 variabil între –25.15 şi + 4.85 (600 de paşi, cu pas 0,05), doar zona îngustă din Figura A2-5 este stabilă, oscilantă sau haotică, în rest sistemul fiind asimptotic instabil. Pentru uşurința urmăririi scării, linia din partea de jos a figurii 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 absolută, 10000.

Dicþionarele multimedia ale limbii române. Secvenţe de implementări şi experimentări

Dumitru TODOROI, Diana MICUSA, Zinaida TODOROI, Ion LINGA, Ion COVALENCO, Nicolae OBJELEANU, ªtefan SPÃTARU, Stela LUNGU, Virginia ÞURCANU, Elana COZLOV, Nadejda AMBROZII, Victor SLOBODEANU, Igor COªERU, Cãtãlina SURUCEANU
Academia de studii economice din Moldova, Str. Banulescu-Bodoni, 59-61/503»B», Chiºinău MD 2005, Republica Moldova,
E-mail : todoroi@ase.md

Lucrarea actuală în cadrul punerii Marelui Dicþionar al Limbii Române (MDLR) pe calculator a fost medodologic înfluenþată de ideile subliniate în [1-3] şi este o continuare a cercetărilor [4-7,10-11], effectuate în cadrul procesării lombajului natural. Au fost elaborate un şir de proiecte [8-9,12] de informatizare a Limbii Române. Experimentările cu elaborarea sistemelor computerizate de nivelul unu, care susţin diferite sub-dicţionare ale MDLR pe aşa axe ale lui ca: 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 de cercetări şi experimentări în cadrul platformei, alcătuite din aceste 4 subsisteme, sunt expuse în lucrarea de faţă. Sunt prezentate diferite scenarii [19] şi metodologii de utilizare a sub-dicţionarelor informatizate a limbii române. Clarificarea mijloacelor Hardware-ului şi Software-ului modern, care pot suporta MDLR informatizat [18], constituie o problemă importantă la etapa creării Societăţii Informaţionale – Societate a Cunoaşterii [20].

I. Componenta TEXT a dicþionarului economic MULTIMEDIA al limbii Române [23].

Scopul acestui compartiment computerizat al MDLR constă în crearea subsistemului TEXT de nivelul unu a unei părţi introductive a dicţionarului economic al limbii române şi experimentarea cu acest sistem. Acest dicţionar economic constă din 35.000 – 40.000 cuvinte. Cuvintele, care se conţin în Dicţionarul Enciclopedic Ilustrat (DEI) [21], vor fi definite 100/100. Experimentarea cu subsistemul TEXT al MDLR computerizat este efectuată la moment cu circe 200 articole din DEI.

Baza de date TEXT (BDT) a dicţionarului economic constă dintr-o culegere de texte-articole, alcătuită din cuvinte, fraze, paragrafe, capitole etc. ale DEI. Documentele în

BDT includ nu numai informaţii textuale (definiţii de cuvinte), dar pot conţine informaţii şi de alt tip, de exemplu, prin extindere, imagini. Prin urmare BDT în sistemul computerizat al MDLR conţine nu numai materialul textual, dar şi ilustrativ: diagrame, grafice, fotografii etc.

Prin crearea subsistemului TEXT de nivel unu utilizatorul obţine un mijloc important, prin intermediul căruia informaţia poate fi introdusă şi utilizată în mod complimentar pe cale electronică.

1.1. Capacitãþile 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 funcţiona efectiv e necesar ca SGBDMM să posede următoarele capacităţi:

(a) Capacitatea de a interoga datele, uniform reperezentate în diferite formate;

(b) Capacitatea de a interoga datele, reprezentate în diferite medii;

(c) Capacitatea de a transmite subiectele mediei din dispozitivele de stocare locale într-un mod efectiv;

(d) Capacitatea de a primi răspunsul la o interogare şi de a desfăşura prezentarea acestui răspuns pe baza mediei audiovizuale;

(e) Capacitatea de a furniza această prezentare pe acea cale, care ar satisface calităţile diferitor cerinţe ale serviciului.

1.2. Structura bazei de date TEXT (BDT) a dicþionarului economic MULTIMEDIA

Dicţionarul economic, care este pe cale de a fi pus pe calculator, este o BDT cu posibilitatea de a fi extinsă cu diferite componente ale MULTIMEDIEI. Subsistemul TEXT a dicţionarului economic MULTIMEDIA este un subsistem al SGBDMM, care aprovizionează acestă posibilitate împreună cu utilizarea complimentară a BDT.

Structura BDT e compusă din:

(1) Indice cu caracteristica “număr”;

(2) Termen principal (cuvînt, articol) cu caracteristica “text”;

(3) Variantă(e), derivate, abreviere (concretizare) cu caracteristica “text”;

(4) Categorie gramaticală cu caracteristica “text”;

(5) Domeniu cu caracteristica “text”;

(6) Definiţii pentru termenul principal (şi concretizări) cu caracteristica “text”;

(7) Sinonim(e) cu caracteristica “text”;

(8) Antonim(e) cu caracteristica “text” şi aştele. De asemenea BDT are posibilitatea de a fi extisă cu aşa subdiviziuni MULTIMEDIA ca:

(9) Audio cu caracteristica „OLE”;

(10) Imagini cu caracteristica „OLE”;

(11) Video cu caracteristica „OLE” şi altele.

1.3. Scenarii de utilizări şi interogări a subsistemului TEXT al MDLR unformatizat

Interogarea este o formă, care ajută utilizatorul să prezinte o informaţie anumită într-o structură anumită, definită de utilizător. Spre exemplu, utilizătorul doreşte să obţină informaţii din arhive, articole, sau alte documente, care conţin informaţia despre Uniunea Europeană. Interogarea poate avea următoarea formă: “Găseşte toate dosarele, legate de investiţiile străine, făcute de UE în domeniul educaţiei”. Un simplu cuvînt cheie a acestui dosar nu va permite găsirea răspunsului corect, chiar şi dacă indicile acestui document deja există. Totuşi, sistemul ne va prezenta careva cuvinte, legate de această interogare, dar ele pot să nu fie direct relatate la tema dorită. De aceea textul trebuie să fie indexat nu numai pe cuvintele cheie, dar şi pe conţinitul semantic şi/sau pragmatic al cuvintelor (în cazul BDT, de exemplu, concretizarea).

Soluţionarea problemei utilizătorului, care dorezte să afle definiţia cuvîntului “Academie”, de exemplu, cere întroducerea polisemiei în BDT, care conţine aşa concepte ca precizia şi rechemarea. Întrebarea, propusă de către utilizător în acest context, este: “Cum să aflu din baza de cunoştinţe a MDLR sensul cuvîntului “Academie – ca instituţie de învăţămînt economic”. Pentru aceasta BDT va fi completată cu o nouă coloană “concretizare”, care va preciza şi va face posibilă afişarea pe monitor anume a acelei definiţii a cuvîntului, de care utilizatorul este cointeresat (de exemplu: Academia de studii economice).

Un fragment de structurã schematicã a BDT.

I Indece	I Cuvânt	I Concretizare	I ….. I	Definiþie	I …. I	Traducere cuvânt I
I 03342	I Academia	I de studii economice I		I Nume dat ºcolii de ...	I	I
I 14269	I Banii	I EURO	I	I Denumire a princip...	I	I
I 14271	I Banca	I de economii	I	I	I	I

SGBDMM, ca o extindere a SGBD Ms ACCESS-2000, în baza căruia este creată componenta TEXT a MDLR informatizat, gestionează BDT, utilizînd limbajul SQL. În exemplele următoare utilizătorul este cointeresat în sfera finanţelor. Accesul BDT a MDLR este efectuat prin intermediul următoarelor interogări din SQL (care, în general, constituie aşa comenzi ca 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-dicţionarul explicativ economic al limbii Române [24]

Dictionarul 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 cerinţele de bază către un dicţionar informatizat: prezintă formele exacte ale cuvintelor, accentul, etimologia, definiţia-tezt, definiţia-sunet (audio), definiţia-video (film), definiţia-imagine (grafic, schema, poza etc.) şi corespunde cerinţelor unor categorii foarte largi de utilizatori nu numai elevi şi studenţi, dar şi funcţionari şi profesionişti, contribuind la ridicarea nivelului de cultura.

Compartimentul AUDIO al MDLR informatizat furnizează informaţii necesare ale articolului respectiv (cuvîntul, definiţia lui) în forma AUDIO. Subsistemul AUDIO de nivelul unu al SGBDMM oferă posibilitatea de AUDIO-utilizare a dicţionarului. Acest AUDIO-dicţionar va contribui din plin la ridicarea pe o treaptă superioară a societaţii noastre în exprimarea economică corectă orală şi scrisă. Conţinutul datei AUDIO poate fi caracterizată prin două metode: (a) folosind metadata prin explicarea conţinutului unui fişier AUDIO sau (b) prin extragerea tipului potrivit de date AUDIO, folosind procesorul tehnic.

2.1. Componenta AUDIO a metadatelor

Cu un fisier AUDIO se procedeaza la fel ca ºi în cazul unei date VIDEO: acestui fiºier i se asociază un set (grup) de segmente, toate referindu-se la o perioadă de timp. Fiecărui segment i se atribuie un set de activităþi, care au decurs în acea perioadă de timp, subliniate prin aceste segmente. În general, metadata utilizează reprezentarea AUDIO, care este sesizată ca un set de obiecte marcate în timp.

Utilizarea componentei **AUDIO** a metadatei din MDLR informatizat este mai mult recomandată cînd este o modalitate de creare a acestei metadate, modificarea ei ºi, în deosebi, la interogarea **AUDIO** –dicþionarului de utilizatorii, care necesită această formă de comunicare om-maºină.

Crearea componentei **AUDIO** a metadatei este un lucru mai complex decît alte forme de dicþionare informatizate, deoarice identitatea indivizilor, ce vorbesc, nu poate fi uºor cunoscută; chiar ºi conþinutul discursului poate fi neclar.

Conceptul despre conþinut este descris în termeni de metadată a procesului. Ca rezultat data **AUDIO** este consideratã ca un semnal ?DELTA?(x) în timpul x. Trăsăturile utilizatorice ale acestui semnal DELTA(x) sunt: (a) extragerea, (b) indicarea ºi (c) depozitarea.

O undă constă dintr-un set de vîrfuri (creste) ºi adînciturі (văi). Perioada vibraþiei T este definită ca timpul, pentru care o parte a undei să revenă la poziþia iniþială. Alte caracteristici utilizate de componenta AUDIO în crearea metadatei sunt: (1) frecvenþa, (2) viteza ºi (3) amplituda.

Baza de date **AUDIO (BDA)** poate fi interacþionatã ºi gestionatã, utilizînd sunetul auditiv prin intermediul procesiunii de segmentare, memorizare ºi extragere a informaþiei.

2.1.1. Segmentarea

Segmentarea e o procedurã de separare a sunetului auditiv în cîteva ferestruici egale. Aceastã procedurã poate fi utilizatã conform urmãtoarelor douã metode:

(a) Utilizatorul specificã dimensiunile ferestrei, presupunînd cã proprietãþile undei ºi a ferestrei se vor obþine prin medie;

(b) Utilizatorul segmenteazã sunetul în acelaºi mod ca ºi imaginiile, folosind predicatul de omogenitate H.

2.1.2. Extragerea

La extragere cel mai des utilizate sunt facilitãþile de indicare a intensitãþii, zgomotului, înãlþimii ºi strãlucirii.

2.2. Unele sisteme de utilizare a BDA

Din punct de vedere a MULTIMEDIEI, AUDIO - baza de date (BDA) poate fi interpretatã ca o sursã auditivã, ca un fiºier cu o fereastrã auditivã ºi cu trasãturile respective, asociate aceastei ferestre.

Scenariile de utilizare a BDA cuprind toate formele MULTIMEDIEI, 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 sã acceseze baza de date, bazîndu-se pe nesiguranþa conþinutului.

Baza de date DB2, un alt exemplu, utilizatã cu calculatorul de tip IBM, necesitã cuplarea cu un sistem adãugãtor, care permite lãsarea mesajelor vocale pe robot. DB2 poate importa ºi menþine clipurile, care pot fi cautate printr-un nume sau descriere.

Putem reaudia mesajele, lãsate pe robot, prin intermediului **Internetului**. Un exemplu în plus constituie o utilizare a unui cuvînt din AUDIO-dicþionarul economic al limbii române prin intermedioul AUDIO-VIDEO-robotului, care este un sistem autorizat ºi care acþioneazã pe baza unui program de lucru stabilit sau care reacþioneazã la anumite influienþe exteriore.

Un exemplu de interogare a componentei AUDIO a subdicþionarului economic al MDLR prin indermediul limbagului SQL ºi al subsitemului AUDIO de nivel unu al SGBDMM poate avea forma :

SELECT	Robot
FROM	Ţurcan-Mutruc
WHERE	Attribute **IS** Definiþie **AND** Attribute **IS** Audio

Ca rezultat al acestei interogări utilizatorul prin intermediul răspunsului prietenos obþine pe ecran definiþia TEXT a cuvîntului Robot ºi în paralel acest subsistem AUDIO al SGBDMM difuzează acestă definiþie cu voce feminină sau masculină (la dorinþa utilizatorului).

III. Subsistemul IMAGINI de nivel unu al dicţionarului economic informatizat al limbii române [25]

Scopul acestui capitol constă în descrierea posibilităţilor de întroducere a imaginilor în baza de date a MDLR informatizat şi de utilizare a acestora în viaţa cotidiană. Daza de date IMAGINI (BDI) a subdicţionarului economic al MDLR informatizat constituie o componentă, care oferă posibilitatea de extindere a procesului de înţelegere a sensului cuvîntului dat. Din cele aproximativ 35000-40000 de articole ale dicţionarului economic din MDLR doar numai 50-60%, după părerea noastra, pot fi prezentate în forma de imagini.

Experienţa, obţinută pe baza cîtorva zeci de articole din DEI în cadrul evaluării subsistemului IMAGINI al SGBDMM, ne confirmă întru totul conţinutul zicalei : «Mai bine odată să vezi decît de o sută ori să auzi» şi a zicalei «Un tablou este egal cu o mie de cuvinte». Aceste facilităţi utilizatorice din evoluţia creatorică şi utilizatorică a MDLR informatizat le confirmă şi lucrările din [22] chiar şi prim 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 video they've seen

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 povesti mai mult despre un obiect decît câteva pagini (Vezi Tabelul 3.1) de descrieri textuale. Pentru un chirurg este cu mult mai uºor sã-ºi gãseascã un pacient potenþial prin investigarea diferitor imagini. Imaginile pot fi combinate cu corpusuri, text-definiþii, sunet-definiþii, traduceri etc.

În afara de datele IMAGINI ale dicþionarului economic MULTIMEDIA în MDLR informatizat vor fi prezente aºa tipuri de date ca video, audio, document, manuscrise ºi altele. Datele VIDEO sunt des folosite în domeniul învãþãmîntului. Datele AUDIO sunt importante în domeniul criminalisticii, de exemplu, în identificarea vocilor celor suspectaþi. Datele documentare diferã de datele TEXT prin aceia, cã pot sã conþinã nu numai informaþii textuale, dar ºi imagini încadrate. Datele manuscrise se presupune cã în viitorul apropiat vor prevala înregistrãrile electronice.

Sunt cunoscute diferite formate electronice, care dau posibilitatea de a vizualiza imaginea (fiºierele de tip GIF, TIFF, PCX, de exemplu). Subsistemul IMAGINI a SGBDMM are anumite trãsãturi specifice necesitãþii de utilizare a imaginilor ca o componentã vitalã a MDLR informatizat.

3.2. Subsistemul IMAGINI

Baza de date IMAGINI diferã de bazele de date TEXT ºi AUDIO prin complicitatea imaginilor, necesitatea de a diviza, combina ºi utiliza diferite parþi componente ale imaginii, care deseori la interogare se complicã ºi prin utilizarea incorectã ºi analiza neprecizatã a tehnicilor de manipulare a imaginilor. Aceasta se complicã ºi prin aceia, cã diferite organizaþii adunã date fotografice, hãrþi, scheme ºi alte imagini de tip universal sau specializat (cum ar fi, de exemplu, NASA). Interogãrile datelor de tip IMAGINI sunt efectuate în baza datelor de tip TEXT, cautate în baza de date de tip IMAGINI ºi vizualizate în formã de text ºi 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 sunt prevãzute un set larg de proceduri cu imaginile.

3.2.1. Plasa imaginii

Conþinutul imaginii constã din toate obiectele acestei imagini ºi caracteristicile lor, care reprezintã interes din punct de vedere a programului aplicativ. Imaginea poate avea o mulþime de proprietãþi, asa ca descrierea formei, prezentarea vectorului subdiviziunilor, prezentarea vectorului ordinii de descompunere ºi compunere a imaginii ºi altele. Fiecare imagine “I” are o pereche asociatã schimbãtoare de numere pozitive (m,n), care se numeºte **plasa imaginii** . Ea este compusã din m*n celule de mãsuri 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 similării imaginilor: Abordarea metrică şi abordarea de transformare.

Abordarea de transformare este mai generală decît abordarea metrică. Această abordare utilizează aşa operaţiuni ca: transformarea, transferarea, rotaţia, scalarea, simetrizarea şi a.

3.3. Utilizarea imaginii

În prezent multe instituţii de învăţământ oferă programe de studii individuale. Unele persoane studiază cursuri speciale de sinestătător pentru dezvoltarea intelectului uman. Astfel de cursuri pot fi reprezenatate sub formă de imagini speciale.

Imaginile pot fi utilizate în industria filmelor. Specialiºtii au posibilitatea de a vizualiza imaginile, alese de ei, lucrînd la calculator.

Imaginile sunt importante ºi în industria turizmului. Pentru informaþii despre imaginile necesare la fel se poate apela la sistemul de tip IMAGINE al SGBDMM.

Interogãrile de imagini în dicþionarul economic al MDLR informatuizat pot fi efectuate la fel ca ºi în subsistemele de acelas nivel unu de tip TEXT ºi AUDIO prin intermediul limbagului SQL al SGBD. Rezultatul în forma textualã a articolului ºi imaginea într-o formã complimentarã este reprezentatã utilizatorului în formã de Soft-copy sau Hard-copy.

Obþinerea imaginei cuvîntului «bancã», de exemplu, în subsistemul IMAGINI al SGBDMM al MDLR informatizat se efectuiazã prin intermediul urmãtoarelor acþiuni. Se deschide baza de date IMAGINI al dicþionarului economic al limbii române (în care sunt acumulate la momentul experimentării cu SGBDMM al MDLR informatizat doar numai 25 de cuvinte cu imagini respective). Se alege cuvîntul «banca». În înregistrarea respectivã a bãncii în compartimentul imagini se gãseºte OLE al imaginii cuvîntului ales. Se efectuiazã clic pe ea ºi vizualizãm pe ecran imaginea respectivã. Analogic se procedeazã ºi cu alte cuvinte din BDI.

IV. VIDEO-dictionarul economic al limbii romane[26]

În ultimii ani a crescut imens necesitatea de a putea chestiona şi procesa cantităţi mari de date, care nu sunt întotdeauna uşor de reprezentat prin intermediul simbolurilor. Exemple de astfel de date sunt: informaţia în formă de imagini, informaţia-video, datele-audio, informaţia textuală, notiţe şi altele. Ăn continuare vor fi examinate unele momente de realizare a dicţionarului economic informatizat cu VIDEO clipuri. A fost iniţiată baza de date VIDEO (BDV) a dicţionarului 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 unitãþile media dintr-o diviziune localã de depozitare, asigurând continuitatea acestui proces;

(d) SGBDMM trebuie sã primeascã rãspunsul, generat de o chestionare ºi sã poatã genera o prezentare a acelui rãspuns utilizând audiovizualul;

(e) Capacitatea de a oferi prezentarea într-un mod, care ar satisface diferite cerinþe ale utilizatorului.

Tehnologiile, legate de bazele de date, au dezvoltat în ultimii 40 de ani baza, pe care ar trebui să fie creatã o BDMM. În prezent sunt create limbaje de chestionare, tehnicile de aranjare, algoritmii de aportare pentru o mulþime de baze de date de tip relaþional, spaþial, temporal ºi altele. Fiecare din aceste mijloace extind posibilitãþile limbajelor ºi algoritmii precedenþi pentru a face faþã noilor tipuri de date sau pentru a argumenta paradigmele respective.

În acest capitol se va analiza informaþia de tip VIDEO. Necesitatea de a accesa o bază de date VIDEO (BDV) poate apărea într-o mulțime largă de 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 de răspuns la un set de întrebări de tipul:

(A) Ce aspecte posibilele ale filmului pot cointeresa utilizatorii BDV?

(B) Cum pot fi aceste aspecte ale filmului eficient depozitate, încât să minimalizeze timpul necesar subsistemului VIDEO al SGBDMM pentru a răspunde interogãrilor utilizatorilor?

(C) Cum ar trebui să fie limbajul de interogare a datelor VIDEO ºi cum ar trebui schimbat modelul relaþional pentru a corespunde informaþiei VIDEO?

(D) Poate fi oare automatizat procesul de extragere a informaþiei în baza contextului?

Aceste probleme au fost abordate în procesul creării ºi experimentării cu BDV ºi subsistemul VIDEO de nivel unu al SGBDMM.

4.2. Definiþiile datelor de tip VIDEO

De obicei un film este caracterizat prin personajele sale, atributele acestora ºi activitãþile, în care sunt angajate aceste personaje. Principalele surse de interes într-un film includ: (a) oameni, (b) obiecte neînsufleţite, (c) fiinţe însufleţite şi (d) activitãþi.

De observat, cã tema generalã, care se repetã în fiecare cadru, constã în aceea, cã existã un grup de obiecte ºi activitãþi asociate. Astfel vom încerca sã definim o bazã de date VIDEO printr-un ºir de definiþii.

Definiþie 1: O *proprietate* VIDEO este o pereche (*pname, Values*), unde *pname* este numele proprietãþii ºi *Values* este o mulþime. O *instanþã* a proprietãþii (*pname, Values*), este o expresie de forma *pname=v,* unde v⊂ *Values.*

Definiţie 2: O *schemă obiect* este o pereche (*fd, fi*), unde:

fd este o mulþime de proprietãþi cadru-dependente,

fi este o mulþime de proprietãþi cadru-independente (*fi* ºi *fd* sînt mulþimi disjunctive).

Definiþie 3: O *instanþã obiect* este un triplet (*oid, os, ip*), unde:

oid este o frazã numitã identitatea obiectului,

os= (*fd, fi)* este o schemã obiect ºi

ip este o mulþime de afirmaþii de tip:

(a) pentru fiecare proprietate (*pname, Values*), în *fi*, *ip* conþine cel mult o instanþã a proprietãþii (*pname, Values*),

(b) pentru fiecare proprietate (*pname, Values*) în *fd* ºi pentru fiecare cadru *f* al filmului, *ip* conþine cel mult o proprietate instanþã (*pname, Values*). Această proprietate instanţă este notată prin *pname* = v IN f.

Definiţie 4: O schemă activitate *ACT_SCH* este o mulţime finită de proprietăţi astfel încât, dacă (*pname, Values1*), şi (*pname, Values2*) ambele aparţin *ACT_SCH*, atunci *Values*1= *Values*2.

Definiţie 5: O activitate este o pereche, care constă din

(a) AcID, indecele schemei activitate *ACT_SCH* şi

(b) pentru fiecare pereche (*pname, Values*) ⊂ *ACT_SCH* este valabilă ecuaţia de forma *pname*= v, unde v ⊂ *Values*.

Oricărei activităţi i se asociază o schemă de activitate şi fiecărei proprietăţi i se asociază o valoare din mulţimea valorilor posibile.

Fiind dată o singură dată VIDEO v, putem defini "conţinutul" filmului v.

Definiţie 6: Fie că *framenum*(v) specifică numărul total de cadre din filmul v. *Conţinutul* lui *v* constă dintr-un triplet *(OBJ,AC,λ),* unde:

1. *OBJ*={oid_1,...,oid_n} este o mulþime finitã de instanþe ale obiectului,
2. *AC*={$AcID_1$,...,$AcID_k$} este o mulþime finitã de activitãþi/evenimente ºi
3. λ este o hartã de la {1,...,framename(v)} pânã la 2 $^{OBJ\ U\ AC.}$

Intuitiv, conþinutul unei date VIDEO v este teoretic descris de tripletul *(OBJ,AC,λ),* unde:

1. *OBJ* reprezintã mulþimea obiectelor de interes în film,
2. *AC* reprezintã mulþimea activitãþilor de interes din film ºi
3. λ reprezintã obiectele ºi activitãþile, care sunt asociate cu fiecare cadru *f* al filmului.

4.3. VIDEO biblioteca

O persoanã interesatã de obþinerea unei lecþii imprimate pe o casetã video ar dori sã chestioneze o VIDEO bibliotecã, care gãzduieºte o colecþie de casete video, referitoare la un anumit subiect. De exemplu, Universitatea Maryland oferã cursuri, utilizând contactul prin satelit. În viitor casetele video, create în acest fel, vor putea fi accesate cu ajutorul unui calculator, oferind astfel studenþilor prelegeri pentru diferite obiecte adunate în mai mulþi ani ºi þinute de diferiþi lectori. Chestionarea bazei de date VIDEO de un student individual ar presupune accesarea unui numãr foarte mare de casete video.

O bibliotecã VIDEO este o colecþie, care specificã: (a) totalitatea filmelor din bibliotecã, (b) conþinutul fiecãrui film ºi (c) memorizarea fizicã a filmelor.

Definiþie 7: O *VIDEO bibliotecã VidLib* constã dintr-o mulþime finitã de cvintete de tip (*VidContent, Vid_Id, framenum, R, plm*), unde:

(a) *VidContent* este conþinutul filmului,

(b) *Vid_Id,* este numele filmului,

(c) *Framenum* este numãrul 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) a*portarea 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 bibliotacă, 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 (VideoId, StartFrame, EndFrame),

FindVideoWithActivity (*a*)

FindVideoWithActivityandProp(*a,p,z*)

FindVideoWithObjectandProp(*o,p,z*)

FindObjectsInVideo(*v,s,e*)

FindActivitiesInVideo (*v,s,e*)

FindActivitiesAndPropsinVideo (*v,s,e*)

FindObjectAndPropsInVideo (*v,s,e*)

O chestionare standardă a VIDEO-bibliotecii, utilizînd SQL are forma:

SELECT câmp$_1$,…, câmp$_n$

FROM relația$_1$(*R1*), relația$_2$(*R2*),…, relația$_k$ (*Rk*)

WHERE condiție.

4.3.3. Ordonarea datelor VIDEO

O problemă importantă este crearea structurilor informaționale, care ar organiza bazele de date VIDEO în aşa fel încât să optimizeze procesarea celor opt funcții enumerate mai sus. Este imposibil de a depozita conținuturi al VIDEO-datelor cadru cu cadru, deoarece un singur film de 90 minute conține 162,000 cadre. Astfel este necesar de a crea reprezentări compacte a conceptului de conþinut video. În acest sens vom prezenta două astfel de structuri: (a) arborii segment cadru, ºi (b) arborii R-segment.

4.3.4. Arborii segment cadru

Ideea de bază a arborelui segment cadru este foarte simplă. La început se creează două tabele unidimensionale: OBJECTARRAY ºi ACTIVITYARRAY.

În acest context arborele poate fi creat în 2 etape:

La prima etapă presupunem, că [s_1,e_1),…, [s_w,e_w) sunt toate intervalele în coloana "Segment" a tabelei segment. Fie q_1,…,q_z o enumerație ascendentă a tuturor membrelor $\{s_i,e_i \mid 1?i?w\}$. Dacă z nu este exponent al numărului 2, atunci se procedează astfeel: fie r cel mai mic număr intreg aşa ca 2^r>z şi 2^r>*framenum(v)*. Se adaugă noi elemente q_{z+1},…,q_2r în aşa fel, că q_2r = *framenum(v)*+1 şi $q_{z+j} = q_z$+j (j>0, z+j<2^r).

La a doua etapă arborele este unul binar format după cum urmează:

1. În fiecare nod arborele segment cadru reprezintă o secvenþă de cadru [x,y).
2. Fiecare frunză este la nivelul r. Prima frunză din stânga marchează intervalul [z_1,z_2), a doua [z_2,z_3) ºi aºa mai departe.
3. Numărul din interiorul fiecărui nod este adresa acelui nod.
4. Mulþimea de numere de lângă nod marchează numărul de identitate al VIDEO-obiectelor ºi a VIDEO-activitãþilor, care apar în întreaga secvenþă de cadru asociată cu nodul dat.

Definiþie 8: O *secvenþã de cadru* este o pereche [i,j), unde 1?i?n şi [i,j) reprezintã mulþimea tuturor cadrelor între i (inclusiv) ºi j.

Definiþie 9: O *ordonare* parþialã $\subseteq$ asupra mulþimii tuturor secvenþelor de cadru este definitã ca $[i_1,j_1) \subseteq [i_2,j_2)$ cu condiția, cã $i_1<j_1=i_2<j_2$.

Definiþie 10: O mulþime X de secvenþe de cadru este *bine aranjatã* dacã:

1. X este finitã (adicã X= {$[i_1,j_1)$,..., $[ir_2,jr_2)$}, pentru oricare r_2) ºi
2. $[i_1,j_1) \subseteq [i_2,j_2) \subseteq \ldots \subseteq [ir_2,jr_2)$

Definiþie 11: O mulþime X de secvenþe de cadru este *solidã* dacã:

1. X este bine ordonatã ºi
2. Nu existã nici o pereche de secvenþe de cadru în X de forma $[i_1,i_2)$ ºi $[i_2,i_3)$

4.3.5. Operaþii 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 $v_1,\ldots, v_n$, 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 v_i, *fst(v_i) este* asociat cu filmul v_i.

De asemenea pot fi exprimate cele 8 funcþii, întrodise în SQL mai sus. De exemplu, una din aceste funcþii FindVideoWithObject(o), poate fi implementatã cu arborii segment cadru PRINTR-o operaþie de selecþie, efectuatã asupra INTOBJECTARRAY DE TIP:

```
SELECT   VIDEO_ID
FROM     INTOBJECTARRAY
WHERE OBJ = o.
```

4.3.6. Arborii R-segment (RS-arbori)

Arborii R-segment sunt foarte asemãnãtori cu arborii segment cadru, cu o singurã deosebire. Deºi conceptele de OBJECTARRAY ºi ACTIVITYARRAY rãmân aceleaºi, în locul utilizãrii unui arbore segment cadru pentru a reprezenta secvenþa de cadru profitãm de faptul cã o secvenþã [s,e) este un dreptunghi cu lungimea laturii (e-s) ºi lãþimea 0. Fiecare nod va avea o structurã specialã pentru a specifica, pentru fiecare dreptunghi, care obiect sau activitate este asociatã acestuia.

4.4. Operaţii cu VIDEO-clipuri

Un film este creat prin filmarea unor secvenţe şi combinarea lor, utilizând un operator de combinare. O secvenþã este de obicei filmatã de mai multe camere, fiecare având o vitezã relativã de rotaþie constantã. În general o secvenþã poate avea mai multe atribute asociate aºa ca durata filmării, tipul de camerã utilizat ºi altele.

Un operator de combinare a filmărilor, deseori numit *edit effect*, este o operaþie care în baza a douã filmări S_1 ºi S_2, ºi a unui interval de timp t efectuiazã o secvenþã compusã în timpul t. Aºadar un film este creat prin combinarea unei mulþimi de secvenþe filmate, utilizând un ºir finit de operaþii de compunere. Exemple de astfel de operaþii de compunere a filmelor includ:

1. Concatenarea filmărilor,
2. Compoziţia spaţială şi
3. Compoziţia cromatică.

4.5. Standardele video

Deşi în general standardele industriale nu sunt parte componentă a fundaţiei cadrului MULTIMEDIA este important în linii generale să explicăm ideea de bază a standardelor MPEG.

Toate standardele de comprimare a informaţiei VIDEO încearcă să comprime filmele prin executarea unei analize intra-cadru: fiecare cadru este divizat în blocuri, diferite cadre sînt comparate pentru a vedea, dacă informaţia conţinută de acestea, nu se repetă în două cadre. Calitatea tehnicii de compresie este măsuratã conform următorilor trei parametri de bazã:

(a) Fidelitatea hãrþii color: cât de multe culori ale filmului original sînt prezente dupã comprimare?

(b) Rezoluþia pixel pe cadru: câte pixele au fost abandonate?

(c) Numãrul de cadre pe secundã: cîte cadre au fost abandonate?

4.6. Scenarii de utilizare a VIDEO-dicþionarului

Dicþionarul MULTIMEDIA al imbii române cuprinde peste 70000 de cuvinte din cele mai diverse domenii. Dicþionarul este conceput atât pentru studenþi cât ºi pentru cercul larg al vorbitorilor limbii române, care doresc să cunoascã sensul propriu care trebuie conferit cuvintelor. Dicþionarul MULTIMEDIA satisface cerinþele de bazã: dã definiþia exactã a cuvântului, ºi, dacã e cazul, genul, numãrul, sinonimele, antonimele, imagini, secvenþe VIDEO ºi AUDIO, care exprimã sensul exact ºi limpede, deplin accesibil, ceea ce constituie partea cea mai importantã de utilizare. Acest dicþionar este una din pietrele de temelie ale culturii tineretului, care va contribui la opera de culturalizare a maselor prin iniþierea în folosirea limbii române informatizatã corectã, exactã ºi unitarã.

Compartimentul VIDEO al acestui dicþionar MULTIMEDIA al limbii române conþine, dupã pronosticurele noastre, peste 12000 cuvinte. Acest compartiment furnizează informaþii necesare referitoare la cuvintele cãutate, secvenþe video ce oferă posibilitatea de a percepe mai bine esenþa cuvintelor. Diviziunea video face dicþionarul mult mai accesibil ºi atractiv utilizatorilor de toate vârstele ºi interesele.

Necesitatea utilizării VIDEO-dicþionarului poate apărea în cele mai diverse situaþii. Să considerăm situaþia, în care un student este nevoit să scrie un referat la merceologia ºi 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.

4.5.1. 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 sau chiar şeful departamentului de producere. Formularea întrebării ar fi: “Găseşte toate secvenţele, în care apare cuptorul”, “Găseşte toate secvenţele, în care apare banda rulantă”, sau “Găseşte toate secvenţele, în care apare şeful departamentului de producere”.

(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”.

4.5.2. Utilizarea bazelor de date VIDEO on diferite domenii.

După cum am menþionat anterior scopul baze de date VIDEO este de a satisface cele mai diverse cerinþe. Astfel aceste BDV îºi gãsesc aplicarea în cele mai diverse domenii.

4.5.2.1. Educaþie. Bazele de datele **VIDEO** au o aplicare largã în educaþie ºi cercetare. Universitãþile pot acorda aºa servicii ca studii la distanþã prin satelit, sau utilizând Internetul. Acestea pot pune la dispoziþia studenþilor un set de casete **VIDEO** cu înregistrări ale cursurilor. Dicþionarul **VIDEO,** fiind ºi el o bazã de date **VIDEO** pune la dispoziþia utilizatorilor secvenþe **VIDEO,** care pot fi utilizate în cadrul comunicărilor, pentru pregătirea unor prezentări, lecþii deschise, rapoarte.

4.5.2.2. Sport. Sãlile de Sãnãtate oferã baze de date, în care sunt înregistrate casete **VIDEO**, care conþin diferite programe de antrenament, utilizatorului oferindu-se posibilitatea de a alege între programe de slãbire, fortificare sau menþinere a condiþiei fizice.

4.5.2.3. Agriculturã. Institutele de cercetãri ºtiinþifice în domeniul agriculturii din þarã ar putea utiliza **VIDEO** dicþionarul pentru a studia mai aprofundat procesul de plantare, condiþiile de creºtere ºi dezvoltare a plantelor, specificul dezvoltãrii plantelor în diferite regiuni sau þãri, aclimatizarea plantelor la condiþiile tãrii în cauzã.

4.5.2.4. Economie. VIDEO-dicþionarul poate fi utilizat în foarte multe domenii ale economiei: finanþe, contabilitate, management, marketing, statisticã, turism. Vocabularul economic cuprinde destul de muþi termeni, care pot fi redaþi printr-un limbaj **VIDEO** mai accesibil atât speciali ºtilor cât ºi utilizatorilor de rând.

V. Concluzii

5.1. Compartimentul TEXT. Dicţionarul economic TEXT al limbii române în forma sa de BDT, ca o subdiviziune a MDLR, are posibilităţile de a fi extins cu caracteristicele respective ale MULTIMEDIA: Imagine, Audio, Video etc. Această BDT va ocupa aproximativ 18 MB memorie. La conferinţa tinerilor savanţi ai ASEM din 4-5 aprilie 2002 în baza cîtorva sute de articole din DEI au fost demonstrate caracteristicele de utilizare prietenoasă a subsistemului TEXT al SGBDMM, utilizînd sistemele Ms ACCESS – 2000, Ms WORD - 2000 şi Ms PowerPoint – 2000 ca componente ale Software-ului Ms OFFICE –2000 şi WINDOWS – 2000, expluatate în baza Harware-ului de tip PC Pentiun II, conectat la reţelele Intranet, Externet şi Internet.

5.2. Subsistemul AUDIO. Subsistemul AUDIO interacţionează cu celelalte subsisteme de nivel unu (TEXT, IMAGINI, VIDEO) ale SGBDMM, care susţine evaluarea Marelui Dicţionar al Limbii Române informatizat cu MULTIMEDIE. Acest subsistem AUDIO susţine toate definiţiile celor 61635 de articole din DEI de comun accord cu subsistemul TEXT al SGBDM. Cele 2320 de ilustraţii din DEI sunt susţinute 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, care sunt individual executate în diferite probleme prin intermediul diferitor suporturi ale Software-ului şi Hardware-ului modern. Baza de date BDA al compartimentului AUDIO-dicţionarului 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 expusã pentru analizare ºi discuþii la Conferinþa tinerelor cercetãtori ai ASEM din 4-5 aprilie 2002 în baza cîtorva zeci de articole din DEI. Mijloacele Software-ului ºi Hardware-ului de tip Ms ACCESS-2000, Ms WORD-2000 ºi

Ms PowerPoint-2000 cu dispozitivele respective al PC-ului Pentium II au fost suficiente la etapa iniþialã pentru a demonstra eficienþa ºi eficacitatea mijloacelor ºi metodelor alese pentru realizarea Proiectului "Limba Românã – Limba a Comunitãþii Europene" de catre grupul de cercetãtori – autori ai acestei publicaþii. Volumul BDI de prezentare în Ms ACCESS-2000 fãrã comprimare a 50 articole din DEI ocupã circa 550 MB memorie.

5.4. Subsistemul VIDEO. Dupã cele menþionate mai sus þinem sã subliniem, cã subdicþionarul **VIDEO** are o utilitate mare pentru persoanele ce opereazã în diferite domenii aºa ca: economia, educaþia, sport, agriculturã, industrie, etc. Avantajul acestui dicþionar este cã putem uºor funcþiona cu el ºi este accesibil pentru toþi. Dicþionarul **VIDEO** este o bazã de date, cu care putem opera oricînd avem nevoie ºi oferã posibilitatea de a percepe o informaþie în formã de videoclipuri. În aºa mod persoanele ce se folosesc de astfel de dicþionar înþeleg mai uºor sensul cuvãntului, care este reprezentat în formã **VIDEO**, fiindcã se formeazã o imagine amplã despre cuvãntul dat ºi este uºor de memorizat. Acasta încã odatã confirmã proverbul: «Mai bine o datã sã vezi, decît de o sutã de ori sã auzi».

5.5. Lucrãri paralele ºi perspective. În paralel cu sistemele de nivel unu sunt elaborate sistemele de nivelul doi, care suportã subdiviziunile MDLR în planurile: TEXT&AUDIO, TEXT&IMAGINI şi TEXT&VIDEO.

Elaborarea sistemului, care suportã în comun toate susnumitele compartimente MULTIMEDIA ale MDLR informatizat, constituie a treia platforma, mai complexã, de experimentari ºi implementãri a dicþionarelor computerizate în cadrul elaborãrii MDLR informatizat [17].

Rezultatele evaluãrii preventive a primelor din aceste trei platforme: sistemele unare TEXT, AUDIO, IMAGINI ºi VIDEO au dat posibilitate de a face unele concluzii de evaluare a MDLR informatizat ca o parte componentã a cercetãrilor, fãcute în cadrul Proiectului «Limba românã – limbã a Comunitãþii Europene», care evaluiazã în perioada 2000-2006. Acest Proiect a fost iniþiat [10-11] de cãtre Forumul Internaþional din Chiºinãu,14-15 aprilie anul 2000. Proiectul constituie unul din subiectele de cercetãri, experimentãri ºi evaluãri, efectuate în cadrul Consorþiului Uniunii Latine «Pentru limba românã», a Consorþiului «Pentru informatizarea limbii române» ºi a Comisiei Academiei Române «Pentru informatizarea limbii române».

O serie de aplicaþii a MDLR computerizat este evidenþiatã în [13-16].

Referinţe bibliografice

[1] V. S. Subrahmanian. Principles of Multimedia Database Systems. // Morgan Kaufman Publishers, Inc., San-Francisco, California, USA, 1998, -pp. 442.

[2] D. Todoroi, S. Nazem, T. Jucan, D. Micusha. Transition To A Full Information Society: Stage Development. // Working Paper No. 98-2, UNO, Omaha, USA, March 1998. - 38 p.

[3] D. Todoroi, D. Micuºa, V. Clocotici, I. Linga, V.Tapcov, N. Drucioc, A. Calcatin, M. Morari. **Data Bases** and Communications Tools. Ms ACCESS – 200. // Ed. ASEM, Chisinau 2002, 337 pages. (Eng.)

[4] Dumitru N. Todoroi, Zinaida Todoroi, Diana Micusa. Romanian Computerized Language – One of the European Community Languages. // Proceedings of the 26th Annual Congress of the American Romanian Academy of Arts and Sciences (ARA), Montreal, Quebec, Canada, July 25-29, 2001, pp. 133-137. (Rom)

[5] Diana D. Micusha, Dumitru Todoroi. Natural language processing at the transition to a full information society initial development phase. Part 1. // Studii ºi cercetãri economice. Vol. XXX. Lucrãri prezentate la Sesiunea jubiliarã de comunicãri ºtiinþifice : «Creºtere economicã, dezvoltare, progres», Cluj-Napoca, 2001, pp. 1396-1413.

[6] Diana D. Micusha, Dumitru Todoroi. Natural language processing at the transition to a full information society initial development phase. Part 2. // Studii ºi cercetãri economice. Vol. XXX. Lucrãri prezentate la Sesiunea jubiliarã de comunicãri ºtiinþifice : «Creºtere economicã, dezvoltare, progres», Cluj-Napoca, 2001, pp. 1414-1427.

[7] Sabin-Corneliu Buraga, Dumitru Todoroi. Adaptabilitatea informaþionalã ºi operaþionalã. // Studii ºi cercetãri economice. Vol. XXX. Lucrãri prezentate la Sesiunea jubiliarã de comunicari ºtiinþifice : «Creºtere economicã, 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, Bucuresti, 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, Chisinau, pp. 12-15**.**

[11] Dan Crisrea, Dumitru Todoroi, Dan Tufis. Computational Linguistic: Romanian Language - One of the European Community Languages. // Proc. of the Intern. Sc. Seminar "Strategies and Modalities for Romania and Moldova' European Integration", 28-29 Sept. 2000, V.2, ASEM, Chisinau, pp.276-280.

[12] D. Todoroi, D. Micusa, V. Clocotici, S. Pereteatcu, V. Bordeianu, C. Grigoras, S. Cretu, I. Linga, S. Spataru. Natural Language Processing: IEE-2000 Project. // Proc. of the Intern. Sc. Seminar "Strategies and Modalities for Romania and Moldova' European Integration", 28-29 Sept. 2000, V.2, ASEM, Chisinau, pp.281-285.

[13] Stefan Spataru, Dumitru Todoroi. Distance Education Via Internet, Multimedia and modern System Environment. // Proc. of the Intern. Sc. Seminar "Strategies and Modalities for Romania and Moldova' European Integration", 28-29 Sept. 2000, V.2, ASEM, Chisinau, pp. 307-312.

[14] Ion LINGA. IMPACTUL IMPLEMENTARII COMPUTERULUI ASUPRA PROCESULUI DE ASIMILARE A CUNOSTINTELOR. // Proc. Of the 27th ARA Congress, May 29 – Lune 2, 2002, Oradea, Romania.(To be published).
[15] Ion COVALENCO. Metode adaptabile de evaluare a cuniºtinþelor asistatã de calculator. // Proc. Of the 27th ARA Congress, May 29 – Lune 2, 2002, Oradea, Romania. (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, Romania.(To be published).
[17] Dumitru N. TODOROI, ASEM, Chisinau, Nicolae MARGINEANU, L'Ecole Politechnique, Montreal, Canada.THE ROMANIAN LANGUAGE'MULTIMEDIA – DICTIONARIES IMPLEMENTATION ENVIRONMENT AT THE FULL INFORMATION SOCIETY INITIAL 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 limbajului natural computerizat. // Proc. Of the 27th ARA Congress, May 29 – Lune 2, 2002, Oradea, Romania.(To be published).
[19] Zinaida TODOROI, ULIM, Chisinau, Eugenia MARGINEANU, L'Ecole Politechnique, Montreal, Canada. MULTIMEDIA – dictionaries for Romanian Language. Usage Scenarios on the EAPEC Base. // Proc. Of the 27th ARA Congress, May 29 – Lune 2, 2002, Oradea, Romania.(To be published).
[20] Societatea informaþionalã – Societatea cunoaºterii. Concepte, soluþii ºi strategii pentru România. // ACADEMIA ROMÂNÃ, Editura EXPERT, Bucureºti, decembrie 2001. – 541 pages.
[21] Dicþionar Enciclopedic Ilustrat (DEI). // Editura CARTIER SRL, Chiºinău, Editura CODEX SRL, Bucureºti, 1999, 1808 pages.
[22] Beyond Calculation : The Next Fifty Years of Computing. // Edited by Peter J. Denning and Bob Metcalfe, Copernicus, 1997 Springer-Verlad New York, Inc., 350 pages.

Comunicări la Conferința tinerilor cercetători ASEM, 4-5 aprilie 2002, Chișinău.
Coordonator: Dumitru TODOROI, Prof. Univ., doctor habilitatus.

[23]. AMBROZII Nadejda, GODZINA Irina. Componenta Text a Audio Dicționarului Economic al Limbii Române.
[24]. ȚURCANU Virginia, MUTRUC Carolina. AUDIO-DICȚIONARUL EXPLICATIV ECONOMIC AL LIMBII ROMÂNE
[25]. COZLOV Elena, BABANU Irina. Subsistemul IMAGE al dicționarului economic informatizat al limbii române.
[26]. LUNGU Stela,CIOBANU Diana, GUZUN Oxana. VIDEO-dicționarul economic 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 Romania. The Electronic Linguistic Atlas has features of a multimedia application allowing the user to consult and/or print 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 nuanţele de rostire, în lingvistică se recurge (după practica internaţională) la transcrierea fonetică. Pe lângă transcrierea fonetică internaţională realizată cu Alfabetul Fonetic International (IPA), fiecare ţară îşi are propriile simboluri grafice [1], [2]. Pentru limba română, aceste simboluri sunt realizate 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[98]:
  - 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 (*ALR_Baza*) 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 un fenomen fonetic sau mai multe am realizat familii de fonturi ale căror denumiri le-am dat cu ajutorul fenomenelor fonetice aplicate (ex. *ALR_Semivocale, ALR_Nazalizate, ALR_Seminazalizate, ALR_ScurteNazalizate, ALR_ ….* etc.). Această organizare a fonturilor a fost facută 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.

---

[98] *Formularea "sunete primare", inexactã din punctul de vedere fonetic, este folositã cu înţelesul "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

Vocalele primare folosite în transcrierea fonetică sunt:

simple	diacritice					
a	ä	ă	â	a[99]	Í	
e	ë					
i			î			i
o	ö					
u	ü		ű			

Cu ajutorul acestor "vocale primare" şi al celor trei variante accentuate (a - á A Á) ale fiecăreia dintre ele se obţine seria completă de sunete vocalice care se regăsesc în fontul de bază *ALR_Baza* (17*4=68 semne).

Transformările fonetice care pot modifica cele 17 vocale de bază (68 împreună cu variantele lor accentuate), sunt clasificate în următoarele grupe de fenomene disjuncte[100]:

Grupe	Poziţionare	Notaţie	Fenomen	Exemplu
Durată	Aşezat cel mai sus	(a)	Scurtime	e é E É
		(b)	Semilungime	e é E É
		(c)	Lungime	e é E É
Nazalizare	Aşezat deasupra dar sub a-c	(d)	Seminazalizare	e é E É
		(e)	Nazalizare	e é E É
Ocluzie glotală	aşezat "în umăr", în faţa	(f)	Coup de glotte	e é E É
Deschidere	Aşezat imediat sub vocală	(g)	Închidere	e é E É
		(h)	Semideschidere	e é E É
		(i)	Deschidere	e é E É
		(j)	Deschidere mare	e é E É
Afonizare	Aşezat sub vocală dar şi sub g-j	(k)	Semiafonizare	e é E É
		(l)	Afonizare	e é E É

Prin transformări fonetice se înţeleg toate realizările vocalice obţinute ca urmare a aplicării a cel puţin unui fenomen fonetic (*a)–(l*) asupra vocalelor primare. Din punct de vedere lingvistic sunt impuse următoarele reguli:

Regula [1]

- vocalele a, ä, Í - deschise prin natura lor (cu cel mai mare grad de apertură) - nu pot contacta fenomenele fonetice *h* (semideschidere), *i* (deschidere) şi *j* (deschidere mare);

[99] *Sunetul a nu trebuie marcat cu alt fenomen fonetic*
[100] *Fenomene din aceeaşi grupă nu pot fi aplicate simultan asupra unui sunet*

- vocalele i, □, î, u, ü, ű – închise prin natura lor (cu cel mai mic grad de apertură) - nu pot contacta fenomenul fonetic *g* (închidere).

Sunt excluse deci variantele vocalice a, A, a, ä, ä, ä, Í, Í, Í, ca şi variantele vocalice i, i, î, u, ü, ű. Aceste *15 grafeme*4=60* se scad din cele 17*4*12=816. Astfel prin asocierea vocalelor primare cu câte un fenomen *a-l* apar 756 imagini grafice repartizate în 12 fonturi astfel grupate (convenţional dar extrem de uşor de ţinut minte) 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". Fiecare transformare fonetică exclude prezenţa celorlalte transformări din aceeaşi grupă. În aceste condiţii combinaţiile de câte două sau mai multe fenomene sunt posibile doar între membrii a două grupe diferite.

În plus, cele *15 grafeme*4=60* excluse ca urmare a restricţiei formulate sub **Regula 1**, nu pot participa la combinaţiile de de două, trei, patru fenomene.

1.2. Fenomene fonetice aplicate consoanelor primare

Consoanele primare folosite în transcrierea fonetică sunt:

b, c,□,□,□,□,□, d,π,D,∩, f, g,□,□,□,□, h,ᒍ,○,●, j, k, l,≈,

m,M, n,N,℘,ᒍ, p, r,∝,∉,ρ, s,∇,·,ş,σ, t,v,ţ, v, w, z,|,ʃ, y

Fenomenele fonetice care pot fi aplicate consoanelor primare sunt:

Grupe	Notaţie	Fenomen
Durată	1	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 plus, consoanelor primare le pot fi aplicate numai cel mult două transformări şi numai în anumite combinaţii. În tabelul următor sunt prezentate combinaţiile posibile de fenomene şi consoanele pe care acestea le pot însoţi.

1.2.1. Consoane cu un singur fenomen fonetic:	
semilungime	∩, f , h, □, ○,●, j, l, ≈, m, M, n, N, ℘, ⅃, r, ⊄, ∉, ρ, s, ∇, ·, ş, v, w, z, ⏐, ⌠, y
lungime	∩, f , h, □, ○,●, j, l, ≈, m, M, n, N, ℘, ⅃, r, ⊄, ∉, ρ, s, ∇, ·, ş, v, w, z, ⏐, ⌠, y
semipalatalizare	d, g, h, ○, k, j, l, n, r, ş, t
palatalizare	d, g, h, k, j, l, n, r, ş, t
palatalizare mare	t, d
explozie	c, p, t
caracter silabic	l, m, n, r, s, M
semiafonizare	b, d, π, D, ∩, g, □, □, □, □, □, ●, j, l, ≈, m, M, n, N, ℘, ⅃, r, ⊄, ∉, ρ, v, w, z, ⏐, ⌠, y
afonizare	b, d, π, D, ∩, g, □, □, □, □, □, ●, j, l, ≈, m, M, n, N, ℘, ⅃, r, ⊄, ∉, ρ, v, w, z, ⏐, ⌠, y
1.2.2. Consoane cu două fenomene fonetice:	
semilungime + semipalatalizare	h, ○, j, l, n, r, ş
semilungime + palatalizare	h, j, l, n, r, ş
semilungime + caracter silabic	l, m, M, n, r, s
semilungime+semiafonizare	∩, □, ●, j, l, ≈, m, M, n, N, ℘, ⅃, r, ⊄, ∉, ρ, v, w, z, ⏐, ⌠, y
semilungime+afonizare	∩, □, ●, j, l, ≈, m, M, n, N, ℘, ⅃, r, ⊄, ∉, ρ, v, w, z, ⏐, ⌠, y
lungime + semipalatalizare	h, ○, j, l, n, r, ş
lungime + palatalizare	h, j, l, n, r, ş
lungime + caracter silabic	l, m, M, n, r, s
lungime+semiafonizare	∩, □, ●, j, l, ≈, m, M, n, N, ℘, ⅃, r, ⊄, ∉, ρ, v, w, z, ⏐, ⌠, y
lungime+afonizare	∩, □, ●, j, l, ≈, m, M, n, N, ℘, ⅃, r, ⊄, ∉, ρ, v, w, z, ⏐, ⌠, y
semipalatalizare+semiafonizare	d, g, j, l, n, r
semipalatalizare+afonizare	d, g, j, l, n, r

palatalizare+semiafonizare	d, g, j, l, n, r
palatalizare+afonizare	d, g, j, l, n, r
Palatalizare mare+semiafonizare	D
palatalizare mare+afonizare	D
Explozie+semiafonizare	b, d, g
Explozie+afonizare	b, d, g
Caracter silabic+semiafonizare	l, m, M, n, r
Caracter silabic+afonizare	l, m, M, n, r

2. Mediu pentru editarea transcrierilor fonetice

Interfața realizată pentru editarea transcrierilor fonetice poate fi folosită în mai multe moduri:

- editarea dicționarelor Atalasului Lingvistic;
- editor stand-alone sau ca aplicație de tip server pentru inserarea de obiecte de tip "transcriere fonetica" 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).

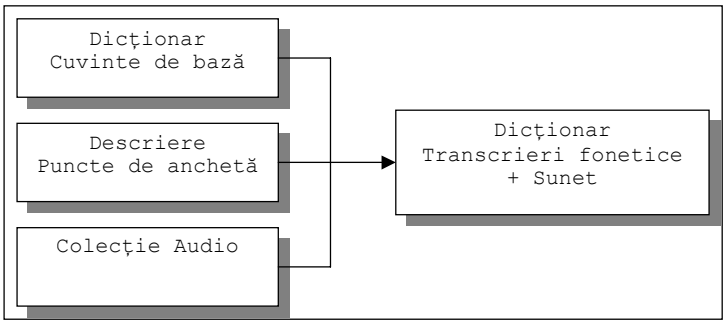


Figura 1. Conexiunile dintre informațiile stocate în dicționarele ALR

Dicționarul "Cuvinte de bază" conține fondul de cuvinte (titlul hărților) din atlasul lingvistic electronic, întrebările care au fost puse la anchetă, note, observații, şi

eventual imagini. Pentru fiecare cuvânt este indicată şi întrebarea corespunzătoare care este pusă în momentul interviului.

În momentul completării acestui dicţionar, utilizatorul poate vedea lista completă a cuvintelor de bază introduse, le poate sorta după diferite criterii, poate modifica articolele introduse anterior, după cum este prezentat în figura 2.

Dicţionarul "Puncte de anchetă" conţine informaţii (cod, nume, observaţii) despre punctele de anchetă prezentate în cadrul atlasului lingvistic. La fel ca la dicţionarul anterior, şi aici, utilizatorul poate vedea lista completă a punctelor de anchetă introduse, le poate sorta după diferite criterii, poate modifica articolele introduse anterior.

2.1. Dicţionar transcrieri fonetice

Dicţionarul de transcrieri fonetice conţine transcrierea fonetică a răspunsului la întrebarea pusă în etapa de interviu pentru fiecare cuvânt din **Dicţionarul Cuvinte de bază** în fiecare din **Punctele de anchetă**, iar acolo unde este posibil şi înregistrarea audio corespunzătoare din **Colecţia Audio**.

Pentru claritatea hărţilor lingvistice, răspunsurile din punctele de anchetă sunt însoţite de note şi comentarii (figura 3).

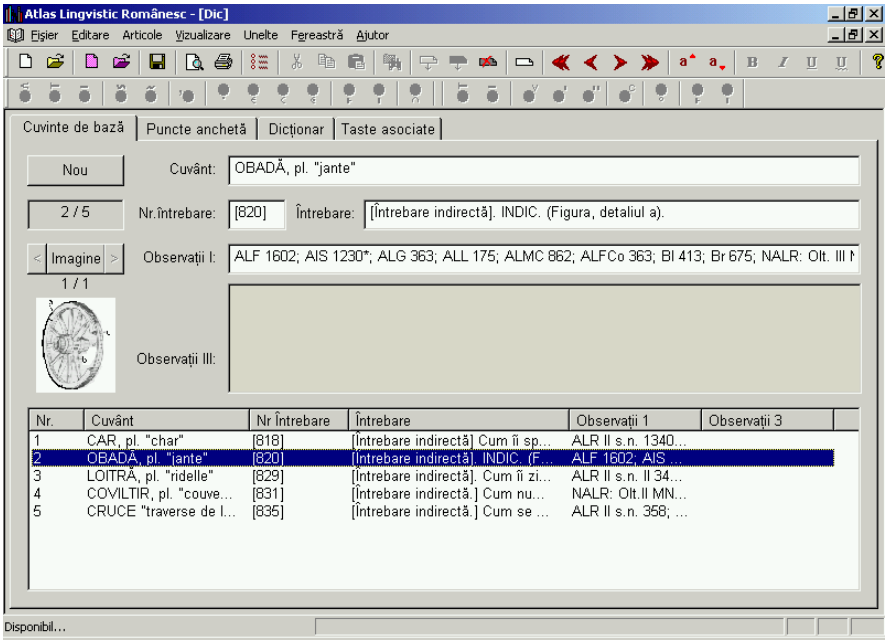


Figura 2. Fereastra de editare a listei cuvintelor de bază

Pentru transcrierea fonetică a cuvintelor din atlasul lingvistic românesc este folosit un număr mare de fonturi, rezultat din numărul mare de combinații posibile ale fenomenelelor fonetice prezentate in 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 şi 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 **CAlrString**. Acestea sunt de fapt şiruri de obecte de tip **CAlrChar**, care la rândul lor au următoarea descriere:

- caracterul corespunzător sunetului primar (pe 16 biți, codificare UNICODE);
  - atribute:
    - poziționare: normal, deasupra sau „în umăr”¨;
    - mod subliniere: linie sau zigzag;
    - cursiv;
    - îngroşat;
  - 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 de fonturi a aplicației, în momentul afişării.

În momentul în care este deschis dicționarul 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 stocate pentru Atlasul Lingvistic Român, descrierea fiecărui cuvânt este compresată folosind un algoritm de compresie LZW. La selecția unui cuvânt de bază, descrierea sa este decompresată în memorie şi 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, la care operatorul trebuie să parcurgă următorii paşi:

- selectează cuvântul titlu;
- selectează punctul de anchetă;

- editează transcrierea fonetică, nota şi comentariul asociat cuvântului pentru punctul de anchetă respectiv.

La editarea transcrierilor fonetice trebuie avute în vedere două aspecte:

- selectarea sunetului primar;
- selectarea fenomenelor asociate.

Selectarea sunetului primar se face prin apăsarea tastei respective, dacă sunetul are un corespondent pe tastatură, sau prin apăsarea unei combinaţii de taste, dacă sunetul nu are corespondent pe tastatură. Combinaţiile de taste sunt prestabilite în aplicaţie (la stabilirea combinaţiilor de taste am folosit recomandările de la Microsoft Word), şi cel puţin deocamdată nu pot fi modificate de utilizator. Pentru a veni în ajutorul celui ce editează dicţionarul, aplicaţia dispune de o fereastră în care sunt afişate combinaţiile de taste predefinite.

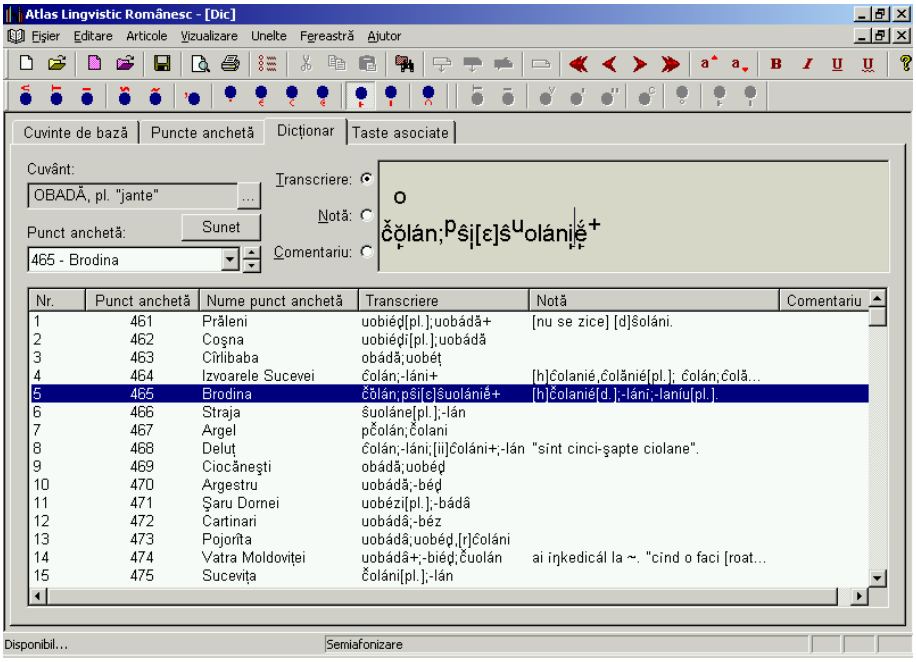


Figura 3. Editarea Dicţionarului de transcrieri fonetice

Pentru selectarea fenomenelor asociate sunetelor, aplicaţia prezintă 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 grupărilor 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.

Dictionarul cu transcrieri fonetice poate conține înregistrările audio în format WAV ale răspunsurilor la întrebările din anchetă pentru cuvintele incluse în dicționarul lingvistic. Acest lucru este posibil dacă la realizarea anchetei pentru atlasele lingvistice se face şi înregistrarea pe bandă a răspunsurilor.

3. Realizarea Atlasului Lingvistic Român pe Regiuni

Sistemul software care modelează atlasul lingvistic electronic, conține module care realizează gestionarea următoarelor grupe de informații:

- simboli 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;

Aceaste componente sunt prezentate în figura 4.

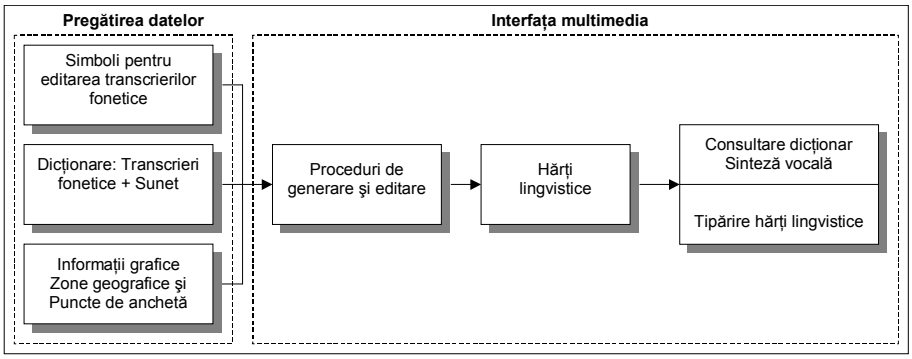


Figura 4. Componentele Atlasului lingvistic Electronic

În acest capitol ne vom referii la facilităţile pe care le oferă modulul software care realizează generarea şi consultarea hărţilor ligvistice. În etapa de proiectare a acestui modul, am avut în vedere modelul interfeţelor multimedia. Din această analiză a rezultat necesitatea existenţei unui modul care să permită:

- generarea unei hărţi noi pe baza informaţiilor din dicţionarele ALR şi a informaţiilor grafice primare cuprinse în fişiere DXF.
- editarea: aranjarea în pagină, selectarea informaţiilor care vor fi vizibile implicit;
- salvarea într-un fişier numit "hartă lingvistică" a selecţiilor şi modificărilor din faza de editare;
- consultarea atlasului electronic:
  - vizualizarea şi ascultarea informaţiilor din punctele de anchetă.
  - tipărirea hărţilor lingvistice;

3.1. Modulul pentru generarea şi editarea hărţilor lingvistice

Pentru reprezentarea hărţilor lingvistice s-au proiectat structuri de date bazate pe obiecte, suficient de flexibile pentru a permite dezvoltări ulterioare. În cele ce urmează vom face o trecere în revistă a pricipalelor structuri de date realizate.

Pentru organizarea informaţiilor grafice primare, am avut în vedere cerinţele impuse de tehnologia de realizare a atlaselor lingvistice. Astfel am apelat la formatul DXF care permite o organizare a obiectelor primare pe straturi. Am realizat fişierul NALRB.DXF care conţine următoarele tipuri de obiecte (straturi):

chenare – limitele paginii şi chenarele hărţii;

frontiere – conturul zonei studiate (Moldova şi Bucovina);

mijloc – indică locul de pliere al hărţii, la legarea în volum;

municipii	– localităţile importante afişate pe hartă;
puncte anchetă	– dreptunghiurile în care se scriu codurile punctelor de anchetă;
transcriere fonetică	– conţine dreptunghiuri pentru transcrierea fonetică;
note	– dreptunghiuri cu poziţiile predefinite pentru Titlu, Nota I, Nota II, Nota III;
zone	– delimitări zonale în jurul punctelor de anchetă

Pentru editarea şi salvarea hărţilor lingvistice din conţinutul ALR, s-a creat o structură de date flexibilă, care să permită în viitor, extinderea editării asistate de calculator a atlaselor lingvistice româneşti regionale la nivel naţional. Astfel, a rezultat o structură de date numită "harta lingvisticγ" de forma următoare:

- header fişier;
- listă cu descrieri obiecte;

Descrierile de obiecte au un antet care este comun pentru toate tipurile de obiecte şi un corp obiect specfic fiecărui tip în parte. Obiectele pot fi simple sau compuse. Un obiect compus conţine la rîndul lui alte obiecte simple sau compuse.

Au fost definite următoarele tipuri de obiecte:

- Text;
- AlrString (obiectul a fost definit pentru editarea dicţionarului cu transcrieri fonetice);
- Dreptunghi;
- Harta cu transcrierile fonetice;
- Harta sintetică (lingvistică sau fonetică);
- Notă referitoare la continuarea transcrierile 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 ;
- Bitmap;
- Strat DXF;

3.2. Modulul pentru consultarea atlasului electronic

Componenta pentru **consultarea atlasului**, permite încărcarea unei hărţi lingvistice generate - editate în etapa anterioară. Sistemul va afişa harta regiunii respective

(în situația studiată este vorba de Moldova şi Bucovina), pe care va plasa transcrierea fonetică a răspunsurilor din punctele de anchetă împreună cu notele şi observațiile introduse anterior în dicționarele ALR sau de operatorul care a realizat harta (figura 5).

Dacă utilizatorul doreşte să analizeze un cuvânt pentru care nu sa realizat în prealabil o hartă, dar care există în dicționarele ALR se realizează generarea automată a hărții pe care se vor plasa transcrierile fonetice a răspunsurilor din punctele de anchetă împreună cu notele şi observațiile introduse anterior în dicționarele ALR. La activarea modului, apare prezentată harta regiunii cu punctele de anchetă şi numele localităților pe care acestea le reprezintă, după care poate fi selectat cuvântul de bază dorit.

După ce harta lingvistică a fost încarcată / generată prin selecția unui punct de anchetă, este posibilă şi redarea audio corespunzătoare transcrierii fonetice asociate acestuia (înregistrarea audio sau cuvântul sintetizat).

Tot cu ajutorul aceastei componente se realizează tiparirea automată a hărților atlasului lingvistic românesc, în vederea includerii lor în volum (figura 6).

Pentru tipărirea hărților au fost prevăzute următoarele facilități:

- posibilități de selectare a informațiilor ce se vor tipări;
- tipărirea pe o pagina sau tipărirea pe două pagini cu respectarea locului de pliere al hărții indicat prin linia “mijloc”;

Dacă utilizatorul doreşte tipărirea într-un mod sintetic (fără hartă), modulul poate asigura crearea unor pagini de tip MN (Material Necartografiat). Folosind această opțiune, va fi tipărită numai lista cu transcrierile fonetice corespunzătoare cuvântului selectat, ordonate dupa criterii de similaritate.

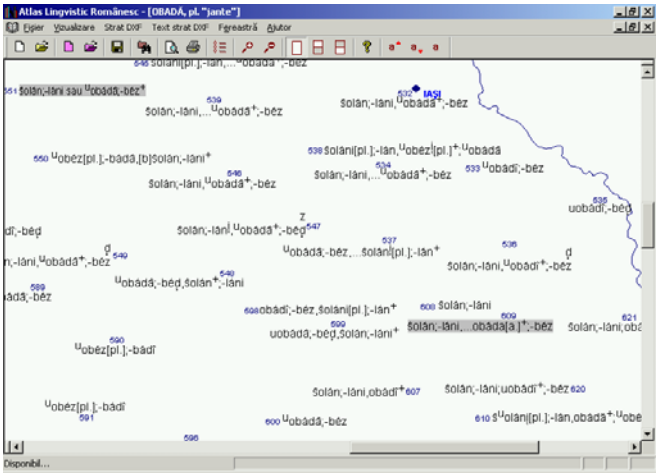


Figura 5. Fereastra de editare – consultare a Atlasului Lingvistic

4. Concluzii

Organizarea prezentată pentru simbolurile grafice facilitează editarea de texte cu transcrieri fonetice şi cu alte editoare de texte care permit folosirea de fonturi multiple în text.

Modul de selectare a fonturilor folosit la editarea transcrierilor fonetice poate fi extins pentru crearea unei aplicații de tip client-server sau la realizarea unui editor simplu, de tip WordPad.

Realizarea acestui sistem de editare a transcrierilor fonetice este în curs de testare şi finalizare. În continuare, ne propunem adăugarea de noi opțiuni şi facilități, care să permită transformarea sistemului într-un instrument util cercetătorilor lingvişti.

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, Iasi, 6 oct. 2000.

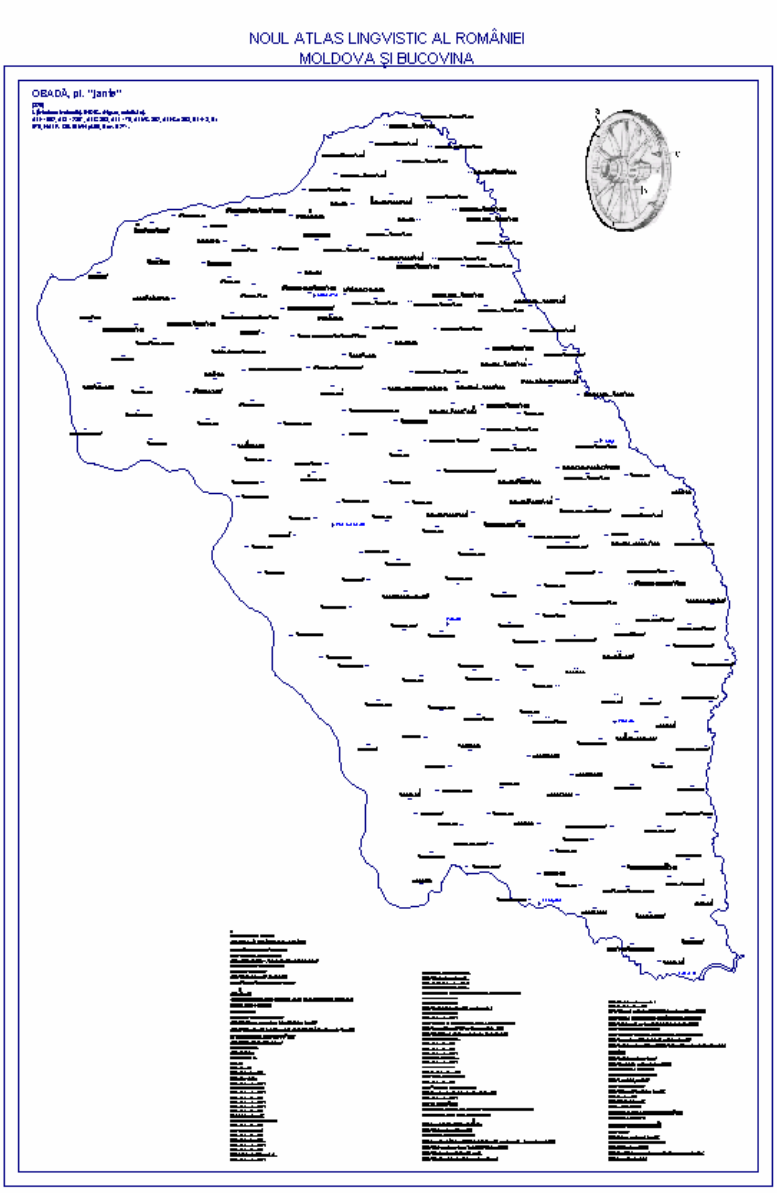


Figura 6. Imaginea unei pagini tipărite din modulul de consultare

SECȚIUNEA VI

Dezbateri şi discuții

Asupra a doi vectori funcționali ai societății cunoaşterii: managementul cunoaşterii şi învățarea electronică. Cultura şi societatea cunoaşterii

Mihai DRĂGĂNESCU
Institutul de Inteligență Artificială
Academia Română

Introducere generală

Acest material, care consituie 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 interimară 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 $(IN)_{structural}$ 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 = $(IN)_{structural}$ se va petre, 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 =$(IN)_{structural}$, automat un asemenea creier electronic va avea şi proprietățile fenomenologice ale intuiției, creativității şi spiritualității. Ceea ce nu credem.

Din momentul în care ce IA > $(IN)_{structural}$ 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 apare 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

I.1 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 *functional.*

Technological vectors are the extended Internet, the e-book and e-document technology, artificial intelligence (with intelligent agents and future Networked Systems of Embedded Computers), nanotechnology and others.

Among the functional vectors of the Knowledge Society, a group of vectors is related to knowledge management:

- *knowledge management for corporations and enterprises, organizations and institutions,* ***local and national administrations****;*
- *the management of the moral use of scientific knowledge at the global level;*
- *e-learning management;*
- *development of a culture of knowledge and innovation;*
- *management of the scientific and technological knowledge for every main domain of activity as health care, sustainable society and others.*

I.2 Knowledge management

The problem of management with respect to knowledge is regarded in two ways:

I. As the management of the organization busy with the use and integration of various types of knowledge;

II. As the management of knowledge itself, for generation of new knowledge, for discovering existing knowledge (tacit or very local, or external to the organization), for combining available knowledge.

Perhaps, what is really needed is a general vision, in a unity, on the management of the organization and the management of knowledge.

Knowledge management for enterprises, organizations, institutions, local and national administrations

In the western literature, in the last years, were elaborated a series of works dedicated to the problems of enterprises and knowledge. In Romania we do not have yet specialists in knowledge management in the context of the knowledge society. We do not have either a knowledge society, but we need experts in knowledge management for building the future knowledge society. A group of members of the Romanian Academy and other wellknown specialists in information technology from Romania and colleagues from USA decided to constitute a Romanian-American Foundation for the Knowledge Society, one of the main aims being to educate in USA a number of young Romanian specialists in the new domain of knowledge management. All is ready for such a Foundation, the contributions of the individual founders are also ready, but no institution or 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 substantial 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 national and 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 the contents of information, especially of knowledge, but these would not be efficient without information technology. The technological vectors of the Knowledge Society are equally important as the functional vectors.

I.3 Points of view for practical knowledge management

Knowledge management is both the management of the organization to use knowledge and the management of all knowledge possible, from inside and outside the organization, to attain the objectives of that organization. Because knowledge is a

special form of information, information technology has to play an essential role in knowledge management. Knowledge and IT are, without any doubt, going hand in hand and have a synergetic effect on the efficiency of organizations.

Lucy Marshall [6] considers that knowledge management refers to the control and utilization of the intellectual capital in an organization. For Lucy Marshall, not the information, but knowledge is the most important asset of an institution. This author recommends a **Chief Knowledge Officer** for an institution, who based on the Intranet of the institution, has to assure the discovery and creation of knowledge in the institution.

Rooney and Mandeville considers the knowledge management at the national level. The abstract of their paper is quoted [7] below:

'As the global economy becomes more knowledge intensive and the wealth of nations more dependent on their knowledge assets being harnessed, it is essential for policy makers of having frameworks for the development and the utilization of national knowledge assets. This article argues that a policy framework can be developed through which policy initiatives in a range of policy areas can be filtered in order to meet the challenges of the knowledge economy. We have developed an approach that has previously been applied to managing intellectual capital in firms and adapted it to the public policy arena. In doing so we question policy orthodoxies such at the assumption that free trade automatically facilitates international knowledge flows, that participation in a global knowledge economy necessarily challenges national sovereignty, and that online delivery of education is necessarily a progressive strategy'.

Peter Drucker (a wellknown professor of social science at Claremont Graduate School and the author of more than thirty books, his most recent book is Management Challenges for the 21st Century, 1999) writes [8] about the **knowledge worker:**

'I am convinced that a drastic change in the social mind-set is required - just as leadership in the industrial economy after the railroad required the drastic 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 traditional 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 traditional "employees" and be treated as such would be tantamount to England's treating its technologists as tradesmen - and likely to have similar consequences.'

I.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 national 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.

I.4 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 knowledge. For science and society, knowledge management will become the most important administration.

References

[1] Mihai Drăgănescu, *Societatea Informațională si a Cunoaşterii. Vectorii Societății Cunoaşterii* (Information Society and Knowledge Society.Vectors of the Knowledge Society)**,** Romanian Academy, July 2001. On the Web, http://www. academiaromana.ro/pro_pri/

[2] Mihai Drăgănescu, Cunoasterea în secolul XXI (Knowledge in the XXI century), communication at the Annual Conference of the Romanian Committee for the History and Philosophy of Science, Romanian Academy, Bucharest, 15 October 2001, to be published.

[3] Mihai Drăgănescu, *The Interdisciplinary Science of Consciousness* (Chapter 5) pp. 46-59, in *Science and the Primacy of Consciousness, Intimation of a 21stCentury Revolution*, Richard L. Amoroso a.o, (eds.), Orinda, California: The Noetic Press, 2000.

[4] See http://www.uts.edu.au/fac/hss/Departments/DIS/km/introduct.htm#Char

[5] Ernst Ulrich von Weiszäcker, Amory B. Lovins, L.Hunter Lovins, *Factor patru. Dublarea prosperității prin înjumătățirea consumului de resurse,* Raport pentru Clubul de la Roma, traducere din limba germană (FAKTOR VIER.Doppelter Wohlstand - halbierter Verbrauch, München, 1995), Bucureşti, Editura tehnică, 1998.

[6] Lucy Marshall, *Facilitating knowledge management and knowledge sharing: New opportunities for information professionals,* Online. 21(5): 92-98. 1997 Sep/Oct.

[7] David Rooney and Thomas Mandeville, *The Knowing Nation: A Framework for Public Policy in a Post-industrial Knowledge Economy*, Prometheus 16 (4) pp. 453-467, 1998.

[8] Peter F. Drucker, *Beyond the Information Revolution,* The Atlantic Monthly, Digital Edition, 1999, http://www.theatlantic.com/issues/99oct/9910drucker3.htm

[9] Menas Kafatos, Mihai Drăgănescu, *Preliminaries to the philosophy of integrative science,* e-book, MSReader format, Academy of Scientists - Romania, Bucharest, 2001, (available free by e-mail: dragam@racai.ro).

II. ÎNVĂȚĂMÂNTUL ELECTRONIC ŞI SOCIETATEA CUNOAŞTERII

II.1 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 societatea 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.

II.2 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 stabilit formatul ediției de cărți electronice de interes

public.Ca urmare a simpozionului a fost înfiinţată şi o librărie de software, cărţi şi documente electronice la Institutul Naţional de Cercetare-Dezvoltare în Informatică (http:// www.e-librarie.ro).

Despre cartea electronică, şi rolul ei pentru societatea cunoaşterii în România, am expus consideraţiile mele în lucrări anterioare [1b], [7] şi nu voi reveni asupra lor. În schimb, voi cita doi autori, unul care a exprimat opinii înainte de apariţia cărţii electronice propriu-zise, altul care a participat la lansarea cărţii electronice. Primul este Paul Saffo, directorul unui elevat Institut al Viitorului din California, care lucrează, foarte scump, numai pentru marile companii americane şi care în anul 1988 prevedea că o carte electronică va fi mai mult decât o carte tipărită datorită posibilităţilor de a introduce elemente audio, video, conexiuni la informaţii pe reţea. El scria [8]:

' The term "electronic book" is misleading because these products are not books at all, but something new. We are living in a moment between two revolutions: one of print, four centuries old and not quite spent and another of electronics, two decades young, and just getting underway. Today's products amount to a bridge between these two revolutions…'

Al doilea este Dick Brass, Vicepreşedinte Microsoft pentru dezvoltare tehnologică, care în anul 2000, an în care cartea electronică propriu-zisă decola, scria [9]:

'If you don't think eBooks will take off, remember that electronic encyclopedias have already outsold all paper encyclopedias. […] They cost less than $100, instead of the $2,000 or more for fine paper encyclopedias. […] Similarly, after the triumph of eBooks, paper books will no longer be the principal means of distributing informatian. But, like horses they will contiue to exist for pleasure…[…] Like all transitions, the move from pBooks to eBooks will be a little painful and tentative at first. Then, in less than 20 years, eBooks will be so pervasive that we won't be able to remember living without them. […] We are on the verge of the most exciting change to the printed word since movable type…'.

Cartea electronică a decolat. Firme precum Amazon şi Barnes and Noble din SUA sunt cunoscute în întreaga lume pentru modul în care au promovat-o. Ele sunt o adevărată şcoală pentru toţi cei care conduc şi vor conduce librării de cărţi electronice şi software, şcoală accesibilă gratuit prin simpla experimentare prin Internet pe web-site-urile acestor firme.

II.3 Procesul de învăţare

În anul 1988 scriam despre procesul de învăţare [10]:

'Înţelegerea profundă a procesului de învăţare depinde de explicarea funcţionării creierului şi a minţii omului, în ultimă instanţă de înţelegerea naturii materiei vii. Cu alte cuvinte, natura intimă a procesului de învăţare nu va putea fi elucidată într-o măsură într-adevăr mulţumitoare decât atunci când ştiinţa va face un nou mare pas în cunoaşterea materiei. Cercetările din domeniile fizicii şi biologiei, esenţiale pentru elucidarea naturii materiei vii, se vor îmbina cu cele din domeniul ştiinţei informaţiei. 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]:

COGNIȚIA:

Anii 1970 ºi 1980 (modelare simbolicã tip calculator)	***Anii 1990*** (efectul conectivismului, reþele neuronice ºi neurale)	***Anii 2000. Ce va urma?*** Efectul ºtiinþei 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 profesor şi le integrează cu cunoştințele anterioare).
- *Învățarea prin analogie.*
- *Învățarea din exemple prin detecție de similarități,* proces esențialmente inductiv (fără a exclude şi procese deductive) prin generalizarea exemplelor pozitive, generalizare care evită exemplele negative.

- *Învățarea prin observare şi descoperire* (spre exemplu a unor regularități în structurări de date).

Gh. Tecuci înclină către o *îmbinare de metode de învățare*. Fără îndoială câteva lucruri credem că se susțin pentru procesul uman de învățare:

- ***Necesitatea unei varietăți de metode,* şi nu o monometodă, lucru deosebit de important** când ar putea apare tendința de a ne baza, în viitor, mai mult pe tehnologie în procesul educațional.
- *Obținerea unui sistem de cunoştințe* sub forma unui model intern de bază (unor modele interne) la care să se poată racorda uşor cunoştințe de detaliu provenite din exterior eventual prin metode informatice.
- *Obținerea sensurilor cunoaşterii,* a sensului 'fizic', al intuiției lucrurilor şi chiar a unui răspuns creativ în procesul învățării, lucru de care automatele nu sunt capabile, adică a pune umanul în starea lui firească.
- *O deschidere firească spre creativitate şi creație,* spre inovare în vederea rezolvării de probleme care nu sunt structurate după tipul sistemului de cunoştințe existemt în modelul intern disponibil la un moment dat.'

Odată cu apariția e-învățării se deschid perspective noi şi pentru studierea experimentală a procesului de învățare şi confruntarea acestui tip important de proces cognitiv cu teoriile ştiinței cognitive care se vor baza pe progresele pe care le va realiza, ceea ce numim, ştiința integrativă. Considerații privind procesul de învățare şi sisteme de e-educație, inclusiv prin folosirea metodelor inteligenței artificiale sunt prezentate într-un grup de trei lucrări recente ale unor cercetători ştiințifici de la Centrul pentru Cercetări Avansate în Învățarea Automată, Prelucrarea Limbajului Natural şi Modelare Conceptuală şi Institutul de Psihologie 'Mihai Ralea' al Academiei Române [14], [15], [16].

National Research Council de pe lângă Academia Națională de Ştiințe din SUA a prezentat, în februarie 2002, un raport [17] privind cercetarea ştiințifică a educației în care despre studiul stiințific al procesului de învățare se arată:

'Much of the controversy about education research relates to its perceived lack of quality. […] Is scientific education research the same as research in social and behavioral science generally or the same as research in the physical sciences? […] A key finding of this NRC committee is that at a fundamental level, 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.

II.4 Învățământul electronic (e-learning)

E-learning este un vector functional 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. Iar performanțele studenților şi elevilor sunt similare (evaluare pentru anul 2000) cu cele ale învățământului în clase de elevi şi studenți..
- Corporațiile industriale recurg masiv la e-educație, iar această tendință nu mai poate fi ignorată. Unele corporații au lansat e-universități pentru personalul propriu, de. Ex. Dell Computer Corp. şi Sun Microsystems.
- Universitățile au început să introducă nu un e-învățământ complet, ci constituirea treptată a acestuia prin unele e-cursuri. Spre exemplu University of California, Berkeley, în domeniul ştiinței şi tehnologiei informației a început (anii 1999-2000) cu patru e-cursuri: sisteme informatice, telecomunicații digitale, e-comerț, sisteme informaționale geografice.
- O serie de firme şi-au dedicat activitatea sau o parte din activitate producerii unor 'e-learning software packages'. Se constituie un segment al pieții software specializat în e-learning. (Astfel se şi explică prezența firmelor

SOFTWIN şi SIVECO la acest simpozion devenite principalele firme româneşti de software educational. Dar asemenea pachete e-software pentru învățământ sunt de aşteptat şi din partea Programului e-şcoală al Ministeului Educației şi Cercetării care urmăreşte o reformă educațională în România.

- Nu se constată deosebiri între rezultatele învățării on-line şi învățarea într-un campus universitar sau o şcoală. Învățarea electronică cere mai multă disciplină şi maturitate decât învățarea convențională [18].
- Pentru experimente de laborator şi pentru viață socială este nevoie totuşi de perioade de lucru în instituțiile de învățământ.
- Odată cu creşterea utilizării metodelor de e-învățământ, construcția de clădiri pentru învățământ se va diminua. În schimb apar cheltuieli pentru noua infrastructură a e-învățământului.
- Modul asincron de acces la cursuri permite e-educația în orice moment şi în orice loc.
- E-învățământul încurajează studenții să-şi asume o mai mare responsabilitate pentru definirea şi organizarea a ceea ce urmăresc să învețe. Studenții sunt mai bine serviți având un acces asistat electronic on-line la cei mai buni instructori decât un contact față în față cu instructori mediocri [19]. In orice caz, nu se neagă rolul instructorilor.
- Discipline ca filosofia şi istoria presupun discuții, iar discipline tehnice presupun proiecte. În aceste cazuri trebuie încă să se găsească soluții mixte de învătământ clasic şi electronic.
- E-învățământul oferă cele mai bune perspective pentru învățarea în 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 consideraţii şi constatări la începutul anului 2002. Valabilitatea unora dintre ele se va confirma, alte constatări vor fi, poate, infirmate, dar vor apare cu siguranţă multe alte aspecte noi.

II.5 Viaţa intelectuală

În timp sunt prevăzute multe schimbări datorită învăţământului electronic [23]. În primul rând, apariţia unor colegii şi universităţi nelocalizate, extinse uneori la scară globală. Siturile acestora pot fi mari sau mici, structurarea socială având loc sub forma unor comunităţi (villages) având facilităţi comune pentru cercetare, proiecte de grup, dar şi pentru activităţi comunitare culturale, sportive etc. O persoană admisă într-o asemenea universitate îi va rămâne ataşată pentru toată viaţa, deoarece educaţia se va extinde pe întreaga viaţă prin perioade discrete (adică necontinue) şi intensive de învăţare. Viaţa 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 canonical body of basic knowledge that defines this notion' [23].

Agenţii inteligenţi de căutare a informaţiei, bibliotecile electronice, vizualizarea informaţiei, 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 umanităţilor în tehnologie şi invers, vor duce la o reorganizare radicală a disciplinelor intelectuale [23].

II.6 Perspective

Cum vor evolua lucrurile în viitor? Ray Kurzweil [24] face următoarele previziuni privind educaţia pentru anii 2009, 2019 şi 2029:

Pentru anul 2009 [24, p. 191-192]:

> '…most effective learning from computers taking place in the home. […] The profound importance of the computer as a knowledge tool is widely recognized. Computers play a central role in all facets of education, as they do in other spheres of life. The majority of reading is done on displays, although the 'installed base' of paper documents is still formidable. The generation of paper documents is dwindling, however, as the books and other papers of largely twentieh century vintage are being rapidly scanned and stored. Documents circa 2009 rotinely include embedded moving images and sounds. Students of all ages typically have a computer of their own, which a thin tabletlike device weighing under a pound

with a very high resolution display suitably for reading. Students interact with their computers primarily by voice and by pointing with a device that looks like a pencil. Keyboards still exist, but most textual language is created by speaking. […]

Intelligent courseware has emerged as a common means of learning. […] The traditional mode of a human teacher instructing a group of children is still prevalent, but schools are increasingly relying on sofware approaches, leaving human teachers to attend primarily to issues of motivation, psychological well-being, and socialization.'

Pentru anul 2019 [24, p.204):

'Paper books and documents are rarely used or accessed. Most twentieth-century papers of interest have been scanned and are available through wireless network. Most learning is accomplished using intelligent software-based simulated teachers.[…] The teachers are viewed more as mentors and counselors than as sources of learning and knowledge. Students continue to gather together to exchange ideas and to socialize, although even this gathering is often physically and geographically remote. […] Most adult human workers spend the majority of their time acquiring new skills and knowledge.'

Pentru anul 2029 [24, p. 221]:

'Human learning is primarily accomplished using virtual teachers and is enhanced by the widely available neural implants. The implants improve memory and perception, but it is not possible to download knowledge directly. Although enhanced through virtual experiences, intelligent interactive instruction, and neural implants, learning still requires time-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 educațională?

II.7 Încheiere. Propuneri

Roger Bohn defineşte, într-un mod specific pentru societatea cunoaşterii, **învățarea drept evoluția cunoaşterii în timp** [25].

Studiul procesului de învățare scoate în relief importanța ştiinței cognitive şi a învățării ca proces cognitiv fundamen-tal. Această ştiință trebuie nu numai cunoscută, atât cât este ea astăzi, ci mai ales dezvoltată de către psihologi, neu-robiologi, sociologi şi specialişti în inteligența artificială.

Este necesară o direcție de cercetare bine susținută pentru a stimula contribuții româneşti în acest domeniu. Am propus şi propun în continuare ca în cadrul programului INFOSOC (Programul național de cercetare-dezvoltare pentru societatea informațională) să se stimuleze cercetări în domeniul ştiinței 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 inteligenței 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 cunoaş-tem ce posibilități şi ce potențial avem în domeniul utilizării agenților inteligenți şi să existe o coordonare şi autocoor-donare a eforturilor. Utilizarea agenților inteligenți pentru toți vectorii societății cunoaşterii, inclusiv pentru e-învăță-mânt va deveni determinantă pentru calitatea şi eficiența acestei societăți. Recenta propunere pentru transformarea Centrului pentru Cercetări Avansate în Învățarea Automată, Prelucrarea Limbajului Natural şi Modelare Conceptuală al Academiei Române într-un Centru de cercetări pentru Inteligența Artificială şi Societatea Cunoaşterii sprijinită de Directorul general ICI, Doina Banciu şi de Ministrul Comunicațiilor şi Tehnologiei Informației, Dan Nica, ar putea să satisfacă acestor cerințe actuale şi de viitor. Sperăm ca şi Academia Română să sprijine această solicitare pentru a putea fi înaintată Guvernului României spre a fi aprobată.

Tot la Academie, Comitetul Român pentru Istoria şi Filosofia Ştiinței şi Tehnicii va acorda o anumită importanță muzeelor virtuale, nu numai pentru istoria ştiinței şi tehnicii, dar şi pentru cunoaştere şi învățare. Ar trebui realizat un web-site de sinteză a tuturor muzeelor virtuale din lume, inclusiv al web-site-urilor unor mari muzee de mare tradiție şi importanță, cu adresele lor pe Internet. Acest web-site ar trebui să fie cunoscut şi accesibil tuturor în România.

Apreciez în mod deosebit eforturile care se fac pentru informatizarea învățământului românesc de către Guvernul României, Ministerul Educației şi Cercetării, firmele SIVECO şi SOFTWIN, ca şi de toate instituțiile reprezentate la acest simpozion dedicat învățământului electronic.

Doresc să mulțumesc tuturor celor care au prezentat comunicări la acest simpozion şi celor care au participat la organizarea lui.

Referinţe bibliografice

[1] Mihai Drăgănescu, *Cunoaşterea şi societatea cunoaşterii,* comunicare la seziunea de lansare a programului strategic SI-SC, Academia Română, 10 aprilie 2001; 1b. Mihai Drăgănescu, *Societatea informaţională şi a cunoaşterii. Vectorii societăţii cunoaşterii,*studiu, Academia Română, 7 iulie 2001, publicat pe Internet şi în vol. coord. Florin Gh. Filip, *Societatea informaţională-Societatea cunoaşterii*. *Concepte, soluţii şi strategii pentru Români,* Academia Română, 2002, p.43 - 112.

[2] Gabriela S. Sabău, *Societatea cunoaşterii. O perspectivă românească,* Editura economică, Bucureşti, 2001; Ion Gh. Roşca, Viorel Petrescu, Beniamin Cotigaru, Gabriela Sabău, Vasilica Ciucă, Oscar Hoffman, Wilhelm Kappel, *Cercetarea pentru dezvoltarea în reconstrucţia durabilă a economiei din perspectiva societăţii cunoaşterii,* Economistul, 4 februarie 2002, nr.270, p. I-III.

[3] Mihai Drăgănescu, *Societatea cunoaşterii,* Diplomat Club, 2001, Nr. 6, p1-2; Mihai Drăgănescu , *Knowledge management, a functional vector of the knowledge society*, Diplomat Club, Nr. 10-11, 2001, p.4; Mihai Drăgănescu, *Factori noi în viaţa cultural-ştiinţifică-politică globală: terorismul şi antiterorismul,* Diplomat Club, 2002, Nr.1, p.7.

[4] Mihai Drăgănescu, *Tendencies of becoming,* Romanian Review, 1986, Nr. 11, p.55-59.

[5] Mihai Drăgănescu, *Spiritualitate, Informaţie, 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 vol. 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/sflibrary.html, 1988.

[9] Dick Brass, Vicepreşedinte Microsoft pentru dezvoltare tehnologică, *E-books,* în vol. Inside/Out, Microsoft- in our own words, Penguin Books, New York, 2000, p.262-263.

[10] Mihai Drăgănescu, *Microelectronica şi învăţământul în domeniul electronicii (I),* Forum, anul XXX, noiembrie 1988, p. 36-48.

[11] Mihai Drăgănescu, *Ştiinţa cognitivă, ştiinţă structurală sau ştiinţă integrativă?*Comunicare la seziunea ştiinţifică 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, *Achiziţia, gestiunea, partajarea şi prelucrarea cunoştinţelor pe web: elemente esenţiale în societatea cunoaşterii,* în vol. coord. Florin Gh. Filip, *Societatea informaţională-Societatea cunoaşterii*. *Concepte, soluţii şi strategii pentru Români,* Academia Română, 2002, p.195-207.

[15] Cristina V. Niculescu, *Noi tipuri de sisteme educaţionale pentru SI-SC,* în vol. coord. Florin Gh. Filip, *Societatea informaţională-Societatea cunoaşterii*. *Concepte, soluţii şi strategii pentru Români,* Academia Română, 2002, p.209-223.

[16] Gheorghe Iosif, Ana Maria Marhan, Ion Juvină, *Strategii de creştere a utizabilităţii şi de dezvoltare a competenţelor de bază ale populaţiei României pentru utilizarea tehnologiei informaţiei,* în vol. coord. Florin Gh. Filip, *Societatea informaţională-Societatea cunoaşterii*. *Concepte, soluţii şi 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 Representatives*, February 28, 2002.

[18] Robert Ubell, *Engineers turn to e-learning,* IEEE Spectrum, October 2000, p.59-63.

[19] Peter Wiesner, *Distance Education: Rebottling or a New Brew?* Proceedings of the IEEE, July 2000, p.1124- 1130.

[20] Ralph B. Ginsberg, Kenneth R. Foster, *The Wired Classroom,* IEEE Spectrum, August 1998, p.44- 51.

[21] Paul G. Shotsberger, Ron Vetter, *Teaching and Learning in the Wireless 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 year 2049,* Proceedings I.E.E.E., September 1999, p.1685 - 1691.

[24] Raz Kurzweil, *The Age of Spiritual Machines,* Penguin Books, New York, 1999.

[25] Roger E. Bohn, *Measuring and Managing Techological Knowledge,* p.295-314, în vol. Eds. Dale Neef a.o., The Economic Impact of Knowledge, Butterworth-Heinemann, Boston, 1998.

[26] Gh. Tecuci, *Building Intelligent Agents,* Academic Press, San Diego, 1998.

III. CULTURA ŞI SOCIETATEA CUNOAŞTERII

Societatea Cunoaşterii

Am prefigurat că va sosi un moment al societății cunoaşterii (chiar cu această sintagmă, Mihai Drăgănescu, 1976, 1986), dar abia în ultimul deceniu al secolului XX conceptul s-a impus în SUA datorită lucrărilor sociologului Peter Drucker şi ale altora, în ultimii 4-5 ani societatea cunoaşterii devenind recunoscută ca o etapă nouă a erei informației, respectiv a societății informaționale. Academia Română a lansat acest concept în România în anul 2001 ca urmare a poziției şi comunicării Mihai Drăgănescu, *Cunoaşterea şi Societatea cunoaşterii,* la seziunea de lansare a programului SI-SC, Academia Română, 10 aprilie 2001 şi a elaborării studiului Mihai Drăgănescu, *Societatea Informațională şi a Cunoaşterii. Vectorii Societății Cunoaşterii,* Academia Română, Bucureşti, 9 iulie 2001, publicat apoi în vol. coord. Florin Gh. Filip, *Societatea informațională-Societatea cunoaşterii*. *Concepte, soluții şi strategii pentru România,* Academia Română, 2002, p.43 - 112.

Spre deosebire de unele puncte de vedere care privesc numai economicul (economia digitală, piața internet) societatea cunoaşterii **nu este numai economia 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 Kassirer [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 omul nu trăieşte numai într-un univers material, ci mai ales într-unul simbolic [3].

Într-o a doua lucrare [1a], *Cultura şi marile tehnologii* (1996) am urmărit linia clasică de definire a culturii ținând însă seamă de obiectele informaționale noi aduse de societatea informațională. **În teoria clasică cultura este definită ca un fenomen social care cuprinde comportamentul socio-uman cu obiectele materiale şi informaționale integrate acestui comportament.** Obiectele informaționale au fost introduse în această definiție la sfârşitul secolului XX.

Pare a fi posibilă o încadrare a teoriei culturii într-o viziune categorială (termenul este utilizat în raport cu teoria categoriilor şi functorilor din matematică, extinsă recent de la domeniul structural la domeniul structural-fenomeno-logic [4]).

Privind comportamentul socio-uman *cultural* ca o categorie, această categorie este o subcategorie majoră a categoriei comportamentului socio-uman *general*. Ultima mai cuprinde şi o subcategorie a comportamentului determinat strict biologic, atât la nivel individual, cât şi social. Într-adevăr, pe lângă comportamente individuale strict biologice există şi comportamente sociale determinate biologic, puse în evidență, în cazul omului, de Gr.T. Popa [5]. Acesta demonstrează cum creierul vechi (primitiv, reptilian, thalamus-hipothalamus) determină comportamente necontrolate cultural care duc mase de oameni la comportamente sălbatice, iar în cazul societăților mai avansate duc la manifestări de semicivilizație, în care impulsivitatea biologică devine colectivă şi sălbatecă.

Cultura

O subcategorie a unei categorii este o categorie. **Categoria cultură** este subcategorie a comportamentului socio-uman general, dar este aceea care deosebeşte specia umană de toate celelalte specii animale, chiar dacă unele dintre acestea pot avea şi rudimente de cultură. Categoria cultură reprezintă *comportamentul socio-uman cultural*, spre deosebire de cel biologic, cu tot ceea construieşte, dar nu se dezvoltă decât datorită, totuşi, anumitor proprietăți biologice remarcabile ale omului, în special ale creierului său

care are o mare disponibilitate informa-țională. De aceea, dacă originea biologică a comportamentului cultural nu poate fi pusă la îndoială, cultura este o construcție care se ridică mult deasupra biologicului, atât cât va putea față de limitele biologice ale omului la un moment dat în istorie.

Poate că alături de cele două subcategorii menționate mai înainte ar trebui să mai adăugăm comportamentului socio-uman încă una, aceea a spiritualității (comportamentul spiritual), pe care nu o tratăm în această lucrare. A privi spiritualitatea ca o a treia subcategorie a comportamentului socio-uman general este o chestiune care trebuie aprofundată, având în vedere că mulți oameni de cultură consideră spiritualitatea a fi un comportament numai cultural. Ținând seama de cercetările şi studiile de filozofie a ştiinței din ultimii 15 ani privind mintea 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 COMPORTAMENTULUI SOCIO-UMAN	**Subcategoria comportamentului strict biologic**	**Notă: există ºi 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 farmacopeea, 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 introdusa sub influența noțiunii de valoare intangibilă din economie care a căpătat o mare importanță pentru societatea cunoaşterii (economia bazata pe cunoaştere). Se mai adaugă aici valori, credințe, cunoaştere tacită.

II. **Cultura umanistă.** Am preluat în acest studiu denumirea tradiţională. Cultura umanistă cuprinde limbajele naturale, literatura, arta, istoria, filosofia, sportul. Cultura umanistă este o cultură tangibilă, ca şi ştiinţa şi tehnologia.

III. **Cultura ştiinţifică: Ştiinţa, tehnologia şi cunoaşterea.** Această categorie a culturii conţine două subcategorii:

III.a **Ştiinţa, cunoaşterea ştiinţifică şi tehnologică,** cunoaşterea tehnologică pentru fabricaţia de produse, dar şi pentru utilizarea acestora, precum şi cunoaşterea organizaţională şi economică, chiar dacă unele obiecte ale cunoaşterii sunt tacite sau fac parte şi din cultura intangibilă. În categoria mare a culturii, anumite obiecte pot aparţine la două sau mai multe subcategorii, acestea nu sunt neapărat disjuncte.

III.b **Uneltele fizice şi informaţionale, obiectele fizice şi informaţionale produse sau fabricate, utilizarea lor, instituţiile şi organizaţiile,** care sunt consecinţe, în cea mai mare măsură, a cunoaşterii ştiinţifice, tehnologice, economice şi organizaţionale, poate chiar şi a culturii intangibile.

Nu numai că unele obiecte culturale pot face parte din mai multe subcategorii ale culturii, dar vor exista şi zone de interferenţă între obiecte ale acestor subcategorii. De exemplu, filosofia ştiinţei, care este un obiect al filosofiei, nu se poarte dezvolta decât în strânsă legătură cu ştiinţa. În teoria categoriilor asemenea legături se numesc morfisme (morphisms sau maps, în limba engleză). Mai mult, pe lângă legăturile dintre obiectele subcategoriilor culturii, din orice sferă a culturii ar proveni, există relaţii între aceste sfere în totalitatea lor. Acestea se numesc functori. Cei mai importanţi functori sunt aceia dintre categoria II şi categoria III de mai sus. Aceşti functori,

F1 : Categoria III (Cultura ştiinţifică) ➔ Categoria II (Cultura umanistă)

F2 : Categoria II (Cultura umanistă) ➔ Categoria III (Cultura ştiinţifică)

reprezintă relaţia şi influenţa reciprocă dintre, în esenţă, cultura umanistă şi ştiinţă (cultura ştiinţifică). Importanţa lor pentru societate şi om nu poate fi subestimată.

Care este mai importantă dintre cele două categorii? Ambele sunt importante, dar motorul dezvoltării provine din sânul categoriei III. Acest lucru a devenit tot mai evident odată cu formularea conceptelor societăţii cunoaşterii [7].

Este adevărat că o altă resursă importantă este viaţa spirituală, ea având şi componenta de creaţie implicând puternic atât cultura umanistă, cât şi cultura ştiinţifică.

Odata cu era informaţiei vor apare desigur multe elemente noi ale culturii datorită tehnologiei informaţiei, cărţii şi documentelor electronice, internetului, tehnologiilor vorbirii, tehnologiilor bioelectronice şi bioinformatice, inteligenţei artificiale şi agenţilor inteligenţi informatici, mediului ambiant inteligent, apariţiei conştiintei artificiale. Vor apare schimbari în viaţa intelectuală, socială şi politică.

Ce se va mai petrece în cultură ?

În secolul XXI sunt posibile câteva evenimente majore care vor schimba viața omenirii:

- Prăbuşirea ecologică a societății şi a speciei umane, datorită deteriorării grave a mediului înconjurător, ceea ce s-ar putea întâmpla la mijlocul sec. XXI (să spunem, anul 2050) dacă nu se trece din timp, 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 va 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 societăţi a conştiinţei 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 fir rezolvată şi sustenabilitatea unei societăţi a conştiinţei.

Este interesant de reluat aici câteva previziuni ale lui Kurzweil [10] privind starea societăţii în anii 2019 şi 2029.

Pentru anul 2019, în domeniul afacerilor şi al economiei, prevede tranzacţii care în majoritate vor folosi persoane simulate, oamenii de afaceri vor avea asistenţi software care vor conduce tranzacţiile în numele lor. Locuinţele vor dispune de roboţoi de întreţinere. Cu aceste artifacte comunicarea se va face prin voce, deoarece vor dispune de o tehnologie a limbajului natural şi a vocii de foarte înaltă calitate. Oamenii vor avea relaţii cu persoane automate inteligente în calitatea acestora de profesori, îngrijitori medicali, persoane de companie etc. Aceste persoane automate au şi calităţi superioare omului în privinţa memoriei, dar, afirmă Kurzweil, 'ele nu sunt încă privite ca fiind egale cu oamenii în toată subtilitatea personalităţii acestora'. Inteligenţa artificială este însă prezentă şi împletită cu toate aspectele societăţii. Responsabilitatea omului va rămâne totuşi pe primul plan şi nu a persoanelor (agenţilor) care îl ajută. Operele de artă se vor realiza prin colaborarea dintre artişti umani şi inteligenţe artificiale. Principalul pericol în societate îl vor constitui micile grupuri de oameni şi inteligenţe artificiale folosind comunicaţii criptate care nu pot fi descifrate. Acestea vor folosi virusuri informatice şi agenţi de îmbolnăvire obţinuţi prin bioinginerie. Pe de altă parte descifrarea relaţiilor dintre genele genomului uman va permite o medicină utilizând inteligenţa artifcială pentru tratamentul şi chiar eradicarea multor boli, inclusiv pentru prelungirea considerabilă a vieţii omului natural.

Pentru anul 2029, Kurzweil prognozează: în domeniul comunicaţiilor va predomina, ca volum, comunicaţia dintre oameni şi maşini. Populaţia umană se va plafona la 12 miliarde de persoane reale, cărora li se asigură toate condiţiile normale de viaţă. Populaţia umană şi a inteligenţelor artificaiale va fi preocupată, în primul rînd, pentru crearea de cunoaştere, într-o puzderie de forme. Va fi greu de să fie menţionate capacităţi ale omului care să nu fie preluate de maşini, de fapt o deosebire netă nu mai există între lumea oamenilor şi lumea maşinilor. Cogniţia umană a fost transferată maşinilor şi multe maşini au personalitate, îndemânări şi baze de cunoaştere preluând şi cunoaş-terea umană. Implanturile neurale cognitive bazate pe inteligenţă artificială vor amplifica funcţiile cognitive ale omului. Kurzweil afirmă: 'A defini ceea ce înseamnă o fiinţă umană devine o chestiune semnificativă politică şi de legislaţie. Creşterea rapidă a posibilităţilor maşinilor este controversată, dar nu există nici o rezistenţă faţă de ea. Deoarece la început maşinile au fost proiectate pentru a fi supuse controlului uman, ele nu au prezentat o faţă ameninţătoare faţă populaţia umană. Oamenii realizează că nu mai este posibilă dezangajarea civilizaţiei devenită om-maşină de dependenţa de inteligenţa maşinilor. Creşte discuţia despre drepturile legale ale maşinilor, în special ale acelor maşini care sunt independente de oameni (care nu sunt introduse într-un creier uman). Cu toate că nu se recunoaşte deplin,

prin lege, influența evidentă a mașinilor la toate nivelele de decizie asigură o protecție importantă a mașinilor'.

Kurzweil consideră calități ale mașinilor inteligente, care încă din anul 2029 pot fi persoane de artă în toate dome-niile artei ('Mulți dintre artiștii de frunte sunt mașini'). Observăm însă că acest lucru ar presupune o stare de conștiință similară omului și prin manifestarea fenomenelor de qualia. Implicit, Kurzweil consideră că mașini inteligente complexe structurale pot avea asemenea stări și pot chiar participa la discuții filosofice pe baza experienței proprii. Vorbind de experiența subiectivă a unor astfel de mașini, aceasta ar însemna că asemenea mașini să fi trecut pragul de la inteligență la conștiință *numai pe baze structurale* încă din anul 2029. Ceea ce nu credem, în principiu, a fi posibil.

Într-adevăr, previziunile pe care oamenii de știință le fac privind dezvoltarea inteligenței artificiale spre conștiință artificială se bazează pe extrapolări ale științei structurale (complexitatea structurală de la un anumit grad în sus generează conștiință, acest lucru fiind considerat valabil începând cu creierul animalelor). Odată cu creșterea complexității artefactelor creiere electronice sau creierelor software se consideră că atunci când acestea ating complexitatea creierului uman se va produce de la sine conștiința 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 științei [15],[16], conștiința nu se poate realiza numai din elemente structurale, fiind nevoie și de elemente fenomenologice [17]. Conștiințele artificiale vor pune probleme foarte mari speciei umane care cred că ar putea fi rezolvate în cadrul unei viitoare societăți a conștiinței. Aceasta va urma atunci societății cunoașterii în cadrul erei informației [18], [19].

Ce va fi cultura în societatea conștiinței, la care vor participa, dacă nu chiar vor predomina conștiințele artificiale? Dacă lucrul cel mai important, în cele din urmă, este continuitatea conștiinței create de omenire, atunci și culturii create de ea trebuie să i se asigure o continuitate.

Aceste considerații arată, dacă mai era nevoie de subliniat, cât de importantă vor fi în sec. XXI, cultura științifică și cultura umanistă, ambele având nevoie de o cultură filosofică adecvată.

Culturi, cultură pozitivă și cultură negativă. Polarizarea culturii în jurul cunoașterii

O cultură poate fi apreciată *pozitiv sau negativ*, în raport cu anumite criterii. Se pierde prea mult din vedere acest lucru. Există astăzi și o cultură a teroriștilor (chiar și o știință a terorismului) o cultură a corupției care ne pune nouă românilor atâtea probleme, o cultură 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 civilizației socio-umane.

În ultimii 12 ani, în societatea românească, pe lânga multe lucruri pozitive, s-au accentuat, din nefericire, şi fenomene negative îngrijorătoare: corupţie, imoralitate, injustiţie. Creşterea imoralităţii şi a injustiţiei, a influenţat până şi viaţa academică din ţara noastră. Avem nevoie şi de un efort cultural pentru a reduce aceste flageluri din societatea noastră, pe lângă efortul dezvoltării economice.

Un exemplu de cultură pozitivă este arta. A cunoaşte arta înseamnă cunoaştere,dar a simţi arta, a trăi arta, a avea nevoie de ea, a fi o bucurie interioară, acestea înseamnă cultură umanistă adevărată.

Dar dacă cele de mai sus nu sunt însoţite de comportament civilizat, de civilizaţie socio-umană, cultura poate fi denaturată (rapturile de opere de artă în scopuri personale sau statale). Natura firească a culturii pozitive este aceea de a susţine civilizaţia socio-umană, spiritualitatea, cunoaşterea şi conştiinţa, în cele din urmă societatea cunoaşterii şi societatea conştiinţei.

În privinţa relaţiei dintre cultura umanistă şi cultura ştiinţifică, astăzi nu se mai poate vorbi de cultură, cu înţelesul de cultură - în general, dar de fapt cu gândul la cultura umanistă.

Cultura - în general, are o mult prea puternică componentă ştiinţifică (inclusiv tehnologică, economică, organizaţională, politică) pentru a mai accepta o asemenea simplificare, este adevărat, continuatoarea unei tradiţii care astăzi este complet depăşită. Cultura, respectiv cultura - în general, este cultura umanistă ***şi*** *cultura ştiinţifică, împreună, ultima având, ca şi prima, un conţinut extrem de bogat.*

În spatele confuziei care se menţine astăzi atunci când vorbim de cultură se întreţine schisma dintre cele două culturi, datorită unor interese de grup. În etapa actuală a societăţii, cultura umanistă nu-şi mai poate erija numele general de cultură, de fapt nu ea, ci slujitorii ei care nu s-au adaptat la vremurile cunoaşterii. În societatea cunoaşterii, înainte de trecerea la societatea conştiinţei, cultura se va concentra în jurul cunoaşterii. Iar tehnologia va fi un factor cultural atât de covârşitor încât va reveni poate la poziţiile ei mitologice din antichitate. În [1a] remarcam:

În antichitate, la egipteni, zeul Ptah era privit ca patronul lucrătorilor de metal (metalurgişti şi fierari) şi al artizanilor. Ptah era însă unul dintre cei mai mari zei, creatorul pământului, părintele zeilor şi al începuturilor. Interesant acest "al începuturilor".

La grecii antici, echivalentul lui Ptah era Hefaistos, zeul focului şi meştesugurilor, protectorul artizanilor. El nu mai era o divinitate primordială, dar divinitate, fiul lui Zeus şi al Herei, fiind căsătorit cu Afrodita. Se pare că de la egipteni la greci, tehnologia nu mai păstra poziţia începuturilor, dar avea totuşi un reprezentant divin. La romani, echivalentul lui Hefaistos era Vulcan, considerat zeul focului.

Decăderea poziţiei 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 si chiar despre o filosofie a tehnologiei, de care o serie de gânditori şi filosofi au scris lucrări deosebit de interesante: Ernst Kapp, Friedrich Desauer, José 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 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:*

I. Mihai Drăgănescu, ***Cultura şi marile tehnologii,*** conferință, Universitatea Populară de Vară "Nicolae Iorga", Vălenii de Munte -30 august 1996.

II. Mihai Drăgănescu, ***Perspectiva informațională a culturii,*** Contemporanul, 27 mai 1983.

III. Mihai Drăgănescu, ***Dimensiunile europene ale culturii române,*** expunere, Vălenii de Munte, 1992, publicată în Academica, 1992.

IV. Mihai Drăgănescu, ***Arta şi societatea,*** cuvânt, Ploieşti, 4 noiembrie 1991, publicat în Academica, 1991.

V. Mihai Drăgănescu, ***Criterii transpolitice si transmafiote în cultură***, 18 mai 1997, Caiete Critice, 1997, nr.3-4, p.145-147.

VI. Mihai Drăgănescu, ***Spirit enciclopedic şi enciclopedism,*** conferință, Vălenii de Munte, 22 august 1993 (publicată în Academica şi în volumul autorului, Cariatidele gândului, Ed. Academiei Române, 1996, p. 163-168).

[2] Ernst Cassirer, ***Substanzbegriff und Funktionbegriff***, 1910; ***Die Philosophie der Symbolischen Formen***, 1923-1929 (3 vol).

[3] Oltea Miºcol, Elena Gheorghe, ***Repere istorice în filosofia culturii***, Revista de 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 spiritului,*** volum în editare, conþinând lucrări ale acestui autor prezentate ºi publicate la Academia Română în anii 1940 (vezi ºi prefaþa: Mihai Drăgănescu, *O gândire asupra conºtiinþei, moralităþii ºi societăþii*).
[6] UNESCO, definiþia culturii intangibile, web-site UNESCO.
[7] Mihai Drăgănescu, ***Societatea Informaþionalã ºi a Cunoaºterii. Vectorii Societăþii Cunoaºterii,*** Academia Română, Bucureºti, 9 iulie 2001, publicat în vol. coord. Florin Gh. Filip, *Societatea informaþionalã-Societatea cunoaºterii. Concepte, soluþii ºi strategii pentru România,* Academia Română, 2002, p.43 - 112).
[8] Moravec H., ***Rise of the Robots,*** Scientific American, December 1999, p. 86-93.
[9] Moravec H., ***Robot. Mere Machines to Transcendent Mind,*** Oxford Universitz Press, Oxford, 1999.
[10] Kurzweil R., ***The Age of Spiritual Machines***, Penguin Books, 2000.
[11] Broderick D.,***The 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 new species,*** MIT Press, Cambridge, Massachusetts, 2002.
[14] Drăgănescu M., Din l*ucrările despre minte ºi conºtiinþă*:
A. Mihai Drăgănescu, ***The Interdisciplinary Science of Consciousness***, Noetic Journal, Vol.3, No.1, Jan.2000, p. 37-46; republicat in eds. Richard L. Amoroso et al, *Science and the Primacy of Consciousness, Intimation of a 21st Century Revolution*, Chapter 5, pp. 46-59, Orinda: The Noetic Press, 2000.
B. Mihai Drăgănescu, ***Theories of Brain, Mind and Consciousness: Still Great Divergences,*** Noetic Journal, vol.3, No. 2, Apr. 2000, 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., Draganescu M., ***Preliminaries to the Philosophy of Integrative Science,*** E-book (Microsoft Reader), ISBN 973-10-02510-X, Editura ICI, Bucharest, 2001.
[16] Draganescu 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 vol. *Science and the Primacy of Consciousness, Intimation of a 21st 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, ***Conştiința, frontieră a ştiinței, frontieră a omenirii,*** Revista de Filosofie, XLVII, nr. 1-2, 2000, p.15-22.

[19] Mihai Drăgănescu, ***Societatea conştiinței, o viitoare etapă a erei informației. Vectorii societății conştiinței,*** studiu, Academia Română, în pregătire.

[20] După Alain Fienckielkrant, apud [3], p.458-459.

Între lingvistica matematică şi cea computaţională

Acad. Solomon MARCUS
Secţia de Ştiinţe Matematice a Academiei Române
solomon.marcus@imar.ro

Mă simt obligat să reacţionez la un anumit mod de prezentare a evoluţiei 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 Române. Precizez de la început că nu contest interesul şi utilitatea direcţiei de preocupări prezentate în [1]; am în vedere numai modul în care această direcţie este pusă în relaţie cu alte cercetări dedicate limbajului.

Citǎm din [1: 133]:

"*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 soluţii abstracte de modelare generativă de tip universal a ceea ce se presupunea (la nivelul cunoaşterii ştiinţifice a anilor 1960) a fi facultatea limbajului*".

Nu ştiu ce înţelege DT prin "lingvistica formală", o sintagmă nu prea folosită în perioada de emergenţă 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.), 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 în engleză (de exemplu, "machine translation"), franceză, rusă, germană, italiană etc. Din aceste preocupări s–au inspirat primele modele care au constituit noua disciplină a LM.

Vorbesc despre lucruri trăite. Punctul meu de plecare s–a aflat în lucrările unor Kulagina şi Melciuk, puternic implicaţi în studiile de traducere automată rusă–franceză, Yves Lecerf, implicat în problemele de documentare automată, D. G. Hays, implicat în traducerea automată din rusă în engleză şi reciproc, B. Vauqois, cu preocupări de informatică lingvistică la Grenoble. De la ei, ca şi de la alţi 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 mulţi cercetători din domeniul LM din anii 1950 şi 1960, cum ar fi

Maurice Gross, Masami Ito, A. Trybulec şi mulţi alţii. Dubioasă mi se pare sintagma "soluţii 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 echivalenţă, 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 interacţiunea cu calculatoarele electronice, este individualizat sub numele de lingvistica computaţională*".

Chestiunea cu atingerea limitelor ţine 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 ca între LM şi LC ar putea avea loc o competiţie, prima definindu–se prin metoda (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 definiţie 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 relevanţă, este însă simptomatic pentru viziunea sa limitativă în problema în discuţie.

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 sau 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 atenţiei în 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 ţinta programării în limbaj natural se află în actualitate, nu este şi el semnificativ ? LC are mai multe părţi, mai multe orientări, mai multe niveluri de abstracţie, care comportă criterii diferite de evaluare. DT îl asociază pe D. Hays la ideea sa privind falimentul LM şi lansarea, drept consecinţă, a LC. Ca unul care a cunoscut bine cercetările lui Hays (a se vedea frecvenţa citărilor numelui său în lucrările subsemnatului) şi l–a cunoscut şi personal foarte bine, fiind invitatul său ca "plenary speaker" la Institutul de lingvistică al Americii (SUNY, Buffalo, 1971), pot depune mărturie că acest autor vedea în LM şi LC două domenii solidare, două feţe ale aceleiaşi medalii, aşa cum se va vedea din citatul pe care–l vom da mai jos. Desigur, Hays a avut un rol important în anii de pionierat ai LM şi LC, dar ideea unei competiţii între ele i–a fost străină. Voi evoca aici intervenţia sa la cea de a treia Conferinţă Internaţională de LC

(COLING, September 1971): "The field and scope of Computational Linguistics" [2]. Citǎm ([2]:p.23):

"*Solomon Marcus says that formal linguistics is a pilot science, emphasizing at the same time that the ordinary field of linguistics is not. But that is to say that linguistics as a branch of mathematics will supply methods to many fields of science, whereas linguistics as a descriptive field, a branch of natural history or natural science, does not. [...] A four–way scheme can be arranged, with psychology, computation, formal linguistics, and descriptive linguistics at the poles. Psychology and computation are about performance, formal and descriptive linguistics are about competence, computation and formal linguistics are abstract, and psychology and descriptive linguistics are sciences. But two other fields have to find places in this scheme: psycholinguistics joins psychology with linguistics and seems at this time a most fruitful field, one in which great progress can be made with benefit to both parent fields. Correspondingly, on the abstract side, COMPUTATIONAL LINGUISTICS JOINS COMPUTATION WITH FORMAL LINGUISTICS (subl. mea, S. M.) and also seems a fruitful area, one in which RAPID PROGRESS CAN BE EXPECTED WITH BENEFIT TO BOTH PARENT FIELDS (subl. mea, S. M.) and with beneficial application to psycholinguistics*".

Referirea pe care o face Hays la subsemnatul are în vedere sloganul, pe care l–am folosit de mai multe ori, "formal linguistics as a pilot science", unde sintagma "formal linguistics" era folosită ca un echivalent al LM. Iată deci că Hays vedea în LC o alianţă a LM cu computaţionalul, alianţă de natura să imprime un progres rapid atât în LM cât şi în domeniul computaţional. Cei 30 de ani scurşi de atunci au confirmat–o pe deplin. Denumirile folosite pentru preocupările la interferenţa limbajelor, informaticii şi matematicii au variat tot timpul şi nu cred că acest aspect merită prea multă atenţie. Lingvistica matematică? computaţională? inginerească? algebrică? cognitivă? aplicată? cantitativă? teoretică? statistică? probleme matematice ale semioticii? tehnologia limbajului? limbajul în inteligenţa artificială? lingvistica inginerească? procesarea limbajului natural? "information storage retrieval"? lingvistica cibernetică? pe fiecare dintre acestea am întâlnit–o şi propriile mele articole au fost publicate aproape sub fiecare dintre etichetele de mai sus. Iată şi câteva detalii semnificative ale istoriei.

În 1962 s–a înfiinţat în USA "Association of Computational Linguistics".

În 1963 Ferenc Kiefer a demarat la Budapesta revista "Computational Linguistics", care a trăit peste zece ani. Conferinţa 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 conferinţe au avut loc pe tema "avtomaticeskaja obrabotka tekstov" iar "Sprachkunde und Informationsver- arbeitung" 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. Emergenţa LC s–a produs încă din anii '50, sintagma LC a devenit curenta încă de la începutul anilor '60. Şirul de conferinţe COLING nu a făcut decât să continue această tradiţie. Alţii 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 diferenţă faţă de cele de CL. Desigur, între timp au început să apară şi unele publicaţii mai specializate, cu referire la părţi determinate ale CL (cum ar fi cea relativă la corpusul lingvistic). Etichetele nu au avut importanţă şi nu ştiu să se fi desfăşurat vreo competiţie 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 discordanţă 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 aceasta afirmaţie (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? Indicaţii bibliografice asupra lor sunt date parţial 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 instanţierii unei descrieri lingvistice cât mai complete, mentenabilitatea acestei instanţieri şi, desigur, conformantă cu realitatea uzului limbii". ([1]: 133). Cu un mic efort înţelegem 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 perfecţionate. Dar acest fapt nu ţine, cum crede DT, de alegerea între LM şi LC, ci de progresul general realizat în ştiinţă. Pentru a mă referi la propria noastră experienţă, atunci când, în 1969, prezentăm la COLING–ul din Suedia gramaticile contextuale nu aveam cum să mă ocup de aspectul complexităţii acestor gramatici în maniera în care s–a 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 computaţional. Nu m–am mirat atunci când "Encyclopedia of Microcomputers" şi "Encyclopedia of Computer Science and Technology" mi–au solicitat o contribuţie 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 "Computational Linguistics in the Netherlands 2000"[13]. Nu m–am mirat nici când am văzut că o revista cu titlul "Linguistics and Philosophy" publica articole excelente de LC. Interferenţele merg în toate direcţiile 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 "computational". O revistă ca "Theoretical Linguistics" (1970 2000), publicată de Walter de Gruyter (Berlin–New York) a inclus multe articole vizând aspecte matematice şi/sau computaţionale, deşi numele revistei nu indică acest lucru. Chiar o revistă mai tradiţională, 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 computaţionale. Era internetului impune desigur o problematică nouă, faţă de care abordările anterioare se pot dovedi insuficiente. Salutăm iniţiativa noii generaţii 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 ţine seama în mod critic de experienţa acumulată. Din tot ceea ce am prezentat mai sus rezultă clar 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 and Computational 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: "Computational 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 satisfacţie să trăiesc momentul în care această legătură a fost acceptată de ambii parteneri şi că de multe ori nici nu mai e nevoie de accentul retoric al epitetului "matematica"; LM este acceptată pur şi simplu ca lingvistică. Suntem convinşi că o traiectorie similară o urmează şi LC iar unele semne în această privinţă există de pe acum, aşa cum am arătat mai sus.

LC este de mai mulţi ani o secţiune la congresele internaţionale de lingvistică iar LM şi LC au secţiunea lor în reviste internaţionale de referate ca "Language and Language Behavior Abstracts". În România, minţi luminate ale anilor '60, ca profesorii Al. Rosetti, Grigore Moisil şi Tudor Vianu, au înţeles schimbările care se profilau şi au sprijinit proiectul înfiinţării unei secţiuni de "lingvistica aplicată" la Facultatea de Limbă şi Literatură Română a Universităţii din Bucureşti, dar s–au găsit alţii care să–i torpileze.

La Academia Română a funcţionat mulţi ani "Comisia de Lingvistică Matematică" iar revista "Cahiers de Linguistique Theorique et Appliquee", înfiinţată în 1962, a fost multă vreme expresia colaborării lingvisticii cu matematica şi cu informatica. În ciuda forţelor adverse, s–a reuşit în acei ani atragerea unor studenţi străluciţi ai unor facultăţi umaniste la cercetarea limbii şi literaturii cu mijloacele LM şi LC. Pentru a da numai două exemple de actuali profesori universitari care au susţinut teze de doctorat de acest tip, voi menţiona pe Pia Brinzeu, de la Catedra de Engleză a Universităţii din Timişoara şi pe Mihai Dinu, de la Facultatea de Litere a Universităţii din Bucureşti. Tot în acea perioadă şi-a susţinut teza de doctorat Sorin Cristian Niţă, 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 Computational 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. 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 specificul volumului în care apare articolul [1], DT face numeroase referiri la acte şi documente ale unor organisme europene şi internaționale, cum este şi firesc, pentru a nu mai vorbi de aspectul financiar al colaborării cu organismele respective. Această situație a existat de la începutul LM şi LC (chiar dacă nu a avut amploarea de azi), datorită faptului că LM şi LC au apărut şi ca urmare a unor comandamente sociale, privind precaritatea mijloacelor de prelucrare a informației. Îmi amintesc de faimoasele Rapoarte CETIS care veneau de la EURATOM, Bruxelles, pe teme legate de analiză şi prelucrarea automată a limbajului, traducere automată şi documentare automată. În USA, diferite corporații (cum ar fi RAND Corporation, Santa Monica, Calif.) finanțau cercetări similare. O întâlnire semnificativă a fost aceea din 1962, organizată 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 şi 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 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:

[1] D. Tufiº. *Promovarea limbii române în SI–SC*. În *Societatea Informaþionalã – Societatea cunoaºterii* (coord. F. Gh. Filip). Ed. Expert, Bucureºti, 2001, 131–142.

[2] D. G. Hays. *The field and scope of computational linguistics*. Papers in Computational Linguistics (eds. F. Papp, G. Szepe). Proceedings of the Third International Meeting of Computational 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.* In *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.* In *Current Issues in Mathematical Linguistics* (C. Martin–Vide, ed.). Elsevier – North Holland, Amsterdam, 1994, 3–21.
[7] S. Marcus. *Lingvistica matematica, azi.* In *Matematica in lumea de azi si de maine* (C. Iacob, coord.), Editura Academiei, Bucuresti, 1985, 182–186.
[8] S. Marcus. *Recent Romanian investigations in the field of mathematical and computational linguistics.* Avtomaticeskaja Obrabotka Tekstov, Matem. Fyz. Fakulta, KL Praha, 1973, 15–42.
[9] S. Marcus. *Mathematical and computational linguistics.* In *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*. Computational Linguistics 24, 1998, 2, 245–274.
[11] S. Marcus. *Semiotics and formal artificial languages*. In *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 Formal Languages* (G. Rozenberg, A. Salomaa, eds.), 2, Springer, Berlin, New York, 1997, 215–235.
[13] S. Marcus, C. Martin–Vide, G. Paun. *A new–old class of linguistically motivated regulated grammars. Computational Linguistics in the Netherlands 2000* (W. Daelemans et al., eds.), Selected Papers from the Eleventh CLIN Meeting, Ed. Rodopi, Amsterdam, New York, 2001, 111–125.
[14] B. H. Partee, A. Ter Meulen, R. Wall. *Mathematical Methods in Linguistics*. Kluwer, Dordrecht et al, 1990.
[15] E. F. Beckenbach, Ch. B. Tompkins (eds.). *Concepts of Communication: Interpersonal,* Intrapersonal and Mathematical. John Wiley and Sons, New York, 1976.
[16] D. G. Hays. *Introduction to Computational Linguistics*. American Elsevier, New York, 1967.
[17] R. Thom. *Stabilite Structurelle et Morphogenese*. John Benjamins, New York, 1970.
[18] Y. Bar–Hillel. *Four Lectures on Algebraic Linguistics and Machine Translation* revised version of a series of lectures given in July 1962, before a NATO Advanced Summer Institute, Venezia, Italy.

Între lingvistica matematică şi cea computaţională: o altă perspectivă

Dan TUFIŞ

1. În loc de introducere

Dat fiind că acest articol este un comentariu asupra filipicei de neînţeles ”Între lingvistica matematică şi cea computaţională” 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 afirmaţiilor şi implicaţiilor 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 menţionat îmi oferă posibilitatea de a aduce în discuţie elemente de istorie a domeniului ce 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 secţiunea domniei sale de referinţe bibliografice. Lucrările pe care le-am citat eu sunt documentate în cuprinsul textului, prin includerea referinţei complete între paranteze rotunde. Singura excepţie este lucrarea mea, sursa nemulţumirii domnului Marcus, care este referită de amândoi ca [1]. Cititorul va putea face astfel mai uşor distincţia între cele două categorii de referinţe. Înainte de a proceda la analiza afirmaţiilor domnului Academician Marcus, aş dori să fac unele precizări:

- contextul discuţiei 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 mulţi 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

ştiinţifică şi se referă ironic (şi după cum se va vedea în continuare, în mod nejustificat) la alte două, făcând abstracţie de restul prezentării care nu are nici o contingenţă cu domnul Marcus. Domnul Academician are merite pe care nu i le poate lua nimeni, are contribuţii 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 discuţie pe care am avut-o în anul 1991 la câtva timp după ce mă întorsesem de la Conferinţa Europeană de Lingvistică Computaţională organizată la Berlin de profesorul Jurgen Künze. Cu acea ocazie, domnul Academician Marcus mi-a mărturisit că îl cunoaşte de multă vreme pe profesorul Künze 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 despărţit, profesorul Künze optând pentru noua paradigmă a lingvisticii computaţionale.

– 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 ştiinţifică, 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 discuţie exact acest cadru al investigaţiei tehnologice şi a măsurilor ştiinţifice, tehnice, organizatorice şi chiar legislative pentru a crea o bază perenă a cercetării şi dezvoltării tehnologice privind prelucrarea automată a limbii noastre: resursele computaţionale fundamentale ale limbii române. Societatea Informaţională-Societatea Cunoşterii este caracterizată de vectori tehnologici şi funcţionali [M. Drăgănescu: „*Societatea informaţională-societatea cunoaşterii. Vectorii societăţii cunoaşterii*”, In *Societatea Informaţională – Societatea cunoaşterii* (coord. F. Gh. Filip). Ed. Expert, Bucuresti, 2001, 43–112.] a căror ignorare este nu numai neproductivă dar şi periculoasă. „În era electronică, ***este esenţial pentru supravieţuirea 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 informaţional al societăţii cunoşterii este un obiectiv actual şi de viitor şi nu poate fi subiect de dispută în viaţa ştiinţifică românească;

– deşi este un truism, cred că pentru evitarea unor interpretări greşite este necesar să subliniez faptul că în dezvoltarea programelor de inteligenţă artificială, de prelucrare a limbajului natural sau în general în ingineria software, o mulţime de discipline matematice (teoria algoritmilor, teoria complexităţii, teoria limbajelor formale, teoria categoriilor, statistica

matematică şi multe, multe altele) sunt fundamente indispensabile în avansul ştiinţific şi tehnologic al acestor discipline (şi desigur nu numai al lor). Programarea (ca şi matematica elementară) sau utilizarea de produse informatice sunt activităţi la îndemâna tuturor (de altfel reflectate şi în programele şcolare de învăţământ), dar proiectarea şi realizarea de programe software inteligente necesită o pregătire teoretică solidă, talent şi multă muncă. Diferenţa între două programe care realizează aceleaşi prelucrări dar unul în câteva secunde şi altul în câteva ore, apare tocmai din diferenţa de pregătire teoretică şi talent a autorilor lor.

– domeniul ştiinţei şi tehnologiei informaţiei este poate cel mai dinamic sector al activităţii creative: Bill 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 ştiinţifice obţinute în ultimii 50 de ani) nu a fost nici pe departe egalat de ritmul dezvoltării în domeniul software. În ciuda acestui decalaj, ştiinţa ingineriei software si-a reînnoit instrumentarul teoretic (modele şi/sau formalisme) cu o viteză neîntâlnită în alte domenii ştiinţifice. Dinamica fără precedent a cunoaşterii în ştiinţa şi tehnologia informaţiei obligă omul de ştiinţă 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 informaţia mai veche de 15-20 ani este foarte probabil să fie perimată (desigur cu excepţiile ce întotdeauna confirmă regula). Evoluţia 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 susţinute 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 computaţională" 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 competiţie; pur şi simplu ele sunt subsecvente din punctul de vedere al relevanţei faţă de problemele pe care le discutăm aici. Există fără îndoială o filiaţie între ele, în sensul că lingvistica computaţională a preluat o mare parte din instrumentarul lingvisticii matematice (nici nu se putea altfel) dar ce a adus nou lingvistica computaţională, pe lângă noi modele şi formalisme, este în primul rând de natură metodologică şi tehnologică: experimentul şi evaluarea. Ceace se numeşte astăzi lingvistică computaţională teoretică este în mare măsură asimilată cu lingvistica formală modernă. Acest segment al lingvisticii computaţionale a moştenit de la lingvistica matematică cel mai mult şi adecvându-şi metodele la realităţile 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 a 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 jumatate 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 direcţiei de preocupari prezentate în [1]; am în vedere numai modul în care aceasta direcţie este pusă in relaţie cu alte cercetări dedicate limbajului."

Aşa îşi începe domnul Academician Marcus articolul solicitat de mine pentru volumul „Limba Română în Societatea Informaţională-Societatea Cunoaşterii" rezultat al proiectului INFOSOC „SI-SC: Soluţii şi strategii în România". Să urmărim un prim citat incriminat (care în transcrierea dlui Academician este trunchiat si conţine nişte ghilimele ce nu-mi aparţin; redau mai jos varianta publicată) :

[1: p.133]:

"Din acest punct de vedere (*al folosirii calculatorului în prelucrarea limbajului natural – precizarea mea*), este semnificativ a arăta că însuşi numele domeniului de cercetare a prelucrării automate a limbajului natural a suferit modificări reflectând progresele ştiinţifice şi tehnologice: iniţial, 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), cautând soluţii abstracte de modelare generativă de tip universal a ceea ce se presupunea (la nivelul cunoaşterii ştiintifice a anilor 1960) a fi facultatea limbajului. "

Ce l-a supărat aici pe distinsul polemist? Ne spune chiar domnia sa:

„Nu ştiu ce intelege DT prin "lingvistica formală", o sintagmă nu prea folosită în perioada de emergenţă a lingvisticii matematice; exista lingvistica structurală (altceva decât ceea ce ar putea fi lingvistica formală, adica 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 „argumentaţia", 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*. Holt 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 căpatat 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, vol 3, 1957, pp375-408) la faimoasa carte a lui Chomsky (Syntactic Structures, Mouton, The Hague, 1957): „in a sense, transformational analysis is essentially a **formalization** of a long-accepted, traditional approach…". Citatul apare la pagina 387. Chomsky se pare că a apreciat termenul şi l-a adoptat, cel puţin în raport cu propria filozofie generativistă asupra limbajului.

„dar factorul determinant în naşterea LM, în a doua jumatate 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 preocupari 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 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). Până la sedimentarea elementelor definitorii ale unui domeniu de cercetare, pot coexista sau se pot succeda mai multe direcţii de cercetare. Dintre acestea unele pot dispare sau îşi pot diminua foarte mult influenţa în raport cu motivaţia iniţială. Ele 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 citit eu, cel ce pentru prima dată a sugerat idea folosirii calculatorului şi a tehnicilor de decodificare pentru prelucrarea automată a limbajului natural a fost Waren Weaver în 1946. În 1949 el scrie lucrarea **„Translation"** considerata de toti specialistii în traducere automata ca primul document programatic al acestei discipline. În 1952 a avut loc la Universitatea Georgetown din SUA prima conferinta dedicata exclusiv traducerii automate. În 1954, Peter Toma de la Universitatea Georgetown împreuna cu un grup de cercetatori de la IBM realiza primul experiment de traducere automata (engleza-rusa) folosind un dictionar de 250 de cuvinte si 6 reguli sintactice de rescriere. Acest sistem avea sa constituie nucleul faimosului program de traducere automata Systran pe care Peter Toma îl finalizeaza în 1973.

Punctul meu de plecare s–a aflat în lucrările unor Kulagina şi Melciuk, puternic implicaţi în studiile de traducere automată rusă–franceză, Yves

Lecerf, implicat în problemele de documentare automată, D. G. Hays, implicat în traducerea automată din rusă în engleză şi reciproc, B. Vauqois, cu preocupări de informatică lingvistică la Grenoble. De la ei, ca şi de la alti autori similari, am preluat în bună masură 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.

Traducerea automată, dar mai ales eşecul primelor încercări de rezolvare a acestui obiectiv încă nerezolvat sau nerezolvat complet, a constituit fără îndoială o motivație a „emergenței" LM. Aşa cum voi arăta pe larg mai departe, eşecul proiectelor de traducere automată au fost puse, prin interpretarea unilaterală şi tendențioasă a raportului APLAC, exclusiv pe seama inadecvării teoriilor lingvistice folosite atunci şi a cantonării în fapticul unor limbi particulare. Teoria „facultății înnăscute a limbajului" lansată de Chomsky, opunându-se tradiției tipologice de studiu lingvistic prin diversitatea limbilor, a generat o prodigioasă cercetare în direcția determinării principiilor gramaticii universale, în speranța că identificarea şi caracterizarea lor riguroasă le-ar putea operaționaliza atât pentru explicarea comunicării umane prin limbaj cât şi (un derivat subsidiar al obiectivului lui Chomsky) pentru realizarea de sisteme de traducere automată apropiate de performanța umană.

Dubioasă mi se pare sintagma "soluții abstracte", probabil efectul unui obicei binecunoscut de a diaboliza abstractul.

Remarca de mai sus mă surprinde de două ori: mai întâi pentru că nu este nimic reprobabil în expresia „o soluție abstractă" (ba chiar dimpotrivă: "abstract = Care rezultă din separarea şi generalizarea însuşirilor caracteristice ale unui grup 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, considerand–o oarecum echivalentă cu LM; dar chiar daca nu acceptăm aceasta echivalență, nu putem eluda faptul că lingvistica formală se află în imediata vecinatate 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 incercat", sugerând astfel ca ea a eşuat in 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 aparţin):

- 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): Prefaţa la „Special Issue on Using Large Corpora", Computational 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-based 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 practical 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 offers hope for similar progress in NLP."

- Nancy Ide and Jean Veronis (editori) Computational Linguistics –Special Issue on Word Disambiguation, Vol. 24, No. 1 1998 p.15:

> „Although quantitative methods were embraced in early MT work, in the mid-1960s interest in statistical treatment of language waned among linguists due to the trend toward the discovery of *formal linguistic rules sparkled by the theories of Zellig Harris (1951) and bolstered most notably by the transformational theories of Noam Chomsky (1957). Instead, attention turned toward full linguistic analysis and hence to sentences rather than texts, and toward contrived examples and artificially limited domains instead of general language*."

- Victor Yngve: From Grammar to Science:New Foundations for General Linguistics, John Benjamin Publishing Company, 1996:

> „there seems to be no scientific way of deciding among the many contenders…We fiind positions and methods being promoted like a new movie or defended with withering polemics or taken up like the latest fad…*We should abandon logical-domain theories entirely and move to the physical domain…Because this (notation) can be programmed on a computer it can be used to test large-scale models…Gone will be the babel of arbitrary grammatical notations, each to be discarded in turn*".

Deşi nu împărtăşesc în întregime poziţia extrem de radicală a lui Yngve, ea este simptomatică pentru insatisfacţia generală faţă de abordările tradiţionale ale anilor '60-'80.

- R.F. de Bruine (editor) „Synthesis of Proposal for an RTD Programme by Users, Industry and Research in Language and Technology", DGXIII, Commission of the European Cummunities, September 1992:

> „There is a broad need to further understanding of linguistic phenomena in the context of computerising the analysis and generation of language. General research should be stimulated within the following three main topics:
>
> - research on the linguistic meaning representation at the various level of description, ranging from the lower (e.g. phonetic, morphological and syntactic) and better understood ones to the higher, scientifically more difficult ones (e.g. semantic, pragmatic, contextual and communicative ones). It is forseen that the former must yield results in the short to medium term. Even if the latter are long-term enterprises, they must be organised in way that ensures availability of usable intermediate results.
> - reasearch on *more adequate and efficient computational schemes for natural language processing (e.g. constraints based computing and quantitative aspects) providing the base for robust 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 interactiunea 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 dicţionar şi a unei gramatici computaţionale e mai greu de înţeles remarca mea anterioară. Domnul Academician Marcus nu s-a apropiat niciodată de problemele unei implementări şi prin urmare nu mă surprinde lipsa de înţelegere a diferenţei între o definiţie formală a unei gramatici (de exemplu) care se explicitează în câteva rânduri şi implementarea unei gramatici computaţionale care nu numai că nu încape în câteva sute sau mii de pagini dar reclamă o muncă exprimată convenţional în mii de oameni/an. Gramatica computaţională 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 importanţi şi fiind suplimentate cu numeroşi studenţi 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 societăţi. Şi experienţa a arătat că nici nu este necesar! Ralph Grisman, de la Universitatea din New York a demonstrat că programul sau de inducţie gramaticală, pe baza unui corpus de antrenare a generat o gramatică nucleu, a cărei „finisare" a durat mai puţin de două săptămâni şi, confruntată cu gramatica Alvey pe un text arbitrar a reuşit să analizeze mai multe fraze, cu alte cuvinte a demonstrat o mai mare acoperire lingvistică.

> *Nu mi–am imaginat niciodată că între LM şi LC ar putea avea loc o competiţie, prima definindu–se prin metoda (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 definiţie specială a LM, pe care am dori s–o aflăm.*

Nici nu există această competiţie decât în imaginaţia domnului Academician care sugerează mai sus că LC nu foloseşte matematica sau că atunci când o face, disciplina se numeşte LM. Ceea ce, aşa cum am arătat mai înainte, este fals. Elementele suplimentare, esenţiale şi definitorii sunt calculatorul, algoritmii eficienţi şi cunoştinţele cu care acesta trebuie „hrănit". O formalizare a procesului de înţelegere şi/sau producere a limbajului natural, de orice sorginte ar fi ea, nu este decât o ipoteză asupra unui fenomen încă neelucidat. Validarea acestei ipoteze este cheia care a diferenţiat LC de LM. În anexa acestei lucrări am furnizat două definiţii pentru LM şi LC. Prima definiţie (LM) aparţine lui Geoffrey K. Pullum and Andras Kornai iar cea de a doua (LC) se află în pagina WEB a Asociaţiei de Lingvistică Computaţională (al cărui membru sunt din 1985). Aş mai face precizarea că lingvistica teoretică modernă (în sensul precizat mai înainte) studiază limbajul nu numai cu ajutorul matematicii. Alături de matematică, sociologia, pshihologia, medicina şi ştiinţele cognitive constituie domenii ale cunoaşterii care sunt fundamental implicate în explicarea acestui miracol pe care îl reprezintă comunicarea inter-umană. Incapacitatea actuală de a realiza un procesor artificial de limbaj la nivelul performanţei şi competenţei umane se datorează nedescifrării (încă) a mecanismelor minţii şi creierului omului. Dihotomia structural-fenomenologic şi noile cercetări în direcţia unei ştiinţe integrative (reprezentată între alţii de lucrările de pionierat ale Academicianului Mihai Drăgănescu) sunt fără îndoială porţi deschise spre cunoaşterea, în viitor, mai exactă a minţii şi implicit a facultăţii limbajului. Până atunci, obiectivele LC (realizarea de sisteme automate capabile

să prelucreze limbajul natural) apelează la modele aproximative, a căror acceptabilitate se probează prin implementarea şi evaluarea lor pe date reale. Cum între afirmarea unui obiectiv de LC şi realizarea sa operaţională este o distanţă mare, pe care uneori cercetătorii fără o bază în tehnologia programării fie că o ignoră, fie nu vor (şi de multe ori nici nu sunt interesaţi) să o parcurgă, confuzia ce duce la auto-acreditarea într-un domeniu conex este explicabilă.

> *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 relevanţă, este însă simptomatic pentru viziunea sa limitativă în problema în discuţie.*
> *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 „soluţii 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, ţin 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 ştiinţă, chiar dacă foarte controversat, dar acest statut îi este conferit şi de onestitatea cu care s-a detaşat de creaţiile sale anterioare ce i-au adus notorietatea, dovedite (unele chiar de el însuşi) ca fiind depăşite, propunând soluţii 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 computaţională, TAG este cel mai departe de influenţa 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, invitaţie motivată printre altele şi de o deosebită apreciere pe care o demonstraţie 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 dispoziţie trei volume consistente de lucrări asupra TAG tratând foarte amănunţit motivaţiile lingvistice, proprietăţile computaţionale şi caracterizarea matematică. Aceste volume i le-am pus la dispoziţie şi domnului Academician Marcus. Profesorul Joshi a fost in 1997 invitatul profesorului Dan Cristea şi al meu la Şcoala de Vară EUROLAN unde a susţinut o serie de prelegeri de înaltă ţinută ştiinţifică. 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 lui 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 a lui Chomsky (limbajele naturale nu sunt limbaje independente de context) demontată în anii '80 de Gerald Gazdar (autorul teoriei GPSG), cercetarea a fost orientată pe identificarea de formalisme lingvistice cât mai puternice, cu puterea generativă cât mai apropiată de cea a gramaticilor universale (echivalente deci cu maşina Turing). Formalismul ATN (Augmented Transition Networks) al lui William Woods de la BBN a fost timp de peste 10 ani suportul standard al majorității sistemelor de prelucrare a limbajului natural. Eu însumi am dezvoltat în anii 1984 şi 1985 un mediu de programare lingvistică conținând un editor de gramatici ATN şi un compilator ATN. Din punct de vedere formal ATN-ul este echivalent cu o maşină Turing şi tocmai această putere formală prea mare l-a scos din competiția soluțiilor utile în lingvistica computațională. La sfârşitul anilor '80 obiectivul major al LC (valabil şi astăzi) a devenit identificarea unui formalism de putere generativă cât mai mică dar care să acope cât mai multe din problemele practice puse de prelucrarea automată a limbajului natural. Aşa au revenit în actualitate gramaticile independente de context şi s-au dezvoltat abordările lexicalizate. Cele din urmă au fost propuse tocmai pentru a rezolva, în cadrul scheletelor de gramatici independente de context, idiosincrasiile limbajului natural cel mai adesea localizate la nivelul lexical. Mai mult, după anii '90, odată cu resurecția interesului față de abordările statistice, gramaticile regulate şi automatele finite au căpătat o utilizare foarte largă.

LC are mai multe părți, mai multe orientări, mai multe niveluri de abstracție, care comportă criterii diferite de evaluare.

Este adevărat că actualmente în LC se regăsesc orientări, abordări sau motivații diferite. Dar indiferent de sorginte, ele se plasează (cel puțin declarativ) în contextul computațional prin raportarea la un mediu software de prelucrare. Considerând exemplul HPSG, probabil cea mai în vogă teorie lingvistică computațională actuală, atunci când Ivan Sag analizează sau argumentează adecvarea teoriei sale în descrierea formală a unei limbii naturale (aşa cum a procedat în recentele sale conferințe la Facultatea de Litere a Universității Bucureşti şi în Aula Academiei Române) el se plasează în sfera lingvisticii teoretice. Atunci când prezintă soluțiile de implementare a unui fragment major al limbii engleze şi discută rezultatele generate de analizorul HPSG dezvoltat de grupul sau de la Universitatea Stanford şi modalitățile algoritmice de rezolvare a ambiguităților (aşa cum a făcut în prelegerea susținută la sediul RACAI, el se plasează în sfera LC.

DT îl asociază pe D. Hays la ideea sa privind falimentul LM şi lansarea, drept consecință, a LC.

Afirmația de mai sus conține două lucruri false:

a) nu am vorbit de falimentul LM ci de insuficiența **metodelor** sale la momentul invocat (cred că citatele pe care le-am prezentat şi 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 Academiei Naționale 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 computational 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 computational linguistics, as distinct from automatic language translation"). Dezvoltând ideile din capitolul raportului ALPAC referitor la prelucrarea limbajului natural (concept care şi atunci şi acum este diferit de cel al traducerii automate) Pierce considera că NSF (National Science Foundation) trebuia să asigure fonduri de cercetare pentru dezvoltarea de modele de limbă de dimensiuni mari „**since small-scale experiments and work with miniature models of language have proved seriously deceptive in the past, and one can come to grips with real problems only above a certain scale of grammar size, dictionary size, and available corpus**".

Acesta este contextul în care David Hays, activ cercetător la începutul anilor '60 în domeniul traducerii automate (de altfel unul din membrii comitetului ce au elaborat raportul ALPAC) a propus individualizarea preocupărilor legate de prelucrarea limbajului natural cu ajutorul calculatorului, dezvoltarea de modele de limbă realiste (nu miniaturi la îndemâna cercetării individuale) şi a aplicațiilor „serioase" (în opoziție cu experimentele la scară mică) sub numele de *lingvistică computațională*.

> *Denumirile folosite pentru preocupările la interferenta limbajelor, informaticii şi matematicii au variat tot timpul şi nu cred ca acest aspect merită prea multă atenție. Lingvistică matematică? computațională? inginerească? algebrică? cognitivă? aplicată? cantitativă? teoretică? statistică? probleme matematice ale semioticii? tehnologia limbajului? limbajul în inteligența artificială? lingvistica inginerească? procesarea limbajului natural? "information storage retrieval"? lingvistica cibernetică? pe fiecare dintre acestea am întâlnit–o şi propriile mele articole au fost publicate aproape sub fiecare dintre etichetele de mai sus.*

Citatul de mai sus mi se pare extrem de relevant pentru discuția de față şi defineşte clar diferența de opinii. Dacă de pildă distincția dintre *medicină umană şi medicină veterinară* sau (coborând în taxonomie) între cardiologie şi stomatologie „nu merită prea multă atenție" atunci domnul Academician are dreptate.

Din punctul meu de vedere însă, este o mare diferență între unele denominări ale studiului limbii amintite mai sus (la care se mai poate adăuga o listă la fel de numeroasă), ele definind câteva domenii distincte definite prin propriile obiective, competențe, metode şi modele.

> *În 1962 s–a înființat în USA "Association of Computational Linguistics".*

De fapt în 1962 s-a inființat AMTCL, acronim pentru „Association for Machine Translation and Computational Linguistics", primul preşedinte al AMTCL fiind Victor Ingve (cel pe care l-am citat mai devreme), iar al doilea fiind David Hays. ACL (Association of Computational Linguistics) a apărut abia în 1968.

> *În 1963 Ferenc Kiefer a demarat la Budapesta revista "Computational Linguistics", care a trăit peste zece ani.*

Este adevărat, dar conținutul ei era foarte diferit de al revistei „Mechanical Translation and Computational Linguistics' apărută în 1965 ca revistă oficială a AMTCL.

Şi tot ca un rezultat al diferenţierilor tot mai mari care apăruseră în domeniu, AMTCL îşi încetează activitatea la începutul anilor '70 fiind înlocuită de „American Journal of Computational Linguistics" care în 1984 devine „Computational 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 1965 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 „computational 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 conferinte 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 părti determinate ale CL (cum ar fi cea relativă la corpusul lingvistic). Etichetele nu au avut importanta si nu stiu sa se fi desfasurat vreo competitie 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 aparent foarte curioasă. Nu şi dacă observăm următoarele fapte:

- sintagma „lingvistică matematică" este din ce în ce mai puțin utilizată (o căutare pe internet a termenilor „mathematical linguistics", „computational

linguistics", „natural language processing"şi „language technology" este foarte instructivă: numărul de documente ce îi referă este 4.630, 87.900, 169.000 si respectiv 2.840.000);

- în domeniul strict computațional, la care se referea [1], în România activează de câtva timp o serie de cercetători importanți (majoritatea dintre ei membrii ai Comisiei de Informatizare pentru Limba Română pe care am onoarea să o conduc, şi din care de altfel face parte şi domnul Academician Marcus);
- domnul Academician Marcus fie nu cunoaşte, fie dezavuează rezultatele româneşti obținute în domeniul **prelucrării cu calculatorul** a limbii române (cel puțin aşa poate fi considerată ignorarea completă a acestora în lucrările domniei sale); ori poate consideră că nu 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. Referitor la „contradicția" pe care o semnalează în partea a doua a frazei de mai sus, nu pot să-i recomand domnului Marcus decât să citească încă de câteva ori articolul respectiv (sau să-l citească integral). Este vorba de **NOI modele formale de limbă** (în opoziție cu cele vechi), resurse lingvistice computaționale adecvate momentului actual. Dintre noile teorii care au apărut şi s-au şi impus aş putea să amintesc teoria optimalității în comunicare dezvoltată de Prince and Smolensky în 1993 (cu implementări în domeniul fonologiei şi morfologiei computaționale şi cu promițătoare rezultate chiar în sintaxă), teoriile sintactice bazate pe unificare şi satisfacerea de restricții, precum şi o întreagă pleiadă de teorii ale discursului. În domeniul prelucrării automate a limbajului natural există standarde, există tehnologii specifice, există organizații mondiale specializate, mai toate apărute în ultimii 10-15 ani. Dacă domnul Academician Marcus poate afirma că pentru limba română în domeniul resurselor lingvistice computaționale s-a făcut (sau a făcut) ceva înainte de anii '90 înseamnă că domnia sa are o imagine complet diferită de a 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 românesti 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 menţionate ca argument pentru concepte ce au apărut la începutul anilor '90 mă scuteşte de comentarii. Pe de altă parte, avansul ştiinţific în orice domeniu se clădeşte pe cunoaşterea anterioară iar cazurile de „frângere cognitivă", când salturile ştiinţifice neagă cunoaşterea anterioară sunt rare şi ele de regulă definesc revoluţiile în ştiinţă. Filiaţia sau influenţele în dezvoltarea unui domeniu ştiinţific (atunci când ele pot fi depistate cu obiectivitate) constituie preocuparea istoricilor ştiinţei. Lucrările tehnice, de regulă se raportează la contemporaneitate, ceea ce în termeni temporali poate însemna, în funcţie de dinamica domeniului, câţiva ani, un deceniu, mai multe decenii sau perioade chiar mai mari. De pildă, puţine lucrări tehnice în domeniul lingvisticii teoretice, al fonologiei se referă la marele gânditor Panini, considerat de mulţi oameni de ştiinţă creatorul ştiinţei limbii. Lucrarea sa fundamentală *Astaka,* cunoscută şi sub numele de „gramatica lui Panini" conţine descrieri formale ale regulilor de producţie 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 puţine cărţi sau lucrări de referinţă în aceste domenii menţionează 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.

Obstinaţia 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 imaginaţia domnului Academician. Citatul corect este: „în **general** al aspectului lor sintactic, gramatical".

> *DT defineşte "dimensiunea fundamentală" a LC prin "*fezabilitatea instanţierii unei descrieri lingvistice cât mai complete, mentenabilitatea acestei instanţieri şi, desigur, conformanţa cu realitatea uzului limbii*". ([1]: 133). Cu un mic efort intelegem 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 jumatate a anilor ``70, când instrumentele elaborate în informatica matematică deveniseră mult mai perfecţionate. Dar acest fapt nu tine, cum crede DT, de alegerea între LM şi LC, ci de progresul general realizat în ştiinţă. Pentru a mş referi la propria noastră experienţă, atunci când, in 1969, prezentam la COLING–ul din Suedia gramaticile contextuale nu aveam cum sa mă ocup de aspectul complexităţii acestor gramatici în maniera in care s–a putut face acest lucru ulterior (a se vedea, de exemplu, [10]). Dar acest fapt nu are nici o legatură cu eticheta folosită.*

Efortul (chiar mic) este probabil generat de unii termeni de specialitate nefamiliari domnului Academician. Voi furniza lămuririle necesare mai jos.

Eu mă refer la perioada actuală când invoc ca dimensiune fundamentală *fezabilitatea instanţierii* unei descrieri lingvistice cât mai complete. Instanţierea unei descrieri lingvistice înseamnă altceva decât complexitatea formală, de care de altfel şi amintesc în secţiunea trunchiată a citatului folosit de domnul Academician Marcus mai sus. Este un termen tehnic care se referă la construcţia propriu-zisă, în baza unui formalism sau teorii lingvistice, a unei gramatici şi a dicţionarului aferent, care furnizate ca resurse unui program de prelucrare a limbajului natural, permit acestuia să analizeze sau să genereze un text arbitrar. O astfel de instanţiere este fezabilă dacă ea se poate realiza în condiţii de timp şi resurse umane rezonabile.

> *Nu m–am mirat atunci când "Encyclopedia of Microcomputers" şi "Encyclopedia of Computer Science and Technology" mi–au solicitat o contribuţie cu tema "Semiotics and Formal Artificial Languages" (a se vedea [11]) si 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 "Computational Linguistics in the Netherlands 2000"[13].*

Nu văd rostul acestor lămuriri. Toată lumea îl ştie, îl recunoaşte şi nimeni dintre cercetătorii adevăraţi nu-l contestă pe omul de ştiinţă Marcus, important reprezentant român al lingvisticii matematice, creatorul acestei şcoli în România. În articolul [1] nu m-am referit nici direct nici indirect la domnia sa. Faptul că am evocat criticile pe care le-am comentat anterior la adresa **metodelor** lingvisticii matematice ale începutului deceniului şapte nu are nici o legătură cu realizările (încă o dată, excepţionale) ale domnului profesor. Însă probabil că identificându-se cu LM mondială, domnia sa a considerat critica asupra **metodelor LM** din anii '60 un atac la persoana sa, adevărat act de blasfemie.

În anii din urmă, domnul Academician încearcă să transfere în contextul noilor tendinţe şi tehnologii ale limbajului, ignorând o realitate existentă, tot portofoliul de rezultate pe care le-a obţinut 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 acţiunile recente declanşate de domnul Academician Marcus, pretinse a fi iscate de conţinutul articolului [1], nu fac decât să-mi întărească această impresie. Eu nu am nimic de împărţit cu domnul Academician.

> *Nu m–am mirat nici când am văzut că o revistă cu titlul "Linguistics and Philosophy" publică articole excelente de LC. Interferenţele merg în toate direcţiile şi ele caracterizează cultura contemporană. În acest orizont trebuie să ne plasăm, cred, atunci cand ne referim la disciplinele cognitive care se dezvoltă sub ochii nostri şi işi 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 sau nu figurează epitetul "computational". O revistă ca "Theoretical Linguistics" (1970*

2000), publicata de Walter de Gruyter (Berlin–New York) a inclus multe articole vizând aspecte matematice şi/sau computaţionale, deşi numele revistei nu indică acest lucru. Chiar o revista mai traditională, ca "Linguistics" a inclus de multe ori articole de LM şi nici "Foundations of Language" nu a procedat altfel. Multe fapte de LM si de LC se plasează în mod natural în orizontul semioticii computaţionale.

Faptul că tratatul amintit nu incorporează în titlu atributul computational 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 computaţională, conţinutul ei tratează exact ce anunţă în titlu: metode matematice folosite în studiul lingvistic. Lingvistica teoretică, puternic formalizată în ultimele decenii apelează inevitabil (ca de altfel marea majoritate a domeniilor ştiinţifice) la metode şi modele matematice.

Era internetului impune desigur o problematică nouă, faţă de care abordările anterioare se pot dovedi insuficiente.

Exact aceasta este esenţa celor 3 paragrafe din [1] incriminate şi combătute pe larg de domnul Academician Marcus: insuficenţa abordărilor anterioare. Conştientizarea acestei insuficienţe însă a precedat cu câtiva ani apariţia internetului.

Salutăm iniţiativa noii generaţii de cercetători de a se dedica noilor probleme.

Nu putem ignora tonul paternalist privind noua generaţie de cercetători care se dedică problemelor ridicate de internet în prelucrarea automată a limbajului natural. INTERNET-ul este o revoluţie! Şi implicaţiile sale sunt atât de mari încât asigurarea accesului universal la Internet a devenit o problemă fundamentală chiar şi pentru o organizaţie de caliblul UNESCO. Am avut onoarea să fac parte din Comisia de Experţi creată de Secretarul General al UNESCO (comisie de cel mai înalt nivel) pentru elaborarea documentului Recommendation on Multilingualism and Universal Access to Cyberspace. Sunt al doilea expert român (după dl. Ambasador Dan Hăulică, Membru Corespondent al Academiei) care a făcut parte dintr-o comisie de experţi UNESCO de acest nivel.

Ignorarea în cercetarea privind prelucrarea automată a limbajului natural a fenomenului INTERNET este de neconceput. Societatea cunoaşterii are ca una din premisele sale fundamentale accesul universal, neîngrădit de bariere lingvistice la 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 decât una care ţine seama în mod critic de experienţa acumulată.

Nimeni nu neagă acest lucru, şi faptul că l-am rugat insistent pe domnul Academician să facă parte din Comisia de Informatizare pentru Limba Română cred că arată buna mea credinţă şi speranţa pe care o nutream (şi care mai supravieţuieşte încă) că experienţa domniei sale va fi pusă în slujba obiectivelor pe care nici eu nici domnul Marcus nu le putem atinge singuri. În acelaşi spirit, i-am propus domnului Academician Marcus să scriem împreună o antologie a cercetărilor româneşti în domeniul lingvisticii formale şi computaţionale, de la inceputurile pe care le evocă domnia sa şi pînă în zilele noastre. Din păcate propunerea a rămas fără răspuns.

Din tot ceea ce am prezentat mai sus rezulta clar ca LM si 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 and Computational Linguistics", din [15], se afirma pur şi simplu (p.86): "Mathematical linguistics has also been called theoretical linguistics and even computational linguistics". Iar mai departe, în acelaşi loc: "Computational 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]).

Nu văd în pasajul pe care l-am citat mai sus nici un argument împotriva a ceea ce am susținut în [1] şi în cele prezentate aici. Notez în treacăt adverbul „even" cu o valoare discursivă în completă consonanță cu considerentele istorice pe care le-am invocat ale evoluției ştiințifice şi tehnologice în domeniul prelucrării limbajului natural.

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 "lingvistica aplicată" la Facultatea de Limba şi Literatura Română a Universității din Bucureşti, dar s–au găsit alții care să–i torpileze.

Aşa este, şi mă bucură elogiul adus acestor corifei ai ştiintei româneşti. Poate şi pentru că alături de câțiva reprezentanți importanți ai lingvisticii româneşti actuale care au înțeles tendințele şi imperativele momentului (Prof. Dan Mazilu-decanul Facultății de Litere, Prof. Alexandra Cornilescu, Conf. Emil Ionescu) am 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 Iaşi) la lansarea în 2001 a Masterat-ului în Lingvistică Computațională al Facultății de Informatică. Nu este uşor să pendulezi între Iaşi şi Bucureşti, dar şi domnul profesor Cristea, şi doamna profesor Cornilescu şi eu o facem pentru ca cele două programe „surori" de master 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 studenti străluciți ai unor facultăți umaniste la cercetarea limbii şi literaturii cu mijloacele LM şi 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 forţelor adverse, s–a reuşit în acei ani atragerea unor studenţi străluciţi ai unor facultăţi umaniste la cercetarea limbii şi literaturii cu mijloacele”* tehnologiei limbajului.

> *Pentru a da numai două exemple de actuali profesori universitari care au susţinut teze de doctorat de acest tip, voi mentiona pe Pia Brinzeu, de la Catedra de Engleză a Universităţii din Timişoara şi pe Mihai Dinu, de la Facultatea de Litere a Universităţii din Bucuresti. Tot în acea perioada şi-a sustinut teza de doctorat Sorin Cristian Niţă, pe o tema de critică textuală automată privind înlănţuirea (filiaţia) diferitelor variante ale “Istoriei Tării Româneşti” (Şerban Cantacuzino).*

Exemple de profesori şi cercetători români valoroşi, cu contribuţii substanţiale în domeniul limbii române se pot da foarte multe. Mulţi dintre ei sunt în străinătate şi fac o bună propagandă ştiinţei 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 computaţională în spiritul actual tocmai pentru că a parcurs acea cale dificilă a instanţierii lingvistice (în cazul său la nivelul componentului lexical).

> *Iată însă că, în pofida realităţilor puse in evidenţă mai sus, in ([1]: 134) se scrie: “În Romania, cercetările în domeniul LC şi al prelucrarii limbajului natural, precum şi primele rezultate practice au aparut 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 ştiinţifică, a aşezării numelor autorilor aceluiaşi articol în ordine alfabetică; dar nu acest fapt este cel care ne interesează în momentul de faţă).*

Inainte de a comenta acest pasaj şi pe cel următor, nu pot să trec peste observaţia absurdă şi falsă pusă între parantezele ce trădează totuşi o ezitare a probităţii omului de ştiinţă în faţa unei răutăţi gratuite. Nu există nici o regulă generală de genul celei afirmate. Ordonarea alfabetică este o convenţie între autorii cu contribuţii egale în redactarea unei lucrări. Am deschis la întâmplare două volume de specialitate, conţinând contribuţii (S. Amstrong et al. (eds) „Natural Language Processing Using Very Large Corpora, Kluwer, 1999 şi T. Strzalkovski (ed) „Natural Language Information Retrieval”, Kluwer, 1999). Din cele 19 lucrări cu mai mai mulţi autori, doar trei urmăresc (probabil din întţmplare) regula generală în lumea ştiinţifică pe care o invocă domnul Academician şi pe care probabil a impus-o şi o impune tuturor celor alături de care publică, indiferent de contribuţia fiecăruia.

> *Să observăm că încă în 1978, în articolul “Mathematical and Computational Linguistics” [9] de prezentare a activităţii din România în domeniul LM şi LC se face referire la peste 400 de articole publicate de 130 de autori români şi sunt menţionaţi peste 300 de autori străini (unii*

dintre ei, nume de vază ale LM şi LC din acea perioada) care au citat şi continuat cercetările românesti. Să mai adăugăm că numeroşi lingvisti români dintre cei mai importanți au citat şi folosit rezultatele şcolii românesti de LM şi LC. Iată ca vine acum DT şi face (deliberat sau nu) din tot acest efort un teren viran care–l astepta pe DT să tragă primele jaloane. Nu e cam mult?

Deşi am repetat de nenumărate ori până în acest moment, o mai fac o dată, precizând că discuția din [1] se referea la **resurse lingvistice computaționale** şi **programe software de dialog în limbaj natural** (în limba română). Acestea erau rezultatele practice pe care le menționam în citatul comentat cu gratuită aciditate. Poate să-mi menționeze domnul Academician vreun sistem de dialog în limba română implementat înaintea sistemelor pe care le-am realizat eu şi colaboratorii mei? Iată câteva repere:

- Sistemul QA (1980) un sistem inferențial de întrebare răspuns în limba română, susținut de un demonstrator original de teoreme în calculul 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;
- IURES (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 IURES a fost omologat internațional în 1988 şi a constituit primul produs de inteligență artificială exportat (în acelaşi an). Sistemele IURES şi SDLR sunt referite printre altele în enciclopedia de lingvistică computațională. Mai important este faptul că sistemele IURES 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 in general şi, în particular, asupra LM şi LC în România. DT a mai publicat, in urma cu câțiva ani, un articol în care se schița o privire istorică asupra LC în România, cu câteva citări la întâmplare, care tradau necunoaşterea situatiei reale.*

Cu rezerve față de prima parte a paragrafului, mă opresc la grija domnului Academician pentru dreptul noilor generații de studenți şi de cercetători asupra „informării corecte" asupra istoriei LM şi LC. Personal, cred că mult mai important pentre ei este să ştie prezentul şi tendințele viitoare ale domeniului. Astfel de cunoştințe le pot asigura un loc de muncă, o direcție de specializare, o carieră viitoare. Noile generații de studenți şi de cercetători sunt utilizatori pasionați ai Internetului. Acest uriaş ocean informațional le asigură un imens volum de cunoştințe, începând cu cursuri on-line (obligatorii pentru profesori la mai toate universitățile importante ale lumii), valome ale conferințelor sau articole extrem de utile, recente şi mai puțin recente, cărți electronice. Chiar şi relevante lucrări de istorie asupra diverselor domenii ştiințifice. Sistemele moderne de regăsire documentară le asigură şi o ierarhizare a acestor surse de informare în raport cu relevanța şi cu interesul manifestat de alți cititori. Listele de discuții sau arhivele de întrebări frecvente (FAQ) le pot oferi răspunsuri avizate şi obiective la întrebările ce-i preocupă. In anexă este furnizat un exemplu.

În ultima parte a citatului de mai sus, domnul Academician Marcus aduce în discuție o lucrare a mea din 1996 şi care arată că frustrările domniei sale sunt mai vechi. Articolul de care aminteşte domnul Academician mai sus, are titlul ***„Resurse lingvistice computaționale: trecut, prezent şi viitor"*** şi a apărut în volumul **„Limbaj şi Tehnologie"**, Ed. Academiei, 1996. Cei interesați, pot găsi articolul respectiv în pagina oficială a RACAI (*http://www.racai.ro* secțiunea publicații). Iar cele „*câteva citări la întâmplare, care trădau necunoaşterea situației reale*" apar în capitolul 2. **„Cercetări şi realizări româneşti în domeniul prelucrării automate a limbajului natural"**. Cred că titlul volumului, al articolului şi al capitolui sunt lămuritoare pentru ceea ce discutam acolo, dar probabil fraza, care trimitea la un volum editat de domnul Marcus, „abordările statistice, revenite acum în actualitate, au avut o tradiție strălucită (în România, adăugarea mea DT)" a fost prea scurtă şi insuficient de laudativă.

> *Mai este un aspect care cere o precizare. În conformitate cu specificul volumului în care apare articolul [1], DT face numeroase referiri la acte şi documente ale unor organisme europene şi internationale, cum este şi firesc, pentru a nu mai vorbi de aspectul financiar al colaborarii cu*

organismele respective.
Aceasta situaţie a existat de la începutul LM şi LC (chiar daca nu a avut amploarea de azi), datorită faptului că LM şi LC au apărut şi ca urmare a unor comandamente sociale, privind precaritatea mijloacelor de prelucrare a informaţiei. Imi amintesc de faimoasele Rapoarte CETIS care veneau de la EURATOM, Bruxelles, pe teme legate de analiza şi prelucrarea automată a limbajului, traducere automată şi documentare automată. În USA, diferite corporaţii (cum ar fi RAND Corporation, Santa Monica, Calif.) finanţau cercetări similare. O intâlnire semnificativă a fost aceea din 1962, organizată de "NATO Advanced Summer Institute", la Veneţia, Italia, privind traducerea automată. De numele acestui Institut este legat un document care a marcat evoluţia cercetărilor de traducere automată: seria de expuneri prezentate de Y. Bar–Hillel [17]. În legătură cu aceste activităti dirijate şi finanţate de diferite organisme europene şi internaţionale, trebuie să observăm că cei implicati au avut înţelepciunea 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 ştiinţific real al problemelor. Pentru a da un prim exemplu, mă voi referi la faptul că mai multe rapoarte CETIS au pus în discuţie un concept care, născut din experimentele de traducere automată, avea să se dovedească de o deosebită semnificatie pentru teoria sintactică în toată generalitatea sa; este vorba de conceptul de proiectivitate sintactică, cu consecinţe bogate în studiul structurilor arborescente şi al gramaticilor de dependenţă. Azi putem spune că si sintaxa limbajului natural şi teoria matematică a grafurilor au profitat esenţial de conceptul respectiv (folosit până şi de Rene Thom, în probleme de morfogeneză [17]). Această expansiune a unui concept sau rezultat dincolo de motivaţia sa initială este testul cel mai convingator al interesului său. Un al doilea exemplu se referă la titlul provocator folosit de Bar–Hillel pentru expunerile sale: "Patru conferinţe despre lingvistica algebrică şi traducerea automată".
Simpla alăturare a celor două sintagme, una foarte teoretiăa, cealaltă aparent tehnologică, avea menirea să–i avertizeze pe cei care presau să se obţină cât mai repede rezultate practice asupra faptului că proiectele de traducere automată nu se pot finaliza de azi pe maine, ci au nevoie de un lung itinerar lingvistic, matematic si computaţional. 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 esenţial cercetările de AI, cu consecinţe benefice pentru aspectele logice şi semantice ale limbajului natural.
Întrebarea pe care ne–o punem, dar o lăsăm deocamdată fără răspuns, deoarece nu suntem pregătiţi pentru a-l da, este urmatoarea: Nu cumva aspectele pe care le–am criticat mai sus sunt consecinţa unui fenomen mai general, acela al unui orizont insuficient de cuprinzător, al unei prea mari

dependenţe de factori utilitari imediaţi? Ştiinţa a oscilat mereu între cognitiv şi utilitar, dar istoria arată că functia utilitară s–a manifestat în toată profunzimea ei atunci când ea a fost fructul unei evoluţii fireşti a funcţiei cognitive, evoluţie 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 obligati totuşi să ţinem treaz şi celălalt ochi, îndreptat spre ceea ce se întamplă pe scena cercetării ştiinţifice vii, aşa cum apare ea în revistele de specialitate şi la întâlnirile ştiinţifice de profil.

Remarcile de mai sus îmi sugerează celebra fabulă cu strugurii cei acri. Cercetarea instituţionalizată (în opoziţie cu cea „de dragul artei") are motivaţii întotdeauna justificabile. Organismele de finanţare a cercetării, naţionale sau internaţionale, nu fac desigur acte de caritate. Obţinerea unei finanţări pentru un proiect de cercetare nu este la îndemâna oricui şi el implică nu numai abordarea unei probleme importante, dar şi credibilitatea grupului de cercetare. Evaluarea propunerilor de proiecte se face de către experţi recunoscuţi în domeniul respectiv, angajaţi şi plătiţi de agenţiile de finanţare a cercetării. În condiţiile unei concurenţe internaţionale acerbe pentru fondurile (din păcate prea mici) destinate cercetării, a lua în derâdere, invocând caracterul utilitar, cercetările ce obţin concurenţial finanţarea arată o desprindere de realitate. În luna martie a.c. am participat la evaluarea propunerilor de proiecte europene din cadrul Programului Cadru 5 (apelul 8), şi în calitate de raportor al direcţiei **„II.1.1 - Exploratory High Risk/Long Term Research"**, pot să afirm că propunerile de proiecte pe care le-am văzut erau foarte departe de a avea caracter utilitar. Domnul Academician Marcus lasă fără răspuns o întrebare cu răspuns sugerat, ridicând o problemă discutată cu ceva timp în urmă, anume a tipului de cunoaştere contemporană: enciclopedică (şi inerent generalistă) sau specializată. Cel puţin în domeniile tehnologice, viteza fără precedent a apariţiei de cunoştinţe noi face imposibilă cunoaşterea enciclopedică şi în acelaşi timp expertă pe toată lărgimea spectrului cunoaşterii actuale chiar şi într-un domeniu aparent îngust. Tehnologia limbajului este actualmente termenul ce subsumă toate preocupările legate de **prelucrarea automată a limbajului natural.** Cred că acest lucru spune totul!

3. In loc de concluzii

Ajungând în acest punct al răspunsului meu la atacul domnului Academician Marcus mărturisesc că mă încearcă un apăsător sentiment al deşertăciunii. Nu am dorit această polemică şi în nici un caz în acest context. Considerând că ea este nepotrivită faţă de obiectivele urmărite de proiectul „SI-SC: Soluţii şi strategii în România", în calitatea mea de director de proiect şi coeditor al volumului de faţă, am discutat cu membrii comitetului director al proiectului oportunitatea publicării polemicii domnului Academician Marcus (şi implicit a răspunsului meu) în volumul destinat unor probleme tehnice. Părerea a fost unanimă că nu este cazul să amestecăm obiectivele proiectului cu discuţia de faţă. Dar transmiţând domnului Academician această opinie şi făcându-i propunerea de a găzdui această polemică pe internet (în pagina oficială a RACAI) domnia sa s-a simţit cenzurat,

insultat şi îndreptăţit să facă o serie de afirmaţii pe care mă abţin să le comentez. Decizia de includere a acestei secţiuni în volumul de faţă am luat-o fără plăcere pentru că pe de o parte, în ciuda părerii domnului Academician Marcus (*Articolul meu se încadrează perfect în obiectivul pe care pretindeţi că-l urmăriţi şi în acest spirit a fost conceput. Realizaţi 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 incadreaza perfect in obiectivul*...îi va da domnului Academician satisfacţia pe care şi-a dorit-o, las cititorii să aprecieze cât de grav ar fi fost pentru obiectivul tehnologiei limbii române în contextul „Societatea Informaţională – Societatea Cunoaşterii: Soluţii şi strategii în România" ca cele două articole să nu fi apărut aici.

Referinte bibliografice (secţiune din lucrarea domnului Academician Marcus):

[1] D. Tufis. *Promovarea limbii romane in SI–SC*. In *Societatea Informationala – Societatea cunoa°terii* (coord. F. Gh. Filip). Ed. Expert, Bucuresti, 2001, 131–142.

[2] D. G. Hays. *The field and scope of computational linguistics*. Papers in Computational Linguistics (eds. F. Papp, G. Szepe). Proceedings of the Third International Meeting of Computational 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.* In *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.* In *Current Issues in Mathematical Linguistics* (C. Martin–Vide, ed.). Elsevier–North Holland, Amsterdam, 1994, 3–21.

[7] S. Marcus. *Lingvistica matematica, azi.* In *Matematica in lumea de azi si de maine* (C. Iacob, coord.), Editura Academiei, Bucuresti, 1985, 182–186.

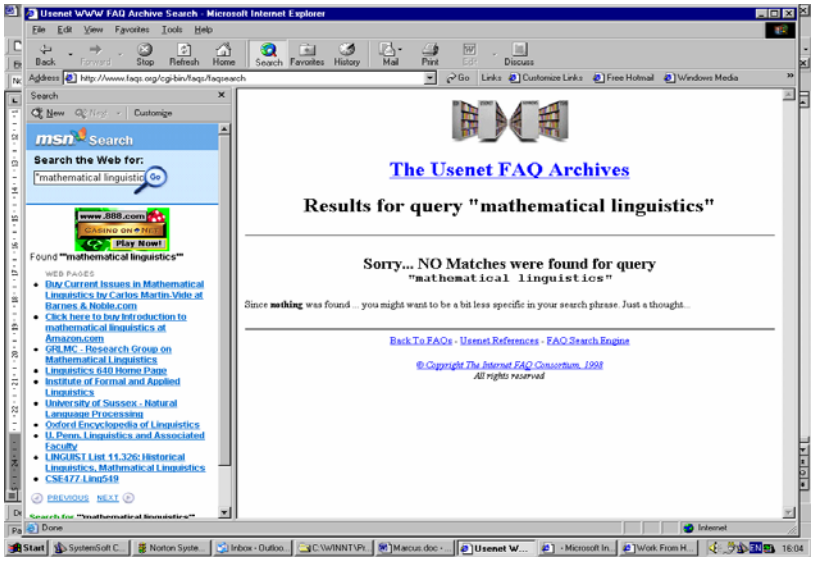
[8] S. Marcus. *Recent Romanian investigations in the field of mathematical and computational linguistics.* Avtomaticeskaja Obrabotka Tekstov, Matem. Fyz. Fakulta, KL Praha, 1973, 15–42.

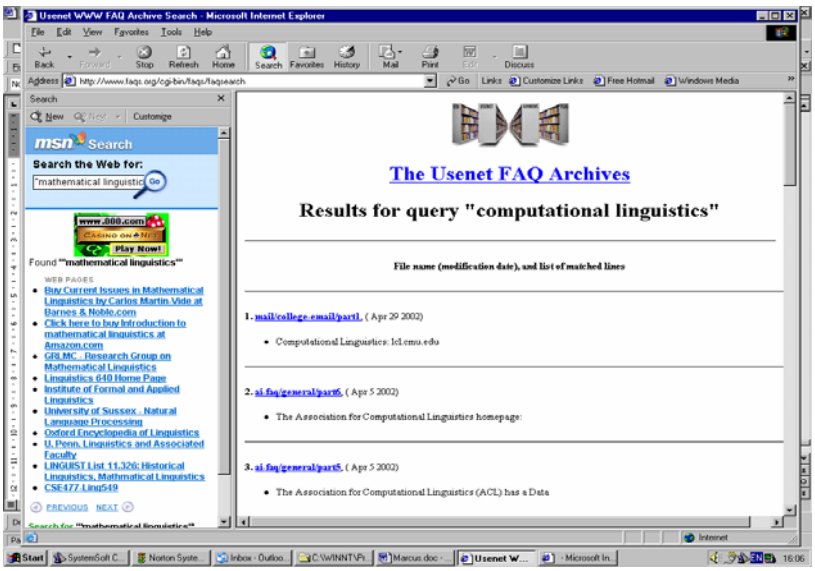
[9] S. Marcus. *Mathematical and computational linguistics.* In *Current Trends in Romanian Linguistics* (A. Rosetti, S. Golopentia Eretescu, eds.). Revue Roumaine de Linguistique 23, 1978, 1–4, 559–588.

[10] S. Marcus, C. Martin–Vide, G. Paun. *Contextual grammars as generative models of*

natural languages. Computational Linguistics 24, 1998, 2, 245–274.
[11] S. Marcus. *Semiotics and formal artificial languages*. In *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 Formal Languages* (G. Rozenberg, A. Salomaa, eds.), 2, Springer, Berlin, New York, 1997, 215–235.
[13] S. Marcus, C. Martin–Vide, G. Paun. *A new–old class of linguistically motivated regulated grammars. Computational Linguistics in the Netherlands 2000* (W. Daelemans et al., eds.), Selected Papers from the Eleventh CLIN Meeting, Ed. Rodopi, Amsterdam, New York, 2001, 111–125.
[14] B. H. Partee, A. Ter Meulen, R. Wall. *Mathematical Methods in Linguistics*. Kluwer, Dordrecht, 1990.
[15] E. F. Beckenbach, Ch. B. Tompkins (eds.). *Concepts of Communication: Interpersonal,* Intrapersonal and Mathematical. John Wiley and Sons, New York, 1976.
[16] D. G. Hays. *Introduction to Computational Linguistics*. American Elsevier, New York, 1967.
[17] R. Thom. *Stabilite Structurelle et Morphogenese*. John Benjamins, New York, 1970.
[18] Y. Bar–Hillel. *Four Lectures on Algebraic Linguistics and Machine Translation* revised version of a series of lectures given in July 1962, before a NATO Advanced Summer Institute, Venezia, Italy.

ANEXA1: Exemple de căutare într-o arhivă de întrebări frecvente (Usenet FAQ)





ANEXA 2: Definiții

What is Mathematical Linguistics?

MATHEMATICAL LINGUISTICS is the study of mathematical structures and methods that are of importance to linguistics. As in other branches of applied mathematics, the influence of the empirical subject matter is somewhat indirect: theorems are often proved more for their inherent mathematical value than for their applicability.

Both in phonology/morphology and in syntax/semantics the choice of linguistic formalism is to some extent influenced by considerations that go beyond the primary issue of descriptive adequacy. One important issue is Recognition Complexity. This concerns the complexity of the decision problem for membership in a language: it is assumed that a grammatical theory should have the property of guaranteeing that there is some reasonably rapid (polynomial in the lenght of the input) computation that will answer the question of whether a given sequence of words is a grammatical expression according to a given grammar. Human beings certainly do much more than this when they listen to an utterance and figure aut the meaning of what was said, so a grammatical theory that cannot even guarantee reasonably rapid confirmation of well-formedness is probably not psycholinguistically realistic. Another one is Learnability, which concerns what sorts of mathematically definable procedures could in principle correctly guess the grammars for languages.

(Geoffrey K. Pullum and Andras Kornai)

What is Computational Linguistics?

Simply put, COMPUTATIONAL LINGUISTICS is the scientific study of language from a computational perspective. Computational linguists are interested in providing computational models of various kinds of linguistic phenomena. These models may be "knowledge-based" ("hand-crafted") or "data-driven" ("statistical" or "empirical"). Work in computational linguistics is in some cases motivated from a scientific perspective in that one is trying to provide a computational explanation for a particular linguistic or psycholinguistic phenomenon; and in other cases the motivation may be more purely technological in that one wants to provide a working component of a speech or natural language system. Indeed, the work of computational linguists is incorporated into many working systems today, including speech recognition systems, text-to-speech synthesizers, automated voice response systems, web search engines, text editors, language instruction materials, to name just a few.

EDITURA Expert

București, România
Licența Ministerului Culturii nr. 1442/1992
Tel.: 411.60.75; Fax: 411.54.86
Consilier editorial: **Valeriu IOAN-FRANC**

ISBN 973-8177 - -