Lucrările atelierului
Resurse lingvistice şi instrumente pentru
Prelucrarea limbii române
Iaşi, 19-21 noiembrie 2008

Volum apărut cu sprijinul Ministerului Educaţiei şi Cercetării, prin Consiliul Naţional al Cercetării Ştiinţifice din Învăţământul Superior (CNCSIS)

ISSN 1843-911X

Lucrările atelierului
Resurse lingvistice şi instrumente pentru Prelucrarea limbii române
Iaşi, 19-21 noiembrie 2008

Editori:
Diana Maria Trandabăţ
Dan Cristea
Dan Tufiş

Organizatori:
Facultatea de Informatică,
Universitatea „Alexandru Ioan Cuza” Iaşi

Institutul de Cercetări pentru Inteligenţă Artificială
Academia Română, Bucureşti

Institutul de Informatică Teoretică
Academia Română, Filiala Iaşi

Editura Universităţii “Alexandru Ioan Cuza” Iaşi

COMITETUL DE PROGRAM:

Corneliu Burileanu, Facultatea de Electronică, Universitatea Politehnica Bucureşti şi Institutul de Cercetări în Inteligenţă Artificială, Academia Română, Bucureşti

Constantin Ciubotaru, Institutul de Matematică şi Informatică, Academia de Ştiinţe a Moldovei, Chişinău, R. Moldova

Svetlana Cojocaru, Institutul de Matematică şi Informatică, Academia de Ştiinţe a Moldovei, Chişinău, R. Moldova

Dan Cristea, Facultatea de Informatică, Universitatea "Al. I. Cuza" şi Institutul de Informatică Teoretică, Academia Română, Iaşi

Nicolae Curteanu, Institutul de Informatică Teoretică, Academia Română, Iaşi

Cristina Florescu, Institutul de Filologie Română "Al. Philippide", Academia Română, Iaşi

Corina Forăscu, Facultatea de Informatică, Universitatea "Al. I. Cuza", Iaşi şi Institutul de Cercetări în Inteligenţă Artificială, Academia Română, Bucureşti

Gabriela Haja, Institutul de Filologie Română "Al. Philippide", Academia Română, Iaşi

Radu Ion, Institutul de Cercetări în Inteligenţă Artificială, Academia Română, Bucureşti

Rada Mihalcea, Universitatea North Texas, SUA

Vivi Năstase, EML Research, Germania

Constantin Orăsan, Universitatea Wolverhampton, Anglia

Oana Postolache, ISI - Universitatea California, SUA

Irina Prodanoff, ILC-Pisa şi Universitatea Pavia, Italia

Georgiana Puşcaşu, Universitatea Wolverhampton, Anglia

Violeta Sereţan, Departamentul de lingvistică, Universitatea Geneva, Elveţia

Valentin Tablan, Universitatea Sheffield, Anglia

Amalia Todiraşcu, Universitatea Marc Bloch, Strasbourg, Franţa

Doina Tătar, Universitatea "Babeş-Bolyai", Cluj-Napoca

Horia-Nicolai Teodorescu, Institutul de Informatică Teoretică, Academia Română şi Universitatea Tehnică „Gh. Asachi”, Iaşi

Dan Tufiş, Institutul de Cercetări în Inteligenţă Artificială, Academia Română, Bucureşti şi Universitatea "Al. I. Cuza", Iaşi

Adriana Vlad, Facultatea de Electronică, Universitatea Politehnica Bucureşti şi Institutul de Cercetări în Inteligenţă Artificială, Academia Română, Bucureşti

COMITETUL DE ORGANIZARE:

Dan Cristea, FII-UAIC şi IIT-AR (dcristea@info.uaic.ro)

Corina Dima, FII-UAIC (cdima@info.uaic.ro)

Maria Husarciuc, LITERE-UAIC şi FII-UAIC (mhusarciuc@gmail.com)

Adrian Iftene, FII-UAIC (adiftene@info.uaic.ro)

Mihai-Alex Moruz, FII-UAIC (mmoruz@info.uaic.ro)

Ionuţ Pistol, FII-UAIC (ipistol@info.uaic.ro)

Diana Trandabăţ, FII-UAIC şi IIT-AR (dtrandabat@info.uaic.ro)

Dan Tufiş, ICIA-AR şi FII-UAIC (tufis@racai.ro)

Cuprins

CUVÂNT ÎNAINTE

Acest volum include lucrările celei de a şasea ediţii a Atelierului, în seria de manifestări ale Consorţiului de Informatizare pentru Limba Română, devenită oglindă a rezultatelor cercetărilor ce sunt dedicate în fiecare an domeniului Tehnologiei Limbajului (TL) din perspectiva limbii române. Volumul identifică principalele direcţii şi progresele realizate în acest domeniu pe parcursul anului 2008.

De data aceasta am grupat cele 15 lucrări acceptate (din 20 primite) în 3 capitole: Resurse lingvistice şi instrumente pentru prelucrarea vorbirii, Platforme, dicţionare şi corpusuri adnotate pentru prelucrarea textelor şi Aplicaţii ale tehnologiilor lingvistice, modificând aşadar structura din anii trecuţi, pentru că am considerat că această nouă împărţire reflectă mai adecvat orientările actuale ale lucrărilor ce au primit acceptul Comitetului de Program.

Lucrările incluse în volum descriu cercetări desfăşurate în diverse proiecte naţionale sau internaţionale precum şi rezultatele obţinute de doctoranzi în elaborarea tezelor lor. Îmbucurator este faptul că un număr din ce în ce mai mare de tineri talentaţi îşi aleg TL ca domeniu de cercetare. În cursul anului 2009, cel puţin şase dintre tinerii cercetători care au lucrări incluse în acest volum îşi vor finaliza tezele de doctorat, având contribuţii majore în prelucrarea limbii române în contextul multilingv al societăţii informaţionale.

La aproximativ o lună după Atelier, în 14 şi 15 ianuarie 2009, a avut loc la Luxemburg manifestarea *Language Technology Days*, organizată de *DG Information Society and Media Unit INFSO/E1 – Language Technologies, Machine Translation* şi prezidată de Roberto Cencioni şi Kimmo Rossi. Această importantă întâlnire a urmărit informarea în rândul potenţialilor propunători europeni de proiecte în domeniul TL asupra noilor oportunităţi de finanţare deschise de apelul 4, publicat în noiembrie 2008, al FP7-ICT (*Framework Program 7 of the Information and Communication Technologies*) şi de apelul 3 al ICT-PSP (*Information and Communication Technologies Policy Support Programme*), ce va fi publicat la sfârşitul lunii ianuarie 2009. Cele două direcţii principale menţionate în aceste apeluri se referă la traducerea automată şi la exploatarea multilingvă a web-ului (modele, arhitecturi şi instrumente pentru sisteme de traducere text şi voce auto-adaptive, standarde *de jure* şi de *facto* în gestiunea multilingvă a web-ului). În prima zi a evenimentului de la Luxemburg au fost invitaţi să prezinte situaţia domeniului TL în ţările lor trei experţi din ţări nou intrate în UE: România, Polonia şi Ungaria. România a fost apreciată de mulţi participanţi ca având un nivel competitiv, raportat chiar faţă de ţări cu tradiţie în domeniul TL, atât din punctul de vedere al învăţământului care pregăteşte specialişti, cât şi din cel al nivelului instrumentelor şi resurselor specifice dezvoltate deja. Credem că la această situaţie favorabilă au contribuit în mare măsură întâlnirile organizate de către ConsILR, schimbarea pozitivă de atitudine a Ministerului Educaţiei, Cercetării şi Tineretului faţă de problemele tehnologiei limbii române prin susţinerea unor proiecte semnificative ca ambiţie ştiinţifică şi nivel de finanţare, precum eDTLR (Dicţionarul Tezaur al Limbii Române în format electronic – https://consilr.info.uaic.ro/edtlr/wiki) ori SIR-RESDEC (Sistem de Întrebare-Răspuns în limbile Română şi Engleză cu Spaţii Deschise de Căutare – https://sir-resdec.racai.ro:450/), precum şi participarea specialiştilor români în proiecte europene importante, cum ar fi CLARIN (*Common Language Resources and Technology Infrastructure* – http://www.clarin.eu/), ALEAR (*Artificial Language Evolution on Autonomous Robots*) ori FlareNet (*Fostering Language Resources Network* – http://www.ilc.cnr.it/flarenet/).

Schimburile de experienţă facilitate de seria de întâlniri ale Atelierului contribuie la apropierea nivelului cercetărilor în tehnologia limbii române de cel mondial, la statornicirea unei terminologii coerente în limba română pentru acest domeniu şi, nu în ultimul rând, la cunoaşterea eforturilor comune desfăşurate de specialişti din diferite centre ale României ori din afara ei. În acest sens, ne propunem ca la ediţiile viitoare ale atelierului să includem o secţiune nouă de prezentări care să fie dedicată descrierilor de proiecte naţionale sau internaţionale, furnizând informaţii la zi (participanţi, obiective, realizări etc.); o altă secţiune nouă ar putea fi dedicată discuţiilor de tip „brainstorming" pe marginea unor propuneri preliminare de proiecte comune ale membrilor Consorţiului.

În deschiderea manifestării din decembrie 2008, ale cărei lucrări le oferim publicului în această carte, Consorţiul a hotărât ca pe viitor manifestarea să se transforme din atelier de lucru în conferinţă cu participare internaţională, pe considerentul maturităţii ştiinţifice dobândite deja şi al notorietăţii ei în rândul cercetătorilor de pretutindeni care se preocupă de limba română prin prisma metodelor computaţionale.

Ianuarie 2009

Editorii

CAPITOLUL 1

RESURSE LINGVISTICE ŞI INSTRUMENTE PENTRU PRELUCRAREA VORBIRII

STUDIUL VARIAŢIEI INTONAŢIONALE ÎN LIMBA ROMÂNĂ LITERARĂ FOLOSIND O MĂSURĂ A DISTANŢEI PROZODICE

ADRIAN TURCULEŢ[1], VASILE APOPEI[2], DOINA JITCĂ[2]

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

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

aturcu@uaic.ro, vapopei@ iit.tuiasi.ro

Rezumat

În lucrare se prezintă posibilitatea de a evalua în mod obiectiv distanţa prozodică dintre contururile melodice observate la vorbitorii de limbă română proveniţi din aceeaşi zonă sau din zone diferite. Calculul distanţei prozodice se bazează pe utilizarea coeficientului de intercorelaţie dintre curbele frecvenţei F0, avându-se în vedere de asemenea, şi cele ale duratelor şi intensităţilor (energia) segmentelor vocalice. Aplicat la un corpus de enunţuri, calculul coeficienţilor de corelaţie confirmă, în general, rezultatele analizei auditive şi vizuale. Cel mai apropiat de perceperea auditivă şi vizuală este coeficientul de corelaţie aplicat contururilor frecvenţei F0.

1. Introducere

În cercetările lingvistice actuale a prozodiei s-a simţit nevoia de a evalua obiectiv (a confirma sau a infirma) concluziile percepţiei auditive asupra diferenţelor melodice ale unor rostiri, cu ajutorul foneticii instrumentale acustice. Fără a deveni superfluă sau a putea fi înlocuită, analiza auditivă a primit, în ultimele decenii, un sprijin solid şi obiectivat în sensul bazării pe cercetarea cantitativă obţinută prin analiza şi sinteza computerizată a semnalului vocal.

Într-o primă etapă a studierii prozodiei, lingviştii au beneficiat de rezultatele metodelor de analiză a semnalului vocal referitoare la extragerea componentelor armonice ale semnalului vocal şi în special a traseului frecvenţei F0. Atât percepţia auditivă, cât şi cea vizuală, ambele categoriale, subiective, par însă a avea nevoie de cuantificarea matematică. Astfel s-a născut ideea de măsurare a „distanţei prozodice" dintre curbele melodice ale vorbirii pe baza variaţiei tonului fundamental, a duratei şi intensităţii segmentelor vocalice din cadrul unităţilor intonaţionale.

În lipsa unei definiţii clare a conceptului de *distanţă prozodică*, acesta se poate raporta la conceptul mai larg de *distanţă lingvistică*, utilizat de cei care se ocupă cu tipologia lingvistică în încercările de a cuantifica similitudinea sau diferenţele dintre limbi diferite sau dintre varietăţile aceleiaşi limbi, de exemplu, între dialectele ei. De exemplu, cercetările auditive şi acustice arată că intonaţia moldovenilor este mai apropiată de cea a muntenilor, în timp ce ardelenii au o intonaţie mai deosebită, având deci o „distanţă prozodică" mai mare.

Compararea curbelor frecvenţei F0 cu ajutorul statisticii matematice[1] oferă posibilitatea

[1] De aceea, utilizarea restrânsă a termenului *distanţă prozodică* se referă la o măsură a acesteia calculată cu ajutorul unor formule matematice şi nu la evaluarea auditivă sau vizuală a acesteia.

de a obiectiva „distanţa prozodică" vizuală. Un sprijin în acest sens a primit cercetarea prozodiei de la domeniul aplicat al învăţării limbilor străine sau al deprinderii intonaţiei de către persoane cu hipoacuzie. Având în vedere situaţiile de antrenament în care se urmăreşte corectitudinea reproducerii unui contur tonal, cercetătorul olandez Dik J. Hermes (1998a, b) propune mai multe tipuri de măsurători pentru evaluarea automată a gradului de corectitudine, printre care şi indicele de corelaţie. El vorbeşte mai ales de „similaritatea/disimilaritatea auditivă şi vizuală a contururilor tonale", dar, uneori, şi de „distanţa" dintre contururi.

În cadrul proiectului AMPER[2] a fost reactivat conceptul de distanţă prozodică în cercetarea variaţiei diatopice a prozodiei, pentru evaluarea matematică a similarităţii/ disimilarităţii a două contururi prozodice aparţinând unor varietăţi diferite. A. Romano (1999*2001: 226-235) utilizează la compararea modelelor intonaţionale ale unor dialecte apropiate coeficientul de intercorelaţie a curbelor după formula lui Pearson; în timp ce indicii de similaritate a curbelor F0, precum şi indicii de corelaţie a energiei au o valoare ridicată (peste 0.90), indicii de corelaţie a duratei ating şi valori ceva mai scăzute. Ulterior, A. Romano şi R. Miotti (2008) utilizează indicii de intercorelaţie a curbelor F0 pentru evaluarea *distanţei prozodice* dintre o varietate veneţiană (levantina) şi varietatea iberică din Malaga. Pentru limba română s-a realizat un studiu al variaţiei diatopice a intonaţiei folosind o metodologie similară (Turculeţ et al. 2008).

În lucrarea de faţă, în secţiunea 2, se va prezenta metodologia de realizare a corpusului de analiza prozodică iar în secţiunea 3, metodologia de prelucrare a datelor. Secţiunile 4 şi 5 conţin câteva rezultate ale analizei prozodice perceptuale-vizuale, şi respectiv, cele ale analizei realizată pe baza coeficienţilor de intercorelaţie.

2. Prezentarea metodologiei de realizare a corpusului de analiză prozodică

Înregistrările corpusului de analiză prozodică au fost realizate prin anchete pe teren în opt centre culturale din ţară: Baia Mare, Braşov, Bucureşti, Cluj, Craiova, Iaşi, Oradea, Sibiu, Timişoara şi unul din Republica Moldova: Chişinău. Subiecţii selectaţi, câte unul de sex masculin şi feminin, au studii universitare şi utilizează în mod curent varietatea cotidiană a limbii române standard. Din anchete s-au selectat înregistrările cu rostiri „neutre" (adică rostiri fără focalizarea expresă a unui constituent şi fără conotaţii expresiv-afective evidente) ale unor enunţuri asertive şi interogative, ambele în variante afirmative şi negative.

Înregistrările au fost stocate în fişiere de tip „wav" al căror nume conţine informaţii despre punctele de anchetă, vorbitori (masculin sau feminin), structura morfologico-sintactică şi silabico-accentuală a enunţului precum şi, de modalităţile de rostire: afirmativă (*a*), negativă (*n*), interogativă (*i*) şi interogativă negativă (*m*). Fiecare enunţ a fost repetat de fiecare subiect de cel puţin trei ori, rezultând, la un corpus de 45 de fraze, un număr de 540 de fişiere de sunet. Aceste fişiere au fost adnotate la nivelul segmentelor vocalice cu ajutorul programului Praat, realizându-se un corpus de 540 de fişiere cu semnal vocal de tip *wav* şi acelaşi număr de fişiere cu etichete de tip *TextGrid*. Pentru calcularea distanţelor prozodice (vezi tabelele de mai jos) s-au selectat trei

[2] Acest proiect are ca obiectiv realizarea primului atlas prozodic romanic: *Atlas Multimédia Prosodique de l'Espace Roman.*

enunţuri cu structuri morfologico-sintactice şi silabico-accentuale diferite, rostite cu cele patru modalităţi diferite (codificate mai anterior cu *a, n, i, m*):

- *Nevasta vede-un căpitan* - având codificarea structurii prin grupul de litere *twk;*
- *Nevasta frumoasă vede-un căpitan* - codificată prin s*wk;*
- *Nevasta vede-un căpitan elegant* - codificată prin *twg*.

3. Prezentarea metodologiei de prelucrare a corpusului

Prelucrările asupra corpusului de voce au fost realizate în mediul de programare Matlab, conform metodologiei stabilite în cadrul proiectului AMPER, şi vizează extragerea următorilor cinci parametrii prozodici pentru fiecare segment vocalic (relaţia 1): durata, valoarea maximă a energiei şi frecvenţa F0 măsurată în câte trei puncte (**i**niţial, **m**ijloc, **f**inal).

$$\{D_1, E_1, (F_1^i, F_1^m, F_1^f); D_2, E_2, (F_2^i, F_{21}^m, F_2^f); ...; D_n, E_n, (F_n^i, F_n^m, F_n^f)\} \qquad (1)$$

unde *n* este numărul de segmente vocalice din cadrul rostirii.

Evoluţia valorilor acestor parametrii pe durata unei rostiri generează următoarele trei contururi pe care au fost folosite în calculul coeficienţilor de corelaţie: curba frecvenţei F0 (un contur stilizat), curba duratelor şi a energiilor segmentelor vocalice. Duratele segmentelor vocalice se calculează prin diferenţa reperelor de timp asociate fiecărei etichete. Din valorile energiei calculate pe ferestre de 160 de eşantioane, cu factor de suprapunere ½, se selectează valoarea maximă pe fiecare segment vocalic.

Din aceste valori se extrag cele trei curbe folosite pentru calculul coeficientului de intercorelaţie:

- curba frecvenţei F0 : $\{(F_1^i, F_1^m, F_1^f); (F_2^i, F_{21}^m, F_2^f); ...; (F_n^i, F_n^m, F_n^f)\}$
- curba duratelor: $\{D_1; D_2; ...; D_n\}$
- curba de energie $\{E_1; E_2; ...; E_n\}$.

Pentru fiecare rostire (fişier tip *wav*) rezultă un fişier cu un număr de parametri egal cu numărul de vocale multiplicat cu 5. Valorile rezultate pentru cele trei repetiţii realizate de fiecare tip de rostire (cu aceeaşi structură, aceeaşi modalitate şi acelaşi tip de subiect) Pentru cele cinci valori de parametri prozodici pentru fiecare vocală s-a calculat media şi dispersia. Valorile medii rezultate sunt salvate în fişiere de tip text, pentru fiecare segment vocalic, împreună cu reperele de timp ale acestora.

Informaţiile din aceste fişiere constituie datele pe baza cărora s-au calculat coeficienţii de similaritate pentru analiza comparativă a prozodiei pe baza textelor cu valorile medii a câte trei repetiţii. Coeficienţii de similaritate s-au calculat cu formula coeficientului de intercorelaţie a lui Pearson, C_{xy}, aplicată valorilor de pe curbele de frecvenţă, durată şi energie din două înregistrări selectate pentru comparaţie (relaţia 2).

$$C_{xy} = \frac{\sum_{i=1}^{N} (x_i - \bar{x}) * (y_i - \bar{y})}{N * \sigma_x * \sigma_y} \qquad (2)$$

unde: N = numărul de puncte ale curbelor pentru care calculează coeficientul de intercorelaţie;
x_i, y_i = valorile cu indicele i de pe curbele supuse analizei;
$\bar{x}$, $\bar{y}$ = valorile medii corespunzătoare curbelor supuse analizei;
σ_x , σ_y = valorile dispersiilor corespunzătoare curbelor supuse analizei

Analiza variaţiei diatopice a intonaţiei la nivelul limbii române standard a fost realizată prin valorile coeficienţilor de corelaţie calculaţi între contururile melodice ale aceluiaşi enunţ (cu aceeaşi structură sintactico-lexicală şi fonologică segmentală) rostit de către fiecare cuplu de subiecţi din celelalte nouă localităţi, raportate la rostirea subiecţilor bucureşteni considerată ca reprezentând intonaţia standard. Unele concluzii ale raportării modelelor intonaţionale stabilite prin analiza auditivă şi acustică la distanţele prozodice reprezentate de coeficienţii de corelaţie sunt prezentate în secţiunea 5.

4. *Câteva rezultate ale analizei prozodice perceptuale*

Analiza acustică atestă prezenţa unor modele intonative regionale în vorbirea literară. Aceste particularităţi prozodice sunt chiar mai persistente decât particularităţile fonetice segmentale sau lexicale, permiţând identificarea zonei de proveniență a vorbitorului.

Intonaţia subiecţilor din Bucureşti, a constituit punctul de reper pentru comparaţia cu modelele intonative folosite de ceilalţi subiecţi. Contururile intonaţionale stilizate prin eliminarea unor efecte de micro-prozodie[3] au fost reprezentate prin grafice care prezintă valorile medii rezultate din prelucrarea a trei rostiri ale fiecărui enunţ (cf. Fig. 1-3). Subiecţii bucureşteni utilizează modele intonaţionale standard (v. Dascălu-Jinga (2001, 2005), distingându-se de vorbitorii din alte zone şi printr-o intonaţie mai „economică", cu o extensiune tonală (pitch range) mică, utilizând doar două accente tonale, cel de pe prima silabă accentuată şi cel ce determină modalitatea de realizare a propoziţiei (asertivă/ interogativă). Alte accente lexicale sunt dezaccentuate în modelele neutrale. O uşoară focalizare se poate observa, uneori, mai ales pe cuvântul final al enunţului.

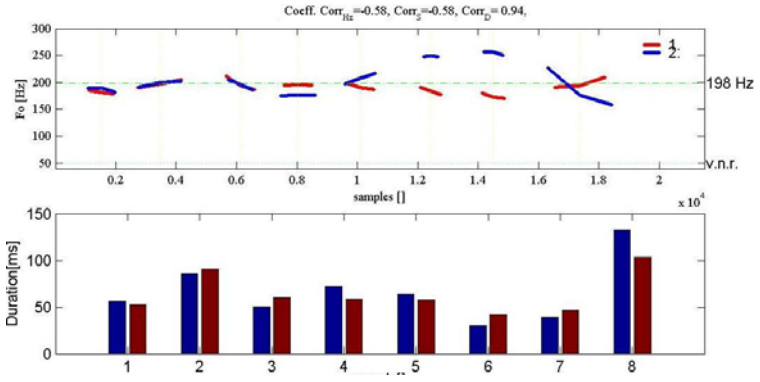

Figura 1: Contururile frecvenţei F0 şi ale duratei în rostirea subiectului feminin din Bucureşti (culoare deschisă), respectiv a subiectului feminin din Timişoara (culoare închisă) a enunţului *Nevasta vede-un căpitan?*

3 Analiza acustică a fost realizată după programe elaborate în cadrul proiectului AMPER de către A. Romano şi A. Rilliard.

La polul opus se află vorbitorii ardeleni, care prezintă aproape regulat o extensiune tonală largă (indiferent de registrul tonal al vorbitorului) în cadrul enunţului şi schimbarea pronunţată (de obicei urcarea) tonului fundamental pe silaba accentuată a fiecărui grup accentual. Aceasta dă impresia de „cântat" şi de „emfază" pe care o percep interlocutorii neardeleni ai vorbitorilor ardeleni[4].

Vorbitorii originari din Transilvania în sens larg (cu provinciile istorice adiacente Banat, Crişana, Maramureş) păstrează, atunci când folosesc stilul colocvial (adesea, chiar cel formal) al limbii standard, contururi intonative specifice, în special la interogativele totale afirmative şi negative. Conturul interogativ ardelenesc, în special în sintagma verbală, are o formă generală concavă, opusă formei convexe observate la vorbitorii din Muntenia şi Moldova, iar conturul melodic terminal este descendent, opus celui ascendent al modelului muntean-moldovean. Pe silaba accentuată a verbului începe o urcare treptată amplă a tonului fundamental, care se extinde asupra întregii sintagme verbale şi se termină printr-o coborâre abruptă pe ultima silabă accentuată a enunţului.

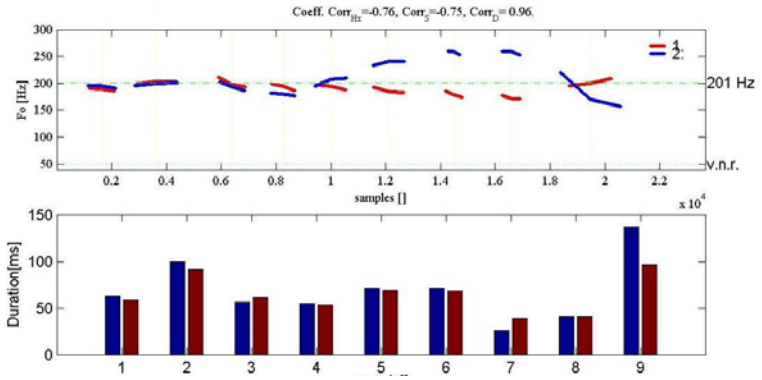

Figura 2. Contururile frecvenţei F0 şi ale duratei în rostirea subiectului feminin din Bucureşti (culoare deschisă), respectiv a subiectului feminin din Timişoara (culoare închisă) a enunţului „*Nevasta nu vede-un căpitan?"*.

În enunţurile cu structura SVO, tonul silabei accentuate a verbului, care poate fi şi plat sau uşor descendent, realizează o cezură între sintagma nominală şi cea verbală; cu un aliniament întârziat, F0 urcă pe sintagma verbală sub formă de cupolă sau de platou până la ultima silabă accentuată. În cazul interogativelor totale negative, adverbul „*nu"* poartă accentul sintagmei verbale, realizând cezura ritmico-sintactică urmată de conturul tonal cunoscut (figura 2). În partea de sud a Transilvaniei (Braşov, Sibiu), dar şi la subiecţii din Oradea, Baia Mare sau Chişinău, modelul acesta „ardelenesc"[5] este concurat de modelul muntenesc-moldovenesc standard.

Moldovenii, mai ales cei din partea de nord (subiecţii din Iaşi), se disting în special la enunţiativele negative printr-o emfază puternică a adverbului „*nu"*, care poartă accentul tonal principal (nuclear) al enunţului (v. figura. 3, culoarea albastră). Această trăsătură,

4 Unele teste de percepţie realizate cu ajutorul studenţilor originari din diferite zone ale ţării vor fi completate şi expuse cu altă ocazie.

[5] Am numit acest model intonational „ardelenesc" (între ghilimele), deoarece apare şi la subiecţii noştri originari din Chişinău.

emfatică la origine, s-a „gramaticalizat”, fiind prezentă în enunţurile rostite obişnuit, neutral.

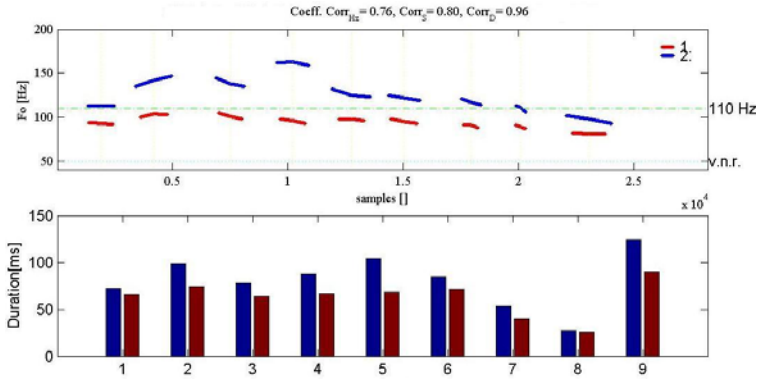

Figura 3: Contururile frecvenţei F0 şi ale duratei în rostirea subiectului masculin din Bucureşti (culoare deschisă), respectiv a subiectului masculin din Iaşi (culoare închisă) a enunţului *Nevasta nu vede-un căpitan.*

5. Rezultatele analizei prozodice pe baza măsurii distanţei prozodice

Înainte de a aplica măsurarea distanţei prozodice la compararea contururilor melodice realizate de vorbitori din zone diferite, pentru a determina mai obiectiv variaţia diatopică a distanţei prozodice, am probat aplicarea distanţei prozodice (măsurată prin coeficienţii la corelaţie ai frecvenţei F0) la: rostirile repetate ale aceluiaşi enunţ de către acelaşi vorbitor; acelaşi enunţ rostit de către fiecare dintre cei doi subiecţi (feminin şi masculin) din aceeaşi localitate; enunţuri cu modalităţi diferite (asertivă şi interogativă, afirmativă şi negativă) rostite de către acelaşi subiect.

Cum era de aşteptat, în primul caz, indicele de corelaţie a F0 este ridicat: peste 0.80, în cele mai multe cazuri peste 0.90; în cel de al doilea caz, există diferenţe (în afara registrului melodic diferit) care fac să coboare indicele de corelaţie spre 0,70 şi chiar spre 0,50. În al treilea caz, indicii sunt mai mari dacă se compară modalităţi cu contururi asemănătoare (de exemplu, *swkn: Nevasta nu vede-un căpitan* şi *swkm: Nevasta nu vede-un căpitan?* la subiectul feminin din Bucureşti: 0.61), dar sunt mici dacă se compară contururi tonale diferite (la acelaşi subiect, *swka: Nevasta vede-un căpitan* şi *swki: Nevasta vede-un căpitan?* : 0.22). Cu ajutorul indicelui de corelaţie a frecvenţei F0 se poate deci preciza dacă două enunţuri rostite de acelaşi vorbitor sau de vorbitori diferiţi au acelaşi *pattern* intonativ sau au *pattern*-uri diferite.

Compararea coeficienţilor de intercorelaţie corespunde, în cea mai mare parte, cu evaluarea distanţei/apropierii prozodice dintre localităţi realizată auditiv (la ascultarea enunţurilor) şi vizual (prin compararea traseelor F0 rezultate din analiza acustică). Rezultatele acestei comparaţii confirmă şi fundamentează concluziile lucrării privind existenţa unor pattern-uri intonaţionale diferite la vorbitori ai limbii literare provenind din zone diferite. Valorile absolute ale acestor coeficienţi se referă la contururile intonaţionale stilizate şi pot fi apreciate ca având o valoare relativă care depinde de reprezentativitatea subiecţilor selectaţi şi de relevanţa enunţurilor „neutrale” obţinute de anchetator.

Cele trei tabele de mai jos, conţin coeficienţii de intercorelaţie pentru contururile frecvenţei F0, duratelor şi energiilor ale celor trei propoziţii selectate (*twka, swka, twga*), rostite în cele patru modalităţi de către doi subiecţi (prima valoare aparţine subiectului feminin, iar cea de a doua subiectului masculin) din cele nouă localităţi menţionate, raportate la rostirile corespunzătoare ale subiecţilor bucureşteni. Aceste valori reprezintă *distanţele prozodice* (de frecvenţă F0, durată şi energie) dintre contururile intonaţionale ale aceloraşi enunţuri rostite de vorbitori bucureşteni şi de vorbitori din celelalte nouă centre culturale româneşti.

Tabel 1: Coeficienţii de corelaţie între contururile frecvenţei F0.

	Baia Mare	**Braşov**	**Cluj**	**Oradea**	**Sibiu**	**Timişoara**	**Iaşi**	**Chişinău**	**Craiova**
twka	0.82/ 0.85	0.88/ 0.94	0.98/ 0.91	0.88/ 0.94	0.98/ 0.91	0.97/ 0.89	0.74/ 0.80	0.75/ 0.93	0.95/ 0.55
twki	0.11/ 0.63	0.72/ -0.36	0.00 -0.06	0.29/ 0.51	0.66/ 0.00	-0.58/ -0.42	0.70/ 0.58	0.61/ 0.01	0.81/ 0.73
twkm	0.43/ -0.37	0.66/ -0.58	-0.80/ -0.62	0.29/ -0.56	0.69/ 0.04	-0.76/ -0.58	0.85/ 0.73	-0.66/ -0.30	0.73/ 0.59
twkn	0.78/ 0.84	0.81/ 0.89	0.84/ 0.88	0.87/ 0.73	0.85/ 0.80	0.94/ 0.83	0.81/ 0.76	0.85/ 0.76	0.92/ 0.30
swka	0.74/ 0.69	0.94/ 0.84	0.87/ 0.87	0.90/ 0.62	0.92/ 0.83	0.93/ 0.87	0.84/ 0.52	0.91/ 0.65	0.89/ 0.64
swki	0.35/ 0.32	0.67/ -0.11	0.11/ 0.20	0.31/ 0.05	0.72/ 0.22	-0.49/ -0.10	0.69/ 0.66	0.01/ -0.01	0.84/ 0.40
swkm	0.07/ -0.52	0.76/ -0.54	-0.62/ -0.53	0.01/ -0.42	0.45/ 0.01	-0.61/ -0.43	0.72/ 0.72	-0.62/ 0.12	0.68/ 0.31
swkn	0.68/ 0.83	0.93/ 0.88	0.79/ 0.92	0.78/ 0.67	0.90/ 0.77	0.66/ 0.86	0.63/ 0.76	0.84/ 0.76	0.94/ 0.79
twga	0.76/ 0.75	0.90/ 0.80	0.58/ 0.83	0.71/ 0.84	0.93/ 0.85	0.94/ 0.61	0.59/ 0.47	0.87/ 0.85	0.63/ 0.71
twgi	0.80/ 0.53	0.62/ -0.35	0.35/ 0.27	0.29/ -0.39	0.60/ 0.33	-0.56/ -0.23	0.62/ 0.48	-0.56/ -0.13	0.45/ 0.60
twgm	0.37/ -0.32	0.73/ -0.46	-0.77/ -0.63	-0.70/ -0.63	0.53/ 0.60	-0.73/ -0.54	0.84/ 0.70	-0.63/ -0.13	0.64/ -0.65
twgn	0.76/ 0.81	0.89/ 0.83	0.78/ 0.85	0.77/ 0.89	0.83/ 0.78	0.95/ 0.88	0.86/ 0.76	0.72/ 0.77	0.90/ 0.65

Tabel 2: Coeficienţii de corelaţie între curbele duratelor vocalelor

	Baia Mare	**Braşov**	**Cluj**	**Oradea**	**Sibiu**	**Timişoara**	**Iaşi**	**Chişinău**	**Craiova**
twka	0.86/ 0.61	0.91/ 0.75	0.88/ 0.57	0.92/ 0.83	0.86/ 0.51	0.79/ 0.63	0.89/ 0.88	0.84/ 0.68	0.89/ 0.72
twki	0.91/ 0.52	0.98/ 0.77	0.85/ 0.43	0.98/ 0.61	0.90/ 0.66	0.94/ 0.55	0.90/ 0.77	0.86/ 0.71	0.91/ 0.61
twkm	0.90/ 0.83	0.86/ 0.84	0.84/ 0.81	0.90/ 0.86	0.88/ 0.78	0.96/ 0.88	0.90/ 0.93	0.80/ 0.89	0.96/ 0.85
twkn	0.84/ 0.88	0.83/ 0.92	0.86/ 0.74	0.91/ 0.83	0.87/ 0.76	0.78/ 0.89	0.80/ 0.96	0.65/ 0.74	0.93/ 0.92
swka	0.87/ 0.86	0.97/ 0.89	0.89/ 0.79	0.85/ 0.82	0.92/ 0.80	0.89/ 0.88	0.91/ 0.95	0.84/ 0.88	0.86/ 0.63
swki	0.93/ 0.82	0.89/ 0.92	0.83/ 0.66	0.88/ 0.78	0.86/ 0.88	0.94/ 0.80	0.95/ 0.92	0.76/ 0.78	0.91/ 0.79
swkm	0.85/ 0.43	0.89/ 0.88	0.74/ 0.84	0.83/ 0.81	0.81/ 0.78	0.90/ 0.92	0.98/ 0.93	0.81/ 0.88	0.90/ 0.85
swkn	0.89/ 0.78	0.93/ 0.80	0.81/ 0.78	0.77/ 0.72	0.83/ 0.83	0.77/ 0.80	0.92/ 0.88	0.83/ 0.84	0.90/ 0.81
twga	0.90/ 0.79	0.93/ 0.70	0.82/ 0.71	0.89/ 0.68	0.88/ 0.78	0.92/ 0.58	0.89/ 0.72	0.89/ 0.67	0.95/ 0.65

twgi	0.88/ 0.87	0.92/ 0.94	0.85/ 0.75	0.94/ 0.86	0.91/ 0.69	0.94/ 0.74	0.84/ 0.83	0.85/ 0.84	0.90/ 0.75
twgm	0.86/ 0.88	0.86/ 0.92	0.85/ 0.78	0.94/ 0.96	0.91/ 0.76	0.89/ 0.84	0.86/ 0.92	0.86/ 0.91	0.91/ 0.74
twgn	0.87/ 0.83	0.82/ 0.87	0.86/ 0.84	0.84/ 0.90	0.82/ 0.79	0.87/ 0.79	0.79/ 0.84	0.79/ 0.89	0.88/ 0.71

Tabel 3: Coeficienţii de corelaţie între curbele energiei vocalelor

	Baia Mare	**Braşov**	**Cluj**	**Oradea**	**Sibiu**	**Timişoara**	**Iaşi**	**Chişinău**	**Craiova**
twka	0.65/ 0.81	0.78/ 0.77	0.82/ 0.90	0.91/ 0.87	0.93/ 0.90	0.54/ 0.74	0.69/ 0.75	0.56/ 0.81	0.67/ 0.78
twki	0.71/ 0.51	0.64/ 0.60	0.37/ 0.35	0.75/ 0.77	0.79/ 0.42	0.35/ 0.47	0.43/ 0.80	0.31/ 0.85	0.71/ 0.84
twkm	0.52/ 0.29	0.76/ 0.42	0.61/ -0.03	0.73/ 0.83	0.88/ 0.17	0.39/ 0.62	0.52/ 0.62	0.35/ 0.83	0.87/ 0.85
twkn	0.86/ 0.79	0.69/ 0.70	0.88/ 0.87	0.95/ 0.91	0.97/ 0.93	0.62/ 0.76	0.71/ 0.69	0.77/ 0.87	0.93/ 0.80
swka	0.65/ 0.54	0.52/ 0.59	0.60/ 0.74	0.96/ 0.71	0.92/ 0.85	0.72/ 0.45	0.53/ 0.48	0.56/ 0.72	0.46/ 0.84
swki	0.31/ 0.17	0.46/ 0.42	0.41/ 0.31	0.53/ 0.60	0.88/ 0.15	0.23/ 0.31	0.19/ 0.30	0.45/ 0.34	0.39/ 0.36
swkm	0.09/ 0.32	0.63/ 0.76	0.30/ 0.47	0.48/ 0.75	0.86/ 0.56	0.01/ 0.76	0.44/ 0.85	0.31/ 0.75	0.71/ 0.64
swkn	0.80/ 0.67	0.85/ 0.67	0.90/ 0.52	0.86/ 0.79	0.93/ 0.71	0.18/ 0.56	0.49/ 0.54	0.57/ 0.67	0.80/ 0.67
twga	0.94/ 0.68	0.83/ 0.62	0.78/ 0.56	0.91/ .51	0.95/ 0.81	0.55/ 0.59	0.57/ 0.59	0.65/ 0.75	0.58/ 0.57
twgi	0.57/ 0.63	0.22/ 0.38	0.58/ 0.64	0.82/ 0.75	0.81/ 0.54	0.20/ 0.73	0.13/ 0.59	0.20/ 0.46	0.62/ 0.59
twgm	0.66/ 0.32	0.66/ 0.57	0.58/ 0.74	0.48/ 0.68	0.90/ 0.69	0.29/ 0.47	0.45/ 0.64	0.04/ 0.66	0.69/ 0.31
twgn	0.81/ 0.66	0.90/ 0.67	0.90/ 0.67	0.56/ 0.86	0.91/ 0.90	0.54/ 0.76	0.64/ 0.80	0.81/ 0.73	0.66/ 0.69

Dintre cele trei tipuri de corelaţii prezentate în tabelele 1-3: pe baza variaţiei frecvenţei F0, a duratei şi a energiei, corelaţia frecvenţei F0 corespunde, în gradul cel mai înalt, evaluării perceptive a asemănărilor / deosebirilor dintre modelele intonative comparate. Pe baza coeficientului de corelaţie general (luând în consideraţie toate cele patru modalităţi) dintre subiecţii bucureşteni, pe de o parte, şi subiecţii din celelalte nouă oraşe cercetate, pe de altă parte), se poate vorbi de o distanţă prozodică relativ mică la subiecţii din Iaşi: 0.71, Sibiu: 0.63, Craiova: 0.62; medie la Braşov: 0.51 şi Baia Mare: 0.41; mare la Chişinău: 0.31, şi foarte mare la Cluj: 0.29, Oradea: 0.23 şi mai ales la Timişoara: 0.18.

O comparaţie între intonaţiile vorbitorilor din Bucureşti şi Ardeal conform tabelului 1 indică grade de similaritate diferite pentru asertive şi interogative, ultimele având o variaţie mult mai mare. Asertivele afirmative şi negative (codificate cu *twka, twkn, swka, swkn, twga, twgn*) au un indice de corelaţie mediu ridicat: 0.81; în schimb interogativele totale afirmative (*twki, swki, twgi*) au un coeficient de corelaţie mediu mic: 0.24, iar cele negative (*twkm, swkm, twgm*) au chiar coeficient negativ: -0.12. Ultimele distanţe prozodice, considerabile, se datorează modelului specific ardelenesc al intonaţiei enunţurilor interogative. Distanţa prozodică mai mică a subiecţilor moldoveni faţă de cei bucureşteni se explică prin menţinerea unui echilibru între toate cele patru modalităţi de rostire avute în studiu.

Un alt procedeu de a calcula coeficientul de corelaţie a curbelor, plecând de la exprimarea frecvenţei F0 în semitonuri, a dat valori identice sau foarte apropiate (cu diferenţe de 1-2 sutimi de procent).

Coeficientului de corelaţie între contururile intonaţionale ca măsură a distanţei prozodice i s-a reproşat faptul de a reflecta *în mod global* similaritatea curbelor F0, fără a ţine seama de punctele cele mai importante din punct de vedere funcţional ale conturului. De exemplu, o deosebire funcţională cum este plasarea tonului nuclear pe *nu* în asertivele negative ale ieşenilor nu se reflectă suficient în coeficientul mediu de corelaţie ridicat: 0.80; în fig.2, coeficientul este ceva mai scăzut: 0.76 (cu frecvenţa exprimată în Hz) şi 0,80 (cu frecvenţa exprimată în semitonuri).

Coeficienţii corelaţiei ai contururilor duratelor vocalelor sunt mult mai mari decât cei ai corelaţiei curbelor de frecvenţă F0, depăşind, în toate cazurile procentul de 0,80. Menţionăm, ca o curiozitate, faptul că subiecţii din Iaşi au şi în privinţa duratei cea mai bună corelaţie (0.88) cu subiecţii bucureşteni.

Datele din tabelul 3 pun în evidenţă o corelaţie scăzută între curbele de energie la nivelul global al rostirilor. Analiza perceptuală pune în evidenţă o posibilă corelaţie a curbelor de energie doar pe anumite porţiuni ale enunţului; de exemplu, la începutul asertivelor, unde se observă o concentrare a energiei sonore care scade treptat odată cu declinaţia, în timp ce la interogativele totale, energia creşte odată cu ridicarea finală a tonului. De asemenea, intensitatea vocalelor poate creşte considerabil pe constituenţii care poartă focusul contrastiv.

6. Câteva concluzii şi perspective

Folosirea coeficientului de intercorelaţie ca măsură a similarităţii contururilor frecvenţei F0 confirmă evaluările perceptive (auditive şi vizuale) cu privire la variabilitatea contururilor melodice utilizate de vorbitorii limbii române literare originari din zone diferite. Cele mai mari valori pentru gradul de similaritate se obţin în cazul contururilor medii care prezintă aceleaşi secvenţe de creşteri-descreşteri pe toată durata lor, indiferent de gama de variaţie în care se desfăşoară aceste secvenţe. Valoarea coeficientului de intercorelaţie este mai scăzută în cazul contururilor intonaţionale care manifestă tendinţe de evoluţie opuse pe perioade mai mari. La valori intermediare ale coeficientului de intercorelaţie avem de a face cu cazuri de similaritate pe grupuri sintactice.

Vom continua cercetările în vederea perfecţionării procedeelor de calculare a diferenţelor prozodice dintre contururile prozodice utilizate de vorbitori în enunţuri care să reflecte atât variaţia diatopică a rostirii, cât şi alte tipuri de variaţie: diastratică, diafazică, utilizând pentru aceasta corpusurile aflate la Seminarul de dialectologie şi sociolingvistică al Facultăţii de Litere şi corpusul SRoL (Teodorescu H.N, ş.a). Interesul actual pentru utilizarea distanţei prozodice şi ritmice între contururile intonative poate prilejui apariţia unei metode noi în cercetarea prozodiei: *intonometria,* cu două direcţii principale de dezvoltare: după modelul dialectometriei, care îşi propune cuantificarea asemănărilor şi deosebirilor dintre dialecte, găsirea unor procedee cât mai adecvate pentru calcularea *distanţelor prozodice* (şi *ritmice*) dintre varietăţi lingvistice în scopul

clasificării tipologice a acestora; după modelul fonometriei, care îşi propune, stabilirea cu ajutorul statisticii, a *normelor de rostire* ale unei comunităţi lingvistice. Intonometria poate avea ca obiectiv precizarea *modelelor intonaţionale neutre* ale unei comunităţi, precum şi a abaterilor de la acestea în scopuri funcţionale, pragmatice, expresiv-afective.

Referinţe bibliografice

Dascălu-Jinga, Laurenţia (2001), *Melodia limbii române vorbite*, Univers Enciclopedic, Bucureşti.

Dascălu-Jinga, Laurenţia, Organizarea prozodică a enunţului, în *Gramatica limbii române, II, Enunţul,* Editura Academiei Române, Bucureşti, 2005, p. 902-946.

Lai, Jean-Pierre & Albert Rilliard (2008): *Distance prosodiques entre les variétés occitanes et sardes* (sub tipar).

Hermes D.J. (1998a), Auditory and Visual Similarity of Pitch Contours, *Journal for Speech, Language, and Hearing Research,vol.* 41, 63-72, p.63-72.

Hermes D.J. (1998b), Measuring the Perceptual Similarity of Pitch Contours, *Journal for Speech, Language, and Hearing Research,vol.* 41, 73-82, p.73-81.

Mairano, Paolo & Antonio Romano (2008), *Distances rythmiques entre variétés romanes* (sub tipar).

Miotti, Renzo & Antonio Romano (2008), *Distanze prosodiche tra varietà friulane, romene e ispaniche* (sub tipar).

Rilliard, Albert & Jean-Pierre Lai (2007), La base de données AMPER et ses interfaces : structure et formats de données, exemple d'utilisation pour une analyse comparative de la prosodie de différents parlers romans*, I Jornadas Científicas AMPER-POR. Actas, Universidade de Aveiro*, 2007, p. 127-139.

Romano, Antonio (1999*2001), *Analyse des structures prosodiques des dialectes et d'italien régional parlés dans le Salento (Italie): approche linguistique et instrumentale*, Presses Universitaires du Septentrion, CEDEX France, 2001.

Romano, Antonio & Renzo, Miotti (2008), *Un contributo per il confronto tra l'intonazione veneta e quella andalusa* (sub tipar).

Teodorescu Horia-Nicolai, ş..a., *SRoL - Proiectul Sunetele Limbii Române*, www.etc.tuiasi.ro/sibm/romanian_spoken_language/index.htm

Turculeţ Adrian, Botoşineanu Luminiţa., Minuţ Ana-Maria, Mladin Ioan-Constantin. (2008), Aspects de la variation diatopique de l'intonation au niveau de la langue roumaine littéraire, Simpozionul internaţional *La Variation diatopique de l'intonation dans le domaine roumain et roman*, Iaşi , 20-21 octombrie 2008 (sub tipar).

DE CE NU PLACE VOCEA SINTETIZATĂ? – CÂTEVA ELEMENTE DE COMPARAŢIE CU VOCEA UMANĂ

HORIA-NICOLAI TEODORESCU[1,2], MONICA SILVIA FERARU[1,2]

[1] *Institutul de Informatică Teoretică al Academiei Române, Filiala Iaşi a Academiei Române*
[2] *Universitatea Tehnică Gheorghe Asachi din Iaşi*

{hteodor ,mferaru}@etc.tuiasi.ro

Rezumat

Prezentăm o scurtă analiză a diferenţelor care apar, la nivel de formanţi, între vocea sintetizată şi vocea naturală. După expunerea scopului analizei şi a metodologiei, prezentăm comparativ date privind valorile medii ale formanţilor pentru mai multe cuvinte şi fraze scurte. Pentru comparaţie s-a utilizat o voce umană din corpusul adnotat SRoL, pentru care frecvenţa fundamentală este foarte apropiată de cea a vocii sintetice. Comparaţia priveşte singurul sintetizor comercial autohton pentru limba română.

1. Introducere

Interesul nostru este, dincolo de cel strict practic, unul legat de aspectele cognitive, anume legat de întrebarea "cât de mult influenţează pattern-urile formantice învăţate de sistemul nervos calitatea perceptivă a vocii"? În măsura în care inteligibilitatea vocii (a mesajului transmis) este bună sau foarte bună, de ce sunt încă respinse în aplicaţii sintetizoarele vocalice? Care calităţi sunt lipsă şi cum sunt metrizabile aceste calităţi? Răspunsurile la aceste întrebări pot avea impact semnificativ de ordin teoretic (îmbunătăţirea modelelor cognitive ale audiţiei), ca şi de ordin aplicativ, privind creşterea calităţii sintetizoarelor comerciale.

2. Metodologie

Metoda de comparaţie priveşte două aspecte la nivel strict formantic static: valorile absolute ale frecvenţelor formanţilor şi valorile relative ale raporturilor frecvenţelor formanţilor raportate la frecvenţa fundamentalei pentru acea vocală. Analiza este statică în sensul că suntem interesaţi pe moment doar de valorile medii, dar nu şi de traseele formantice. Pentru comparaţie, s-a folosit o singură voce umană, aleasă pe criteriul similarităţii valorii frecvenţei fundamentale. Anume, s-a folosit înregistrarea cu codul 55555f din cadrul SRoL (www.etc.tuiasi.ro/sibm/romanian_spoken_language/index.htm). Vocea respectivă este feminină; fişa vorbitorului indică un vorbitor cu educaţie, fără patologie, voce regională din zona Moldovei, fără accent dialectal evident etc.

Rezultatele prezentate în secţiunea a treia se referă punctual la o voce umană şi la una sintetică. O comparaţie mai judicioasă ar trebui făcută pe mai multe niveluri:

- punctual, vocea sintetică cu câteva voci umane cu frecvenţe F0 apropiate, dar cu tonalităţi diferite (cu conţinut diferit de formanţi superiori);

- statistic, vocea artificială cu media şi intervalul de dispersie al valorilor formanţilor superiori;

- aceleaşi, pentru cazul dinamic, al traseelor formantice.

Cuvântul analizat este "Aseară", iar propoziţiile comparate sunt "Vine mama" şi "Cine a făcut asta?".

În finalul lucrării sunt menţionate sumar şi câteva observaţii privind dinamica formanţilor, aşa cu rezultă dintr-o analiză vizuală preliminară (Praat™ şi Wasp™). Aceste rezultate sunt doar calitative.

Pentru analiză, s-a folosit utilitarul Praat™; cuvintele au fost segmentate manual pe foneme şi, folosind zona centrală a vocalelor, am determinat cu utiliatarul Praat™ valorile medii pe vocale ale frecvenţelor fonemelor. Aceste valori au constituit baza analizei raportată aici.

Sintetizorul discutat este unul de tip concatenativ, destinat utilizării de către persoane fără vedere, sau cu vedere redusă. Sintetizorul este descris sumar în cadrul sitului comercial (http://www.baum.ro/index.php). Firma producătoare, BAUM Engineering, se prezintă ca având obiectul de activitate "Dezvoltare de Produse pentru Nevăzători şi Ambliopi"; firma a realizat demonstrativ sintetizorul TTS Online "Voce sintetică românească profesională Ancutza" v3.6.16., disponibil la adresa http://www.baum.ro/index.php. Pentru sintetizor sunt folosite setările default, anume Viteza: 60, Intonaţia: 60, Format: .WAV. Nu sunt date de către autorii parametrii sau modul de realizare ai sintetizorului respectiv şi prin urmare nu putem face aprecieri asupra posibilităţilor de îmbunătăţire a sintezei, la nivel tehnic.

3. Rezultatele analizei şi discuţie

Prezentăm preponderent sub formă grafică, rezultatele comparaţiei la nivel punctual şi static. Nu tratăm dinamica formanţilor.

Rezultatele pentru cuvântul "Aseară" (Teodorescu & Feraru, 2007), pronunţat independent – de exemplu, ca în cazul scurtului dialog "- Când ai fi dorit să mergi la film? – Aseară.", sunt rezumate în Tabelul 1.

Tabelul 1. Valorile absolute şi relative, raportate la F0, ale formanţilor, în cuvântul "Aseară". Voce de tip feminin sintetizată cu TSS comparativ cu vocea umană feminină cu indicativul 55555f din SRoL

	Aseară					
	a	**a**	**diftong ea**	**diftong ea**	**ă**	**ă**
	uman	sintetizat	**uman**	sintetizat	**uman**	sintetizat
F0	254	226	200	228	207	219
F1	593	752	652	769	456	785
F2	1634	1403	1892	1176	1658	1212
F3	2806	1978	2736	1963	3060	1827
F1/F0	2.33	3.33	3.26	3.37	2.20	3.58
F2/F0	6.43	6.21	9.46	5.16	8.01	5.53
F3/F0	11.05	8.75	13.68	8.61	14.78	8.34

Din tabelul 1, se observă că valoarea formantului F3 pentru vocea umană (3060Hz, pentru vocala *ă*, din cuvântul *aseară*) este semnificativ mai mare decât valoarea formatului F3 pentru vocea sintetică (1827Hz, pentru vocala *ă*, din cuvântul *aseară*).

În cazul raportului F1/F0, pentru diftongul *ea*, nu există diferenţe semnificative (3.26 faţă de 3.37) în comparaţie cu raportul F2/F0, care este ca valoare aproape dublu (9.46 faţă de 5.16) în cazul vocii umane comparativ cu vocea sintetizată.

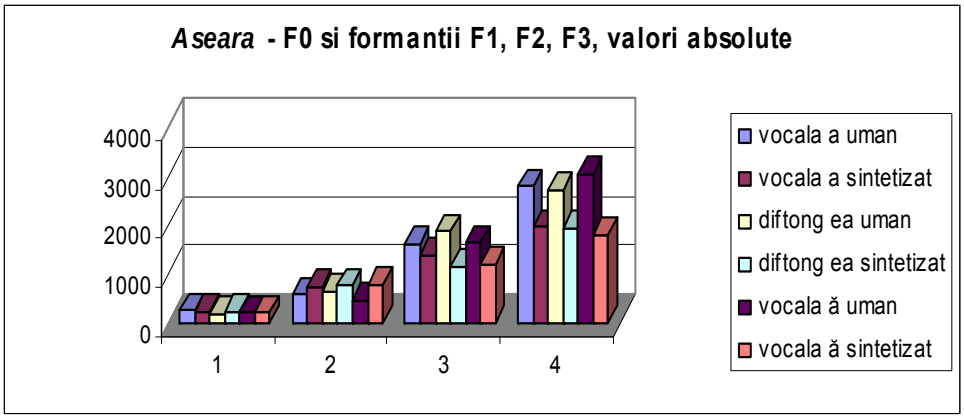

Figura 1 Evoluţia valorilor absolute ale frecvenţei fundamentale şi ale formanţilor pe vocala "a", diftongul "ea" şi vocala "ă" în cuvântul "Aseară"

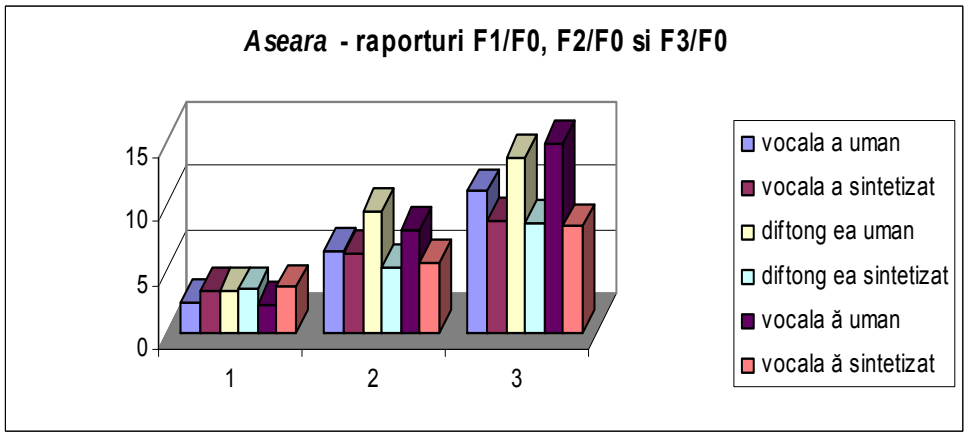

Figura 2 Evoluţia valorilor relative ale raporturilor F1/F0, F2/F0 şi F3/F0 pe vocala "a", diftongul "ea" şi vocala "ă" în cuvântul "Aseară"

În cazul valorilor relative (Teodorescu et al., 2007), se constată mari diferenţe la toate vocalele, la nivelul formantului F3, dar şi pentru diftongul *ea* şi vocala *ă*, între vocea sintetică şi cea umană (v. Fig. 2). Se observă că pentru formantului F1 nu există diferenţe semnificative, iar pentru formantul F2 diferenţe mai semnificative sunt pentru diftongul *ea* şi vocala *ă*, între vocea umană şi cea sintetică.

Tabelul 2. Valorile absolute şi relative, raportate la F0, ale formanţilor, în propoziţia simplă "Vine mama". Voce de tip feminin sintetizată cu TSS comparativ cu vocea umană feminină cu indicativul 55555f din SRoL

	i	**i**	**e**	**e**	**a1**	**a1**	**a2**	**a2**
	uman	sintetizat	**uman**	sintetizat	**uman**	sintetizat	**uman**	sintetizat
F0	222	234	242	226	196	225	200	218
F1	348	463	518	661	873	816	921	807
F2	2690	864	2242	1054	1255	1348	1427	1255
F3	3792	2450	3116	2083	2718	1928	3070	1642
F1/F0	1.57	1.98	2.14	2.92	4.45	3.62	4.60	3.70
F2/F0	12.11	3.69	9.26	4.66	6.40	5.99	7.13	5.75
F3/F0	17.08	10.47	12.87	9.21	13.87	8.56	15.35	7.53

În cazul vocalei *i* din cuvântul *vine*, valoarea raportul F2/F0 pentru vocea umană este depatru ori mai mare decât pentru vocea sintetică; în cazul vocalei *e* din cadrul aceluiaşi cuvânt este de două ori mai mare, iar în cazul ultimei vocale din cuvântul *mama*, valoarea valoarea raportului F3/F0 este dublă.

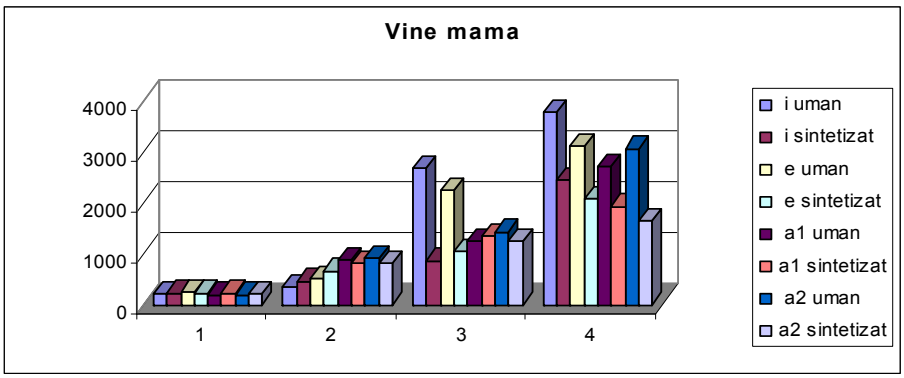

Figura 3 Evoluţia valorilor relative ale raporturilor F1/F0, F2/F0 şi F3/F0 pe vocala "i", "e", "a1"şi vocala "a2" în propoziţia "Vine mama"

Se observă diferenţe semnificative mai mari în cazul formanţilor F2 şi F3, pentru vocea umană comparativ cu cea sintetizată pentru următoarele vocale din cadrul propoziţiei "Vine mama": *i, e* şi ultimul *a* din cadrul cuvântului *mama*.

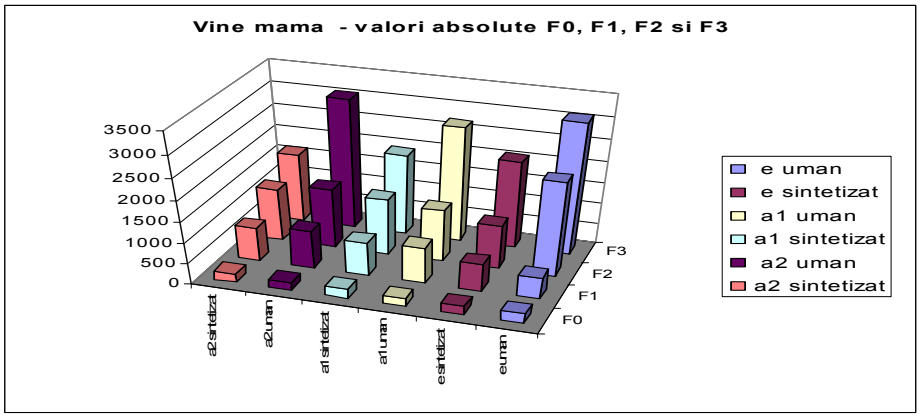

Figura 4 Evoluţia valorilor absolute ale frecvenţei fundamentale şi ale formanţilor pe vocala "i", "e", "a1"şi vocala "a2" în propoziţia "Vine mama"

În figura 4, se observă pentru vocala *a* (primul *a* şi ultimul *a* din cuvântul *mama*) la vocea umană, că valoarea formantului F3 este mult mai mare comparativ cu vocea sintetizată. Referitor la F0, formantul F1 şi F2, diferenţele sunt nesemnificative pentru toate vocalele din propoziţia "Vine mama".

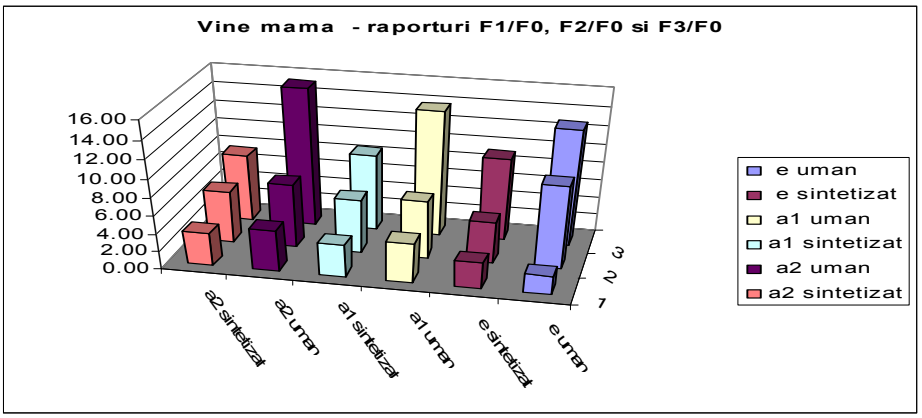

Figura 5 Evoluţia valorilor relative ale raporturilor F1/F0, F2/F0 şi F3/F0 pe vocala "e", "a1"şi vocala "a2" în propoziţia "Vine mama"

Similar, pentru propoziţia interogativă "Cine a făcut asta?" – pe care de altfel sintetizorul discutat nu o poate produce la forma interogativă, rezultatele comparative sunt prezentate în Fig. 6 şi Fig. 7.

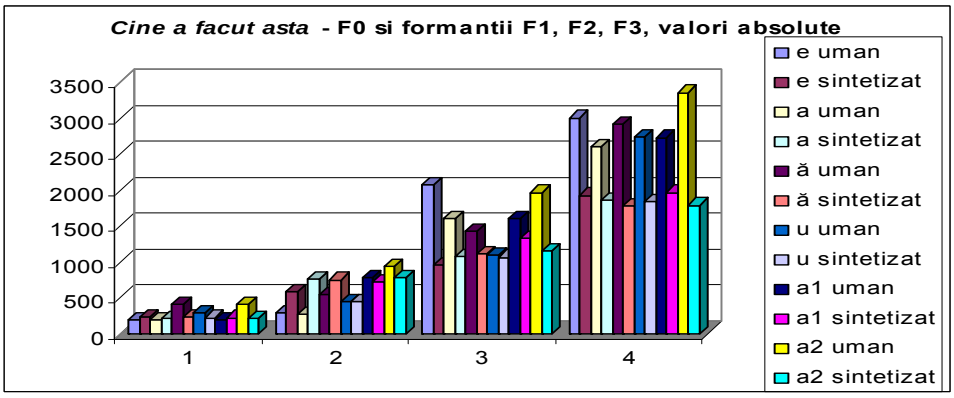

Figura 6 Evoluţia valorilor absolute ale frecvenţei fundamentale şi ale formanţilor pe vocala "e", "a", "ă", "u", "a1"şi vocala "a2" în propoziţia "Cine a făcut asta"

Graficele din Fig. 6, 7 corespund valorilor din Tabelul 3. Valorile F0 sunt semnificativ mari (duble) pentru vocala *ă* (din cuvântul *făcută*) şi vocala *a* (ultimul *a* din cuvântul *asta*) din propoziţia "Cine a făcut asta" în comparaţie cu vocea sintetizată. În cazul formantului F1, valorile sunt aproape duble în cazul vocii sintetizate comparativ cu vocea umană pentru vocalele *e* (din cuvântul *cine*) şi *a* din cadrul propoziţiei "Cine a făcut asta". Valorile F3 sunt ca valoare mari în cazul vocii umane comparativ cu vocea sintetică, iar raporturile F2/F0 şi F3/F0, în cazul vocii umane, pentru vocala *e* (din cuvântul *cine*) sunt aproape duble ca valoare comparativ cu vocea sintetizată.

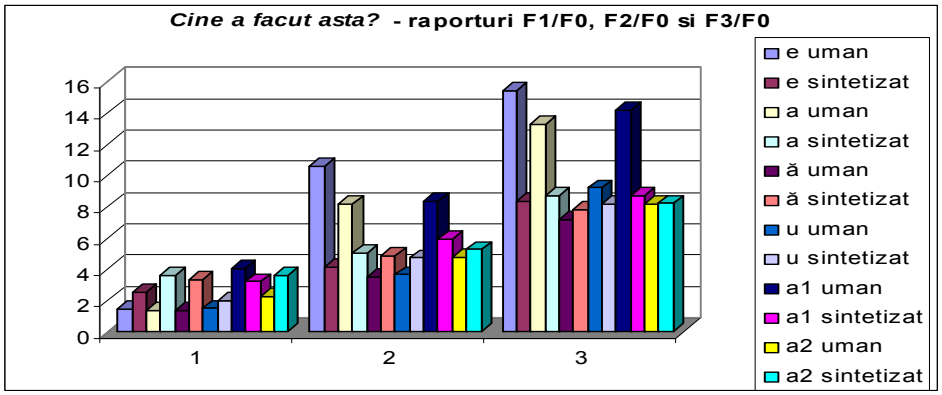

Figura 7 Evoluţia valorilor valorilor relative ale raporturilor F1/F0, F2/F0 şi F3/F0 pe vocala "e", "a", "ă", "u", "a1"şi vocala "a2" în propoziţia "Cine a făcut asta"

Tabelul 3. Valorile absolute şi relative, raportate la F0, ale formanţilor, în propoziţia simplă "Cine a făcut asta". Voce de tip feminin sintetizată cu TSS comparativ cu vocea umană feminină cu indicativul 55555f din SRoL

	e	e	a	a	ă	ă	u	u	a1	a1	a2	a2
	Um	Sint	**Um**	Sint	**Um**	Sint	**Um**	Sint	**Um**	Sint	**Um**	Sint
F0	195	231	196	213	406	227	298	225	192	225	411	218
F1	290	591	268	765	546	753	448	452	776	723	931	783
F2	2068	951	1609	1076	1424	1110	1090	1064	1604	1332	1966	1148
F3	3006	1925	2605	1864	2916	1781	2744	1840	2729	1968	3353	1791
F1/F0	1.49	2.56	1.37	3.59	1.34	3.32	1.50	2.01	4.04	3.21	2.27	3.59
F2/F0	10.61	4.12	8.21	5.05	3.51	4.89	3.66	4.73	8.35	5.92	4.78	5.27
F3/F0	15.42	8.33	13.29	8.75	7.18	7.85	9.21	8.18	14.21	8.75	8.16	8.22

Privitor la dinamica formanţilor (traseele formantice), sunt prezentate în Fig. 8, 9, 10 şi 11 imagini obţinute cu utilitarul Praat™ pentru propoziţiile "Vine mama" şi "Cine a făcut asta" (Feraru & Teodorescu, 2008), pentru vocea umană şi pentru cea sintetizată, iar în Fig. 12 şi 13 imagini similare obţinute cu utilitarul Wasp™, pentru propoziţia "Vine mama."

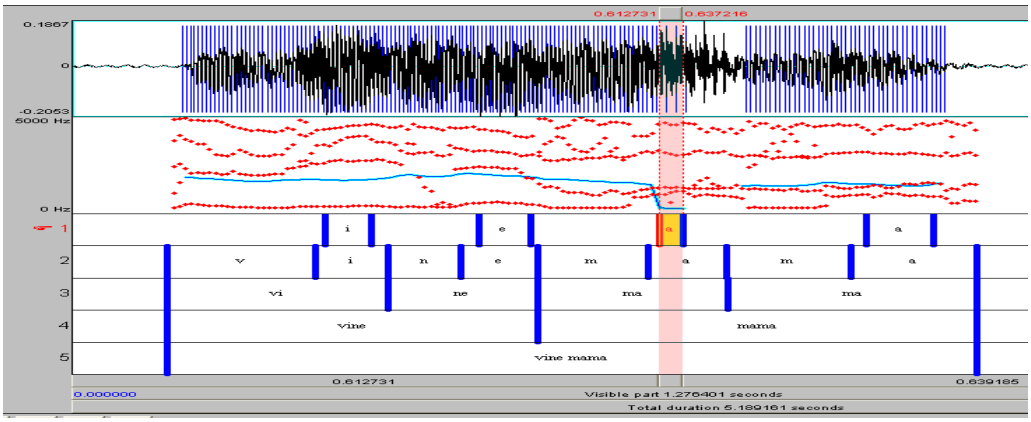

Figura 8 Exemplu de adnotare manuală folosind utilitarului Praat™; voce umană – propoziţia "Vine mama"

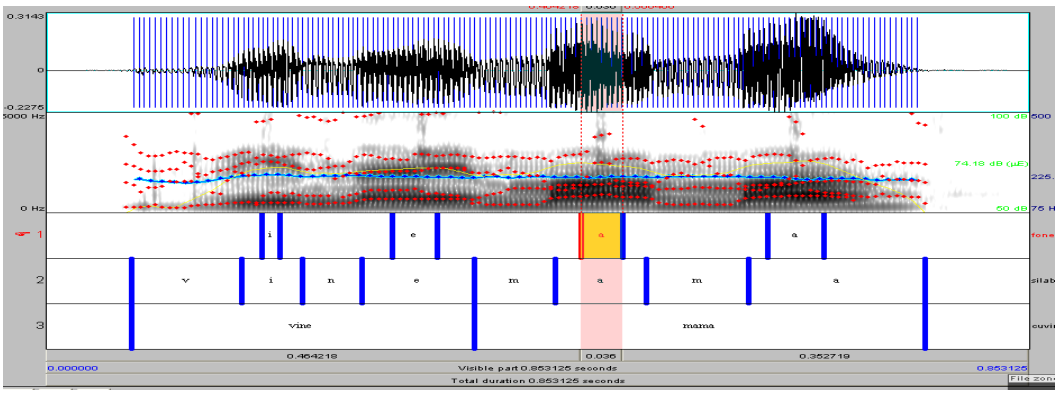

Figura 9 Exemplu de adnotare manuală folosind utilitarului Praat™; voce sintetizată – propoziţia "Vine mama"

În figura 8, în exemplul de adnotare pentru propoziţia "Vine mama", voce umană se observă o bună demarcare pentru formanţii F3, F4 şi o variabilitate bogată a traseelor, indicând o modulare prozodică bună, ne-monotonă.

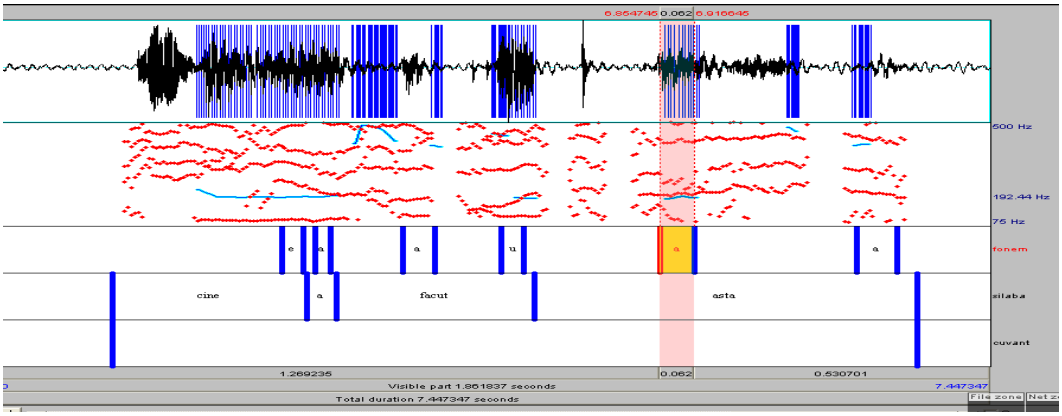

Figura 10 Exemplu de adnotare manuală folosind utilitarului Praat™; voce umană – propoziţia "Cine a făcut asta"

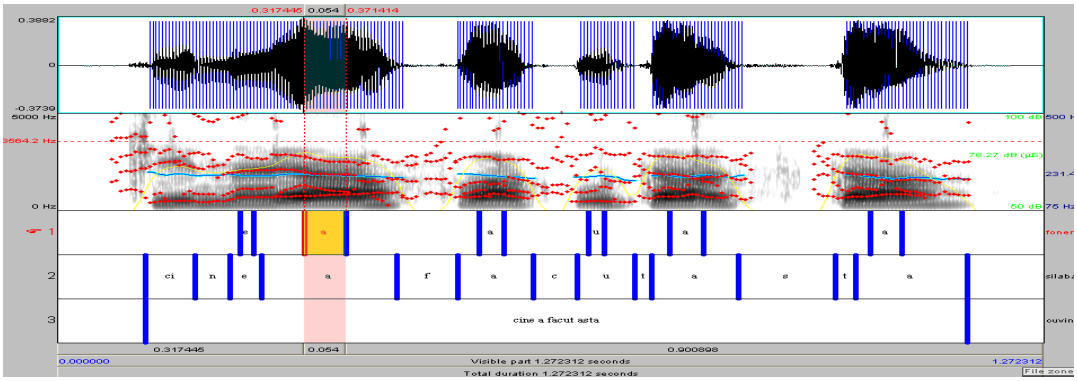

Figura 11 Exemplu de adnotare manuală folosind utilitarului Praat™; voce sintetizată – propoziţia "Cine a făcut asta"

Din compararea traseelor formanţilor superiori în Fig. 8-11, se constată că evoluţia acestor formanţi este discontinuă în cazul vocii sintetice, în timp ce în cazul vocii naturale traseele urmează curbe cu variaţie relativ lină. Chiar şi atunci când traseele lui

F0 şi F1 sunt asemănătoare pentru vocea sintetică şi cea naturală, traseele formanţilor superiori pentru cele două voci sunt total diferite. Considerăm că aceste diferenţe explică cel puţin parţial de ce vocea sintetică este “neplăcută”, induce senzaţia de “nenatural”.

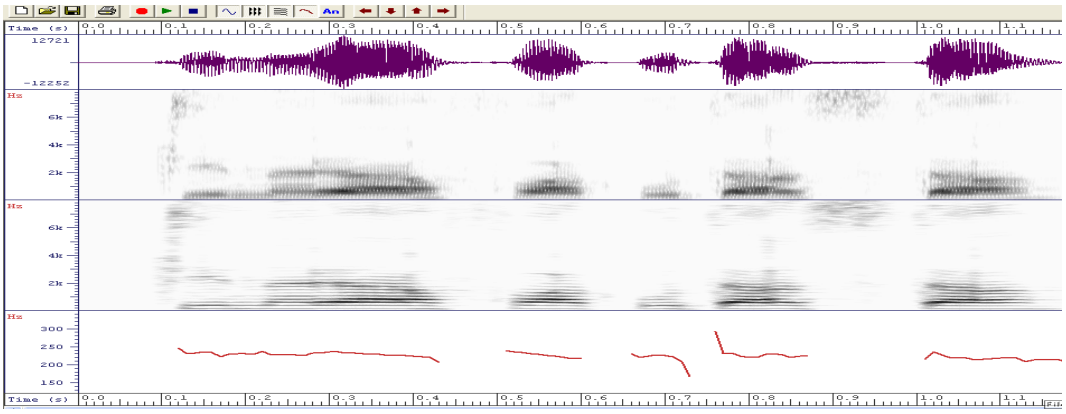

Figura 12 Traseul frecvenţei fundamentale (linia roşie) folosind utilitarul Wasp^{TM}; voce sintetizată - propoziţie “Cine a făcut asta”

În figura 12 se observă pentru vocea sintetică, discontinuităţi pentru F0, F1, F2 precum şi variaţii de tip „ruptură”, indicate de prima săgeată. În figura de mai jos (Fig.13), pentru vocea sintetizată se observă că formanţii superiori (F3, F4) sunt slab demarcaţi, aproape inexistenţi. Benzile de frecvenţă sunt înguste, foarte bine demarcate.

Se observă în Fig. 12 comparativ cu Fig. 13 că traseul F0 este, în cazul vocii sintetice, mai monoton, iar structura de formanţi este mult mai grosieră şi mai sărăcăcioasă. Traseul F0 în cazul vocii umane prezintă variaţii mai mari (probabil datorită accentuării corecte pe cuvinte) comparativ cu traseul F0 în cazul vocii sintetizate, ultimul fiind mai plat, mai uniform (ceea ce poate indica absenţa unui bloc de stabilire a modului corect în care sunt plasate accentele pe cuvinte în cadrul propoziţiilor).

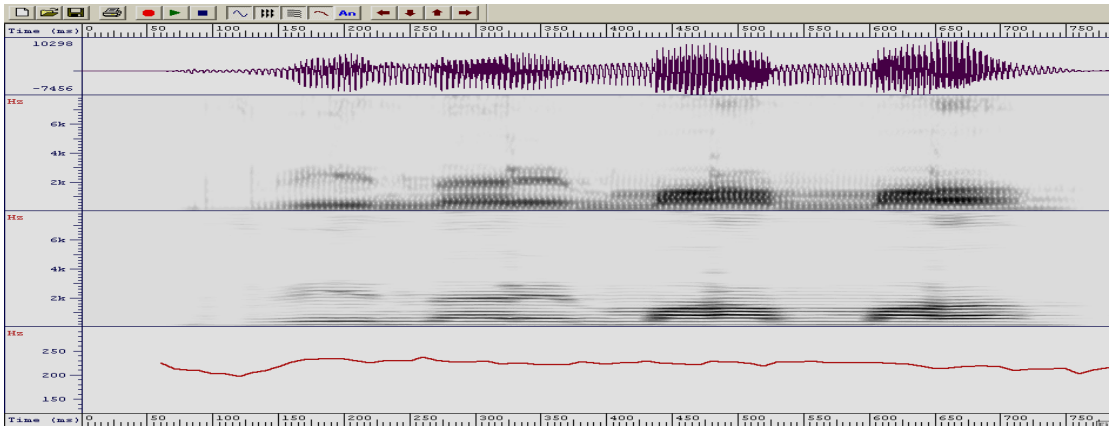

Figura 13 Traseul frecvenţei fundamentale (linia roşie) folosind utilitarul Wasp^{TM}; voce sintetizată - propoziţie “Vine mama”

Spectru vocii umane din Fig. 12 se observă că are un domeniu de variaţie al frecvenţei mult mai larg şi mai difuz; este un spectru bogat în informaţie comparativ cu spectru

vocii sintetizate, care este într-un domeniu de variaţie al frecvenţei mai scăzut, şi sărac în informaţie.

Suplimentar, se constată salturi bruşte (“rupturi”) şi “lărgiri” bruşte ale valorilor respectiv benzilor formanţilor (vezi locurile indicate de săgeţi), salturi care fac vocea neplăcută.

Rezultatele menţionate se confirmă şi pentru alte voci umane. Precizări suplimentare vor fi date într-o lucrare viitoare.

4. Concluzii şi direcţii viitoare

Analiza sumară realizată indică cel puţin o serie de diferenţe majore între formanţii superiori la vocea sintetică faţă de cea umană. Deşi – se ştie – inteligibilitatea este dată doar de formanţii inferiori (F1, F2), calitatea pronunţiei, nuanţele şi bogăţia vocii sunt datorate în mare măsură de formanţii superiori, care sunt sensibil mai puţin “corect” produşi de sintetizor.

Sub nici o formă nu sugerăm că metoda concatenativă nu poate da rezultate excelente în simularea vocii naturale – scopul nostru a fost să arătăm o metodă de verificare cantitativă a “naturaleţii” şi să sugerăm un “benchmark” în verificări. Creşterea naturaleţei necesită, foarte probabil, un număr mai mare de fragmente (elemente) în memorie, astfel încât selecţia să se facă şi ţinând cont de traseele pentru formanţii superiori.

În această lucrare nu ne referim la sintetizatoarele de cercetare, ci numai la cele din domeniul comercial; ca urmare ne-am restrâns doar la singurul produs comercial autohton, realizat pentru limba română. În acest context, amintim că numeroase colective de cercetare s-au preocupat de vocea sintetică şi de îmbunătăţirea calităţii sintezei concatenative în limba română precum, colectivele de la Universitatea Tehnică din Bucureşti, prof. dr. D. Burileanu şi colaboratorii săi, de la Universitatea Tehnică Cluj-Napoca, prof. dr. G. Toderean, de la Academia Tehnică Militară, etc.

Această lucrare reprezintă o raportare preliminară a unor analize începute recent. Ne propunem în viitor să facem un număr mai mare de comparaţii pentru limba română – folosind diverse setări pentru F0 pentru sintetizorul respectiv şi folosind mai multe voci naturale corespunzător alese. De asemenea, este necesar să realizăm analize la nivel dinamic, pentru trasee ale formanţilor (prozodie completă).

Mulţumiri. Primul autor mulţumeşte colegului V. Apopei pentru menţionarea sitului http://www.baum.ro/index.php. Mulţumim de asemenea recenzorilor anonimi pentru observaţiile pertinente făcute.

Analiza prezentată rezumativ în acest scurt raport a fost sprijinită de către Academia Română, Secţia de Ştiinţă şi Tehnologia Informaţiei, în cadrul temei interne „Procese de cogniţie, limbaj şi calcul”, subtema 2.3 “SRoL: optimizare modele statistice disponibile on-line (folosind date din urma analizei formanţilor). Realizarea, în cadrul sitului, a facilităţii de recunoaştere (specificare) on-line a stărilor emoţionale”.

Referinţe bibliografice

H.N. Teodorescu, M. Feraru, D. Trandabăţ, M. Zbancioc, R. Luca, A. Verbuţă, M. Hnatiuc, R. Ganea, O. Voroneanu, L. Pistol, "*Proiectul Sunetele Limbii Române*", www.etc.tuiasi.ro/sibm/romanian_spoken_language/index.htm, http://www.etc.tuiasi.ro/sibm/romanian_spoken_language/ro/fisa_2.htm

BAUM Engineering, *TTS Online, Voce sintetică românească profesională* Ancutza v3.6.16., http://www.baum.ro/index.php?language=ro&pagina=despre_noi, http://www.baum.ro/index.php?language=ro&pagina=ttsonline

P. Boersma, D. Weenink, Institute of Phonetic Science, University of Amsterdam, *Praat: doing phonetics by computer*, www.praat.org

WasP – the Wind Atlas Analysis and Application Program, www.wasp.dk

H.N. Teodorescu, M. Feraru, (2007) A study on Speech with Manifest Emotions, *10th International Conference on Text, Speech and Dialogue*, TSD 2007, Pilsen, Czech Republic, Lecture Notes in Computer Science, Springer Verlag, vol. 4629/2007, ISBN 978-3-540-74627-0, p. 254-262

H.N.Teodorescu, M. Feraru, D. Trandabăţ, (2007) Studies on the Prosody of the Romanian Language: The Emotional Prosody and the Prosody of Double-Subject Sentences. In Corneliu Burileanu and H.N. Teodorescu (Eds.), *Advances in Spoken Language Technology*, The Publishing House of the Romanian Academy, ISBN 978-973-27-1516-1, p.171-182

M. Feraru, H.N. Teodorescu, (2008) *Speech Corpus for the Romanian Language: the Emotional Speech Section*, Inventica 2008, Ed. Performantica, Iaşi, România ISBN 978-973-730-491-9, p. 261-273

BAZA DE DATE ÎN LIMBA ROMÂNĂ PENTRU RECUNOAŞTEREA VORBIRII SPONTANE

DIANA HANES[2], CRISTINA PETREA[2], ANDI BUZO[2], VLADIMIR POPESCU[1, 2], CORNELIU BURILEANU[2]

[1] *Laboratoire d'Informatique de Grenoble, Grenoble INP - France*

[2]*Universitatea Politehnică, Facultatea de Electronică, Telecomunicaţii şi Tehnologia Informaţiei, Bucureşti – România;*

cburileanu@mesnet.pub.ro vladimir.popescu@imag.fr

Rezumat

Recunoaşterea vorbirii spontane reprezintă un domeniu mai puţin cercetat, comparativ cu recunoaşterea vorbirii continue în general. În această lucrare sunt prezentate o serie de rezultate privind proiectarea, achiziţia şi adnotarea unei baze de date de vorbire spontană în limba română. Este propusă o metodologie pentru achiziţia datelor, evidenţiind etapele importante: definirea unui lexic şi a unui dicţionar de unităţi fonetice, achiziţia propriu-zisă a semnalului vocal, precum şi adnotarea materialului vocal în unităţi lingvistice.

1. Introducere

Interacţiunea personalizată între subiectul uman şi calculator constituie o provocare de primă importanţă la ora actuală, în contextul în care serviciile şi aplicaţiile informatice devin din ce în ce mai mult centrate pe utilizator.

În limbile de mare circulaţie (engleză, franceză) există sisteme complete de dialog persoană-calculator. În alte limbi (limba română) realizarea sistemelor de dialog reprezintă un demers de lungă durată.

Recunoaşterea vorbirii poate fi privită drept un proces de recunoaştere a formelor, iar acest lucru se poate realiza fie pe bază de reguli, fie prin metode statistice (Russell & Norvig, 2003).

Un obstacol major în calea realizării unei recunoaşteri fiabile constă în variabilitatea semnalului vocal care provine din: variabilitate lingvistică, variabilitatea vorbitorilor, variabilitatea canalului. Un sistem de recunoaştere a vorbirii spontane trebuie să ia în considerare: independenţa de vorbitor; dimensiunea vocabularului; caracterul continuu al vorbirii; spontaneitatea vorbirii (enunţurile rostite de către utilizator sunt de regulă spontane, neplanificate, caracterizate de disfluenţe, ezitări, interjecţii).

Recunoaşterea vorbirii presupune găsirea unei secvenţe de cuvinte, folosind un ansamblu de modele determinate, achiziţionate într-o fază anterioară de antrenare, şi potrivirea acestor modele cu semnalul de vorbire incident. Sisteme bazate pe abordări statistice sunt disponibile atât în comunităţile academice (sistemul SPHINX de la Universitatea Carnegie Mellon, ansamblul de utilitare HTK – „Hidden Markov Modelling Toolkit" - Universitatea Cambridge, sistemul RAPHAEL - Laboratoire d'Informatique de Grenoble) cât şi în domeniul comercial (sistemele produse de Nuance, Dragon, Microsoft în Statele Unite – (Huang & Acero, 2001)).

2. Arhitectura sistemului de recunoaştere în limba română

Un sistem tipic de recunoaştere a vorbirii funcţionează în două moduri de lucru: antrenare (prin crearea cunoştinţelor necesare funcţionării sale, a modelelor acustice şi lingvistice utilizate) şi recunoaştere (presupune utilizarea resurselor create la antrenare pentru conversia enunţului provenit de la subiectul uman într-o secvenţă de cuvinte).

Figurile 1 şi 2 prezintă propunerea pentru arhitectura fazelor de antrenare şi de testare care stau la baza sistemului de recunoaştere de vorbire spontană. Întrucât scopul acestei lucrări este reprezentat de descrierea procedurii de construire a bazei de date, nu se va insista aici asupra arhitecturii concepute pentru recunoaşterea vorbirii.

La finalul prelucrărilor, rezultatul recunoaşterii este reprezentat de un număr de şiruri de cuvinte alternative pentru un enunţ rostit. Alegerea alternativei celei mai pertinente în raport cu contextul cade în sarcina altor componente ale sistemului de dialog. Sistemul de recunoaştere propus în această lucrare se bazează pe modele Markov ascunse (MMA-uri) antrenate pentru fiecare trifonem.

Sistemul de recunoaştere a cărui dezvoltare se doreşte este bazat pe pachetul de utilitare HTK versiunea 3.0. Mare parte din funcţionalitatea HTK este încapsulată într-un set de biblioteci statice, realizate în limbajul C, care asigură faptul că fiecare aplicaţie se intefaţează cu celelalte într-o manieră puternic controlată şi reproductibilă. De regulă, fiecare aplicaţie HTK implementează un nivel de analiză a vorbirii, însă uneori este necesar ca mai multe aplicaţii să contribuie la realizarea unui nivel de analiză.

Utilitarele HTK, versiunea 3.0, considerate relevante pentru realizarea sistemului de recunoaştere sunt: HLEd, HInit, HCompV, HERest, HCopy, HVite, HResults (Evermann, 2005).

2.1. Etapa de antrenare

În faza de antrenare s-au folosit ca intrare fişiere audio înregistrate. Pentru a construi dicţionarul fonetic s-au folosit convenţiile SAMPA (Munteanu, 2006), (Burileanu, 2002) în scopul realizării transcrierii fonetice. Antrenarea s-a realizat la nivel de trifonem.

Un pas important în faza de antrenare a constat în definirea prototipurilor pentru modelele Markov. Fiecărui trifonem îi corespunde un model Markov cu 6 stări. Ca observaţii pentru modelele Markov s-au folosit două mixturi gaussiene (cu parametrii mediile si varianţele). În prima etapă a fost ales un prototip pentru toate trifonemele.

Modelele Markov ascunse au fost iniţializate folosind o matrice implicită de observaţii şi tranziţii. Parametrii de iniţializare au fost calculaţi folosind toate fişierele de semnal. Aceşti paşi au fost realizaţi prin intermediul utilitarelor HInit şi HCompV.

Etapa de antrenare s-a făcut folosind metoda „embedded" (înglobată) (Evermann, 2005). Această metodă constă în faptul că etichetarea nu se face la nivel de trifonem ci la nivel de fişier. Utilitarul HERest construieşte un MMA care conţine toate MMA-urile din interiorul unui fişier, în ordinea specificată în fişierul de etichetare. Antrenarea MMA-urilor pentru fiecare trifonem s-a realizat folosind parametrii MFCC şi trifonemele etichetate. S-a realizat o aliniere iterativă Viterbi pentru modelele Markov

în scopul obţinerii probabilităţii maxime pentru ca un anumit MMA să reprezinte trifonemul corespunzător.

Parametrizarea semnalului vocal

În cadrul fazei de antrenare o primă subetapă parcursă a fost aceea de parametrizare a semnalului vocal. S-a utilizat parametrizarea cu 12 coeficienţi MFCC, energia şi derivatele corespunzătoare. În total au fost utilizaţi 26 parametri. Lungimea perioadei a fost considerată 10ms. Transformata FFT a folosit fereastră Hamming de 20ms iar semnalului i s-a aplicat un filtru de preaccentuare de ordinul întâi cu coeficientul 0.97. Bancul de filtre are 26 canale şi 12 coeficienţi MFCC la ieşire.

Generarea modelelor Markov prototip

S-au considerat modele Markov ascunse prototip caracterizate printr-o topologie Bakis cu şase stări, cu tranziţii de la stânga la dreapta, dintre care starea iniţială şi starea finală sunt neemisive. Într-o prima fază s-a optat pentru modelarea probabilităţilor de ieşire cu două mixturi gaussiene per stare emisivă.

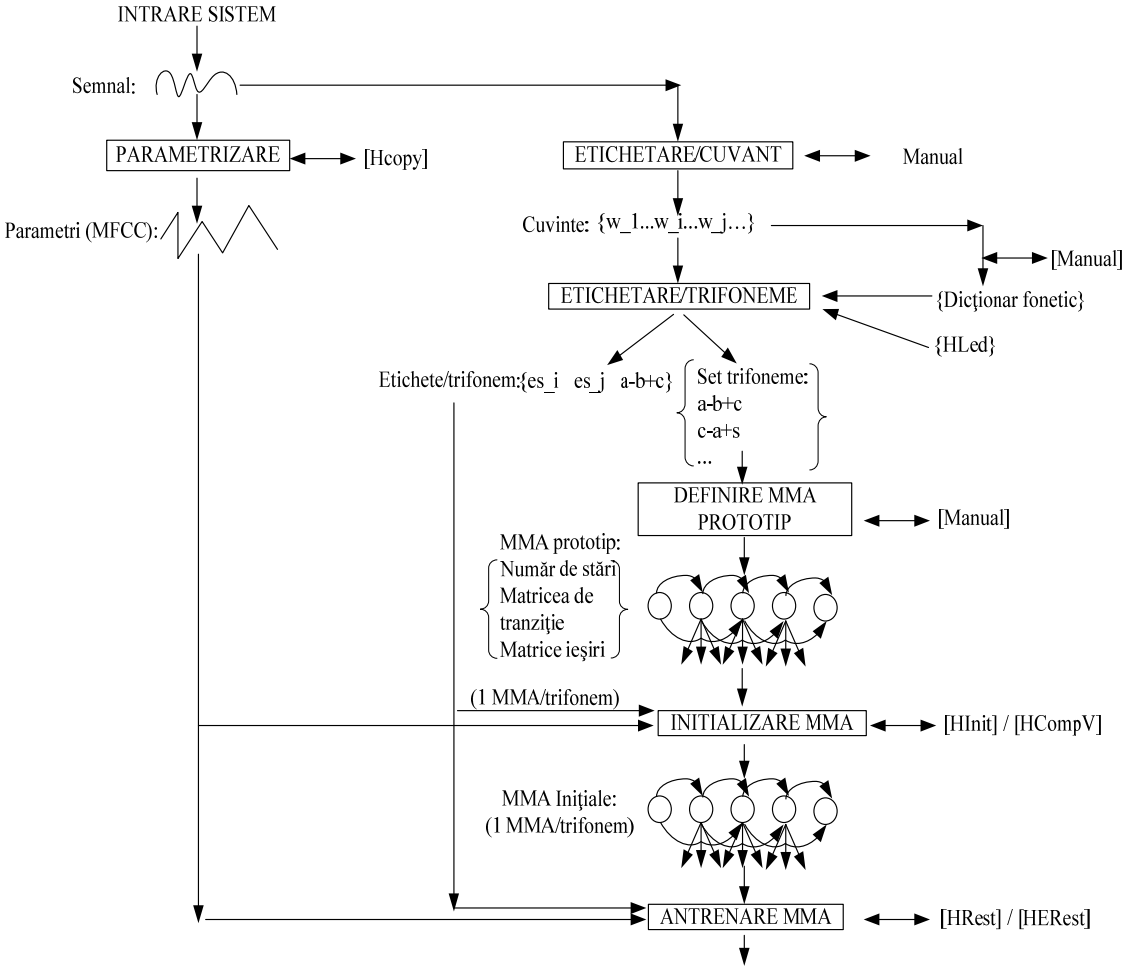

Figura 1: Arhitectura sistemului pentru antrenarea modelelor Markov ascunse

Iniţializarea globală a modelelor Markov ascunse prototip

La iniţializarea globală a modelelor Markov ascunse, mediile mixturilor devin egale cu media globală a vectorilor acustici şi varianţele mixturilor devin egale cu varianţa globală a vectorilor acustici.

Pentru iniţializarea globală s-a folosit utilitarul HCompV (Evermann, 2005). S-au folosit ca intrări: setul de vectori cepstrali filtraţi Mel obţinut cu HCopy; fişierele cu etichetele şi setul de MMA-uri prototip. Rezultatul rulării îl reprezintă un set de MMA-uri iniţializate grupate într-un singur fişier „mmf".

S-a constatat că rularea utilitarului HCompV a durat pentru un singur prototip MMA mai mult de 10s. Numărul de MMA-uri prototip este egal cu numărul de trifoneme obţinut pentru baza de date creată şi anume 5095 trifoneme. Timpul de rulare a iniţializării globale pentru toate trifonemele a fost de aproximativ 50950s, adică 850 minute, ceea ce înseamnă estimativ 14 ore.

Antrenarea propriu-zisă

Pentru antrenarea Baum-Welch „embedded" s-a folosit utilitarul HERest. S-au utilizat ca intrări: fişierele „mfc" care conţin parametrii MFCC, fişierele cu etichete la nivel de trifonem grupate într-un singur fişier „mlf", fişierele MMA rezultate în urma rulării lui HCompV grupate într-un singur fişier „mmf" şi o listă cu toate trifonemele. La ieşire a rezultat un fişier cu extensia „mmf" care conţine toate MMA-urile antrenate.

2.2. Etapa de testare

În faza de testare se folosesc rezultatele obţinute în cadrul etapei de antrenare (modele acustice antrenate), resurse folosite la antrenare (dicţionarul fonetic) precum şi alte resurse (gramatica). Faza de testare împreună cu evaluarea rezultatelor reprezintă faza terminală a procesului de recunoaştere.

Fişierele audio folosite la intrarea sistemului au fost parametrizate şi au fost extraşi parametrii MFCC folosind utilitarul HCopy (Evermann, 2005). Parametrii MFCC constituie intrare pentru decodarea trifonemelor.

Decodarea parametrilor acustici se realizează folosind MMA-urile antrenate la nivel de trifonem. Secvenţa de trifoneme obţinută va fi transformată într-o secvenţa de cuvinte folosind gramatica cu un număr finit de stări.

În faza de testare se va parcurge procesul de „definire gramatici". Astfel se va defini o gramatică-buclă în care se plasează toate cuvintele, care ar putea urma unul după altul cu egale şanse de apariţie. Se foloseşte utilitarul HParse (Evermann, 2005). Decodarea utilizează HVite, HParse şi apoi HResults pentru evaluare.

Utilitarul HVite are ca intrări un set de .mfc-uri care constituie semnalul de test, setul de MMA-uri, gramatica constituită din cuvintele din dicţionarul fonetic, dicţionarul fonetic şi lista de modele Markov ascunse (ca în cazul HERest).

Paşii pentru rularea testării sunt: pregătirea datelor (plasarea într-un acelaşi director a dicţionarului fonetic, a fişierului cu MMA-urile antrenate) şi apoi construirea manuală a gramaticii "grammar-orig" folosind lista de cuvinte din dicţionar. Între fiecare două

cuvinte se introduce separatorul "|". Urmează rularea propriu-zisă şi evaluarea rezultatelor folosind HResults.

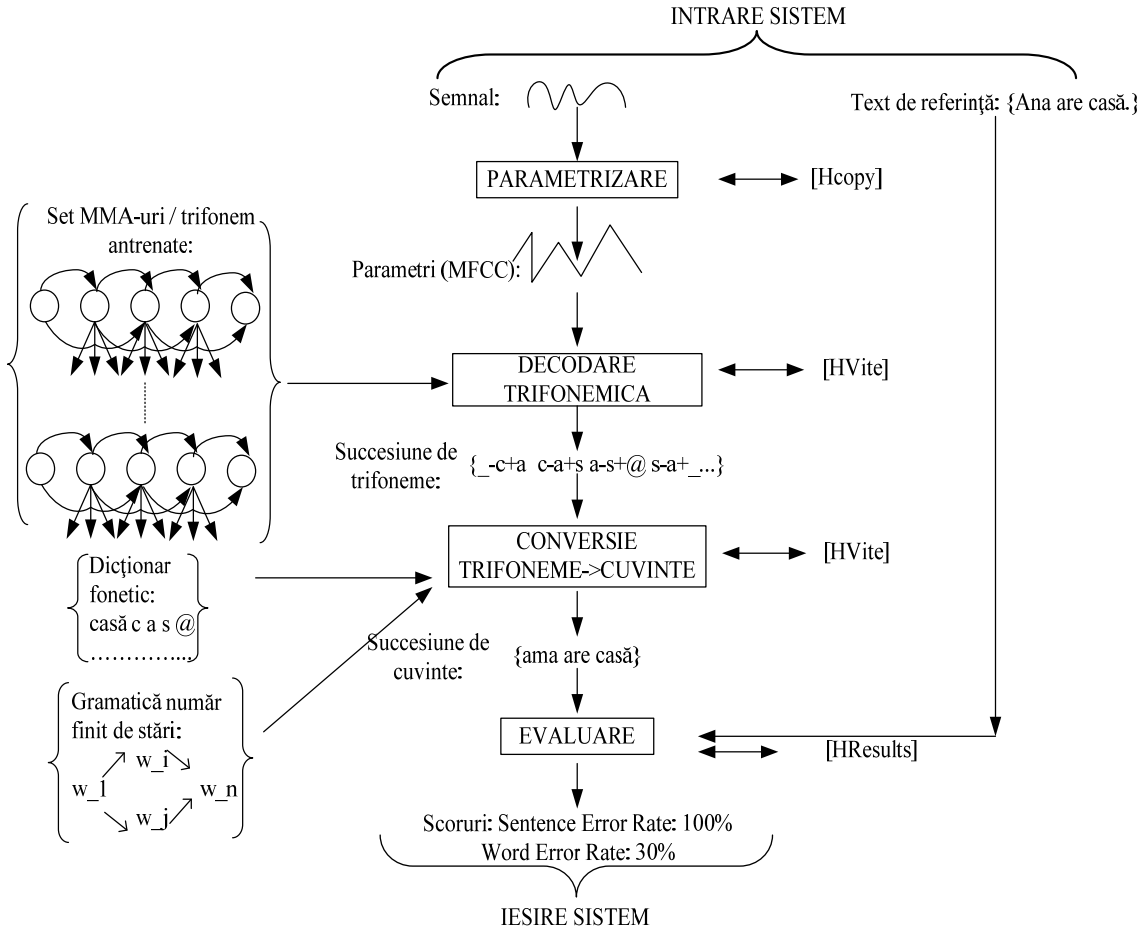

Figura 2: Arhitectura sistemului pentru decodarea vorbirii utilizând modele Markov ascunse

Semnalul de referinţă este comparat cu rezultatele obţinute pentru a determina performanţele utilitarului de recunoaştere construit. Analiza performanţelor acestui utilitar se va face folosind: Sentence Error Rate şi Word Error Rate.

3. Construirea bazei de date în limba română pentru recunoaşterea vorbirii spontane

3.1. Caracteristici

Vocabularul folosit în recunoaşterea de vorbire spontană trebuie să fie pe cât posibil de cuprinzător. De la început s-a urmărit construirea vocabularului într-o manieră scalabilă făcând posibilă adăugarea de cuvinte noi ulterior acestei faze. În faza de creare a bazei de date au fost realizate convenţii care să fie refolosite în etapele ulterioare de lărgire a bazei de date.

3.2. Probleme specifice realizării bazelor de date utilizate în recunoaşterea de vorbire spontană

Construirea bazelor de date pentru recunoaşterea vorbirii spontane se caracterizează prin elemente specifice. Recunoaşterea de vorbire poate fi considerată un proces de recunoaştere a formei iar acest lucru se obţine pe baza unor reguli sau metode statistice (Russell & Norvig, 2003). Cea din urmă variantă este cea preferată în acest moment datorită rezultatelor bune obţinute cu costuri de producţie acceptabile. Metodele statistice presupun folosirea unor date de intrare în procesul de antrenare astfel încât sistemul generează informaţie pe care o foloseşte în etapele ulterioare.

Următoarele caracteristici sunt relevante pentru recunoaşterea vorbirii: tipul sistemului (dependent de vorbitor sau independent de vorbitor), dimensiunea vocabularului (Peinado & Segura, 2006) (vocabular mic: 10-100 cuvinte, vocabular mediu: 100-1000 cuvinte, vocabular mare: 10.000-100.000 cuvinte).

Opţiunile disponibile pentru realizarea bazei de date sunt: înregistrările directe, fişiere audio din cadrul programelor de televiziune sau radio care au fost difuzate pe Internet, utilizarea de fişiere audio înregistrate direct de la radio sau de la televizor.

Parametrii necesari în construirea unei baze de semnal vocal pentru recunoaşterea de vorbire spontană sunt: fişiere audio cu semnal vocal; extragerea caracteristicilor semnalului vocal din cadrul înregistrărilor audio; etichetarea fişierelor audio; parametrii acustici ai fişierelor audio (coeficienţii de predicţie liniară (LPC), coeficienţii cepstrali).

Tabelul 1 indică principalele caracteristici pentru baza de date în limba română.

Tabel 1: Caracteristicile bazei de date realizate.

Proprietate	**Valoare**
Procedura colectare fişiere audio	Înregistrări preluate de pe Internet ale unor emisiuni româneşti
Limbaj folosit	Limba română, vorbire orală/spontană
Durată înregistrări	~ 4 ore
Vorbitori	12 (4 bărbaţi, 8 femei)
Sesiuni per vorbitor	3-20
Număr total de cuvinte	37604
Număr de cuvinte unicat	8068
Frecvenţa de eşantionare semnal vocal	8kHz

Pe baza elementelor menţionate în tabel au fost evidenţiate următoarele aspecte: segmentarea semnalului vocal – este de preferat ca lungimea fişierelor vocale să fie de 60s; etichetarea semnalului – etichetarea poate fi realizată la nivel de cuvânt (proces realizat manual şi consumator de timp) sau la nivel de fonem/trifonem (proces semi-automat care se bazează pe etichetarea iniţială manuală, este lipsit de stabilitate); parametrizarea semnalului vocal – pentru aceasta anumite criterii trebuie îndeplinite: maximizarea dispersiei inter-fonem şi minimizarea dispersiei intra-fonem.

3.3. Culegerea şi structurarea datelor

Pentru a construi baza de date de vorbire spontană în limba română, au fost utilizate înregistrări ce conţin ştiri, poveşti, show-uri de televiziune, discuţii medicale, financiare, previziuni meteo şi alte tipuri de informaţii toate fiind transmise pe Internet sau la radio.

Baza de înregistrări audio conţine fişiere audio care provin de la 12 vorbitori cu pregătiri în diverse domenii; sunt persoane care au diferite stiluri de viaţă, experienţă, obiceiuri. Vorbirea este fluentă şi înregistrările audio conţin diferite tipuri de informaţii. Pentru fiecare vorbitor există în baza de înregistrări audio între 5 şi 38 de fişiere audio.

3.4. Construirea bazei de date

Transcrierea din fişiere audio în fişiere text

Fişierele audio au fost prelucrate înainte de a fi utilizate efectiv la crearea bazei de date. Înregistrările audio au fost divizate pe vorbitori; au fost eliminate zonele în care vocile s-au suprapus; fiecare fişier audio a fost divizat în fişiere audio de durată 60s.

S-a ascultat fiecare fişier audio, s-a extras mulţimea de cuvinte rostite ţinând cont de diacriticele specifice limbii române: ă -> @; â -> i_; î -> i_; ş -> S; ţ -> ts. Cratimele au fost suprimate, cuvintele sub formă de acronim au fost fonetizate astfel: "bcr" devine "becere"; "bvb" devine "bevebe" etc. În cazul onomatopeelor prin convenţie s-a dublat vocala mai lungă astfel "breee...." a devenit "bree".

Întrucât se doreşte construirea unei baze de cuvinte pentru vorbire spontană, au fost incluse în fişierele .txt cuvintele care sunt pronunţate incomplet, incorect şi eventualele bâlbe. Acestea sunt entităţi recurente ce trebuie considerate ca atare.

Din mulţimea de cuvinte rezultată au fost generate două fişiere: un fişier care include totalitatea cuvintelor din înregistrările audio prelucrate (37604 apariţii cuvinte) şi respectiv al doilea fişier care conţine doar o reprezentare a cuvintelor din lista generată iniţial (8068 cuvinte).

S-a constatat că în cadrul vocabularului construit 12147 de cuvinte au peste 100 de apariţii. În figura 3 sunt reprezentate cuvintele care au mai mult de 250 de apariţii în cadrul vocabularului realizat până în acest moment.

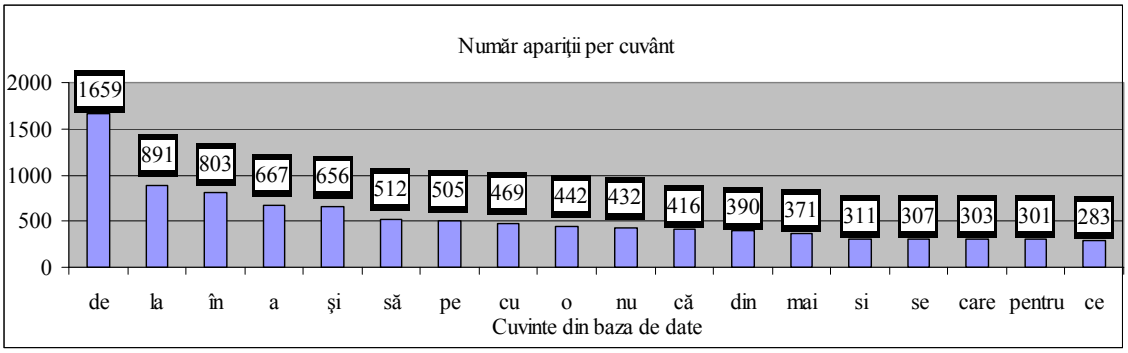

Figura 3: Cuvintele cu cele mai multe apariţii în baza de date.

Transcriere fonetică

Etapa următoare în crearea bazei de date a constat în transcrierea fonetică a fiecărui cuvânt. Transcrierea fonetică s-a realizat pe baza regulilor SAMPA (Burileanu, 2002). Astfel a rezultat dicţionarul fonetic.

Transcriere trifonetică

Conversia transcrierilor fonetice în transcrieri trifonetice s-a realizat cu utilitarul HLEd. Pentru fiecare trifonem s-a construit un model Markov în formatul HTK.

Etichetare la nivel de fişier audio

Etichetarea s-a realizat la nivel de fişier audio folosind utilitarul Wavesurfer disponibil gratuit la: http://www.speech.kth.se/wavesurfer/download.html. În cadrul etichetării la nivel de fişier audio, pauzele scurte între cuvinte au fost marcate explicit prin „sp” şi pauzele mai lungi au fost marcate prin „sil”. S-a obţinut o etichetare mai puţin fiabilă decât cea la nivel de cuvânt însă cu preţul îmbunătăţirii timpului de etichetare.

4. Statistici pentru limba română

Baza de date descrisă în secţiunile anterioare este destinată utilizării în cadrul unei aplicaţii de recunoaştere a vorbirii spontane. Rezultatele sunt prezentate prin intermediul statisticilor care fac referire la numărul de apariţii al cuvintelor şi al trifonemelor. În lucrarea de faţă în locul fonemelor au fost utilizate trifoneme. Având în vedere faptul că între cuvinte a fost folosit ca separator „sp”, care nu reprezintă un fonem, atunci se poate considera că începutul cuvintelor şi sfârşitul acestora sunt realizate ca difoneme.

Exemplu: cuvântul „c a s @” are următoarea reprezentare fonetică: „c-a”, „c-a+s”, „a-s+@”, „s+@”. Motivaţia pentru care se folosesc trifonemele este că acestea permit analizarea contextului, întrucât sunt entităţi care păstrează informaţia despre ce găsim înaintea şi după un anumit fonem.

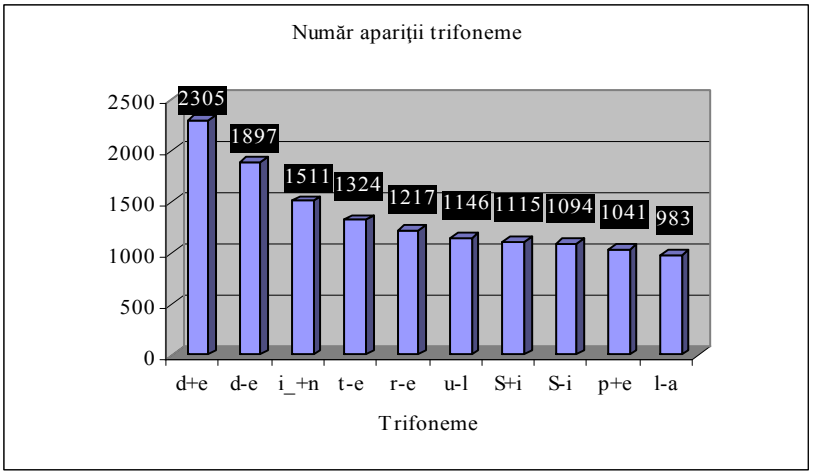

Figura 4: Trifoneme cu cele mai multe apariţii.

În continuare, se vor prezenta statistici referitoare la trifonemele din cadrul bazei de date construite în limba română. Astfel, în figura 4 sunt ilustrate trifonemele care au cele mai multe apariţii în cadrul bazei de date având între 983 şi 2305 de apariţii. Predominante sunt trifonemele „d+e” şi „d-e”.

Figura 5 ilustrează următoarele 10 trifoneme din cadrul bazei de date care au între 982 de apariţii şi 730 apariţii în baza de date. Predominante sunt trifonemele „l+a” şi „l-e”.

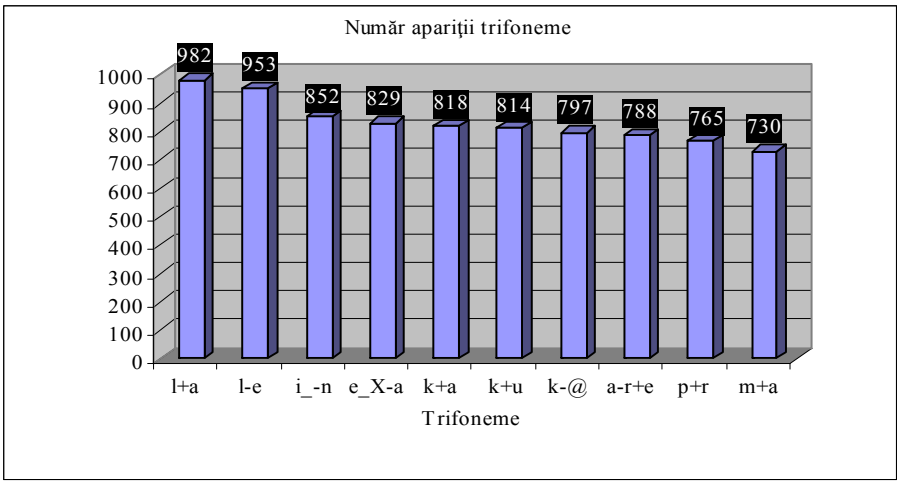

Figura 5: Trifoneme cu apariţii multiple.

Din cele 5095 de trifoneme care caracterizează vocabularul construit se pot extrage următoarele caracteristici: 1% din numărul total de trifoneme au între 500 şi 1000 de apariţii; 7% din trifoneme au între 100 şi 500 de apariţii; 8% din trifoneme au între 50 şi 100 de apariţii; alte 8% din trifoneme au între 30 şi 40 de apariţii; 19% din trifoneme au între 10 şi 30 de apariţii; 16% din trifoneme au între 5 şi 10 apariţii iar 41% din trifoneme au între 1 şi 5 apariţii.

5. Concluzii

Baza de date construită pentru limba română în scopul recunoaşterii vorbirii spontane are o dimensiune relativ medie. Pentru munca de cercetare viitoare şi pentru a putea realiza o bună analiză a caracteristicilor vorbirii spontane, obiectivul principal în viitorul apropiat îl constituie creşterea numărului total de cuvinte şi creşterea numărului de apariţii al cuvintelor.

Caracteristicile principale ale bazei de date în limba română, realizate în scopul utilizării în cadrul proiectului sunt următoarele: durata înregistrărilor este de 4 ore, înregistrările sunt preluate din diverse medii de lucru, vorbitorii folosesc vorbirea spontană. Au fost folosiţi 12 vorbitori diferiţi (8 voci feminine, 4 voci masculine), fiecare vorbitor având mai multe sesiuni de înregistrare, în medie 20 de sesiuni per vorbitor. Baza de cuvinte numără în total 37.604 cuvinte, 8068 de cuvinte au apariţie singulară iar numărul total de trifoneme este 5095.

Într-o primă fază, testele de recunoaştere efectuate au condus la obţinerea unor rezultate cu o performanţă scăzută. Cauzele principale au fost identificate şi localizate după cum urmează: numărul foarte mic de apariţii pentru fiecare trifonem; calitatea redusă a

anumitor fişiere audio; erori ale intervenţiei umane în etapa de prelucrare a fişierelor audio.

Din analiza cauzelor menţionate anterior a rezultat necesitatea parcurgerii următoarelor etape:

- prelucrarea manuală a materialului audio pentru a corecta eventualele erori comise la pregătirea datelor;
- verificarea fişierelor cu etichete;
- antrenarea la nivel de fonem a fonemelor care au mai multe apariţii şi apoi refolosirea anumitor parametri calculaţi la antrenarea fonemelor şi a trifonemelor;
- segmentarea fişierelor etichetate la nivel de fonem în fişiere cu durata sub un minut;
- introducerea unor constrângeri gramaticale care să amelioreze semnificativ calitatea recunoaşterii.

Mulţumiri. Cercetările prezentate în această lucrare au fost finanţate de Guvernul României, prin grantul de cercetare IDEI, nr. 930/2007.

Referinţe bibliografice

Peinado, A., Segura, J. (2006). *Speech Recognition Over Digital Channels: Robustness and Standards*, Chapter 2, Pages 7-30

Russell, S., Norvig, P. (2003). *Artificial Intelligence: A Modern Approach*, Prentice Hall (Second Edition)

Burileanu, D. (September, 2002). Basic Research and Implementation Decisions for a Text-to-Speech Synthesis System in Romanian, *International Journal of Speech Technology*, Volume 5, Number 3, Pages 211-225

Evermann, G., et al. (2005). *The HTK Book*, Version 3.0, Cambridge University Engineering Department;

Munteanu, D. (2006). *Contribuţii la realizarea sistemelor de recunoaştere a vorbirii continue pentru limba română*, Teză de doctorat, Academia Tehnică Militară din Bucureşti

Huang, X., Acero, A., Hon, H.-W. (2001). *Spoken Language Processing - A Guide to Theory, Algorithm and System Development*, Prentice Hall

METODĂ IERARHICĂ DE DETECŢIE A FUNDAMENTALEI

MARIUS-DAN ZBANCIOC[1,2], HORIA-NICOLAI TEODORESCU[1,2]

[1] *Institutul de Informatică Teoretică al Academiei Române - Filiala Iaşi*
[2] *Universitatea Tehnică "Gheorghe Asachi", Iaşi – România*

{zmarius ,hteodor}@etc.tuiasi.ro

1. Introducere

Problematica analizei prozodice nu este încă rezolvată, datorită naturii nestaţionare a semnalului vocal şi inexistenţei unui suport matematic pentru definirea conceptului de frecvenţă fundamentală, F_0, respectiv a conceptului de formant. Colective din institutul nostru de cercetare au implementat mai multe instrumente automate de extragere a lui F_0, pe baza unor metode clasice din literatură (Rowden, 1991; Rabiner & Juang 1993; Rabiner & Schafer 1978; Calliope, 1989; O'Shaughnessy, 1987; Cristea & Valsan, 1999): AMDF, autocorelaţie, HPS şi metoda cepstrală, respectiv a unei metode hibride propuse în (Teodorescu, 2006). Recent s-a realizat şi un instrument de extragere a formanţilor superiori F_1,..,F_4, validarea rezultatelor de ieşire fiind realizată într-un bloc decizional neuro-fuzzy. Sistemul ierarhic hibrid, cu un bloc neuro–fuzzy care înglobează mai multe metode de detecţie a F_0, ponderează diferit fiecare extractor funcţie de performanţele lui şi controlează astfel influenţa fiecărei metode asupra valorilor finale.

Îmbunătăţirile recent aduse de noi instrumentelor de analiză au vizat blocul de pre-procesare, care realizează operaţii de filtrare, respectiv segmentare a zonei de interes (vocalice) de zona de background (consonantică sau de pauze între rostiri). Valorile de prag (threshold) folosite anterior erau determinate empiric, noile valori fiind determinate pe baza regulilor furnizate de un arbore de decizie. Instrumentele de analiză (codurile sursă şi executabilele) sunt disponibile pe site-ul SRoL - „Proiectul Sunetele Limbii Române" (http://iit.iit.tuiasi.ro/romanain_spoken_language/ index.htm). Aceste instrumente de analiză a informaţiei prozodice bazate pe valorile frecvenţei fundamentale şi ale formanţilor sunt utile în aplicaţii de recunoaştere şi sinteză a semnalului vocal (prin modulele de studiu al caracteristicilor fonemelor), în aplicaţii de studiu al intonaţiei şi al altor informaţii paralingvistice, sau în aplicaţii de identificare de limbă şi de particularităţi ale acesteia (dialecte) etc.

2. Descrierea modulelor componente ale aplicaţiei

Pentru implementarea unui instrument pentru detecţie de F_0 şi pentru determinarea valorilor formanţilor, s-au conceput modular mai multe aplicaţii (programe), fiecare dintre acestea fiind apelabilă independent de restul aplicaţiilor (principiul modularităţii). Rezumăm aceste aplicaţii, fiecare fiind asociată uneia dintre etapele de analiză şi principalele lor funcţii:

Modulul de extragere de trăsături din fişiere de sunet prin corelare cu informaţia extrasă din fişierele adnotate „*.TextGrid". Se determină un vector de timp folosit pentru delimitarea fonemelor şi se extrag pattern-uri de trăsături (energie benzi spectrale, valoare medie energie în domeniul timp, deviaţia standard, rata trecerilor

prin zero etc.), care să fie folosite ulterior de o metodă de clasificare automată (arbori de decizie, reţele neuronale, algoritmi genetici etc.).

Modulul de preprocesare

- Filtrarea semnalului (folosind un filtru median sau un filtru de mediere, respectiv un filtru trece bandă [70, 5000 Hz]) pentru a izola mai bine banda de frecvenţe de interes în care se caută fundamentala, respectiv formanţii.

- Segmentarea V/C zonei vocalice de zona consonantică (pe baza energiei din fereastra de analiză în domeniul timp, respectiv a energiei spectrale dintr-o bandă de frecvenţe raportată la valoarea întregii energii spectrale).

Modulul de prelucrare statistică

- Poate fi utilizat pe fişierele de ieşire, pentru determinarea unor valori ce caracterizează un anumit grup de foneme (de exemplu consoanele), sau pentru extragerea unui vector de trăsături specifice unui anumit fonem.

- Este apelat pentru determinarea automată a pragurilor utilizate de modulul de segmentare C/V, atunci când se doreşte o minimizare a erorii de clasificare a zonei de interes, de zona de background (fundal sonor).

Modulul de detecţie de valori formantice F0, F1,…, F4 foloseşte

- metoda autocorelaţiei (analiză în domeniul timp);

- metoda diferenţelor AMDF (analiză în domeniul timp);

- metoda produsului spectrelor armonice HPS (analiză în domeniul frecvenţelor);

- metoda cepstrală (analiză în domeniul que-frecvenţelor) – utilizată şi pentru căutarea formanţilor superiori.

Modulul decizional pentru ponderarea ieşirilor furnizate de fiecare metodă de detecţie de formanţi. Sunt eliminate valorile eronate prin comparare cu un număr de N ieşiri anterioare, respectiv ieşiri ulterioare (fără a mai respecta condiţia de cauzalitate). Printr-un algoritm multicriterial, funcţie de performanţele fiecărei metode se asociază acestora ponderi, determinate astfel încât să se apropie cât mai mult statistic valoarea finală de valorile „reale" ale formanţilor.

3. Modulul de preprocesare

Etapa de preprocesare este una esenţială a instrumentului de analiză a informaţiei prozodice. În această etapă se realizează filtrarea semnalului pentru eliminarea zgomotelor suprapuse peste semnalul util, cum ar fi zgomotul indus de reţea (50 Hz) şi zgomotul echipamentului de înregistrare. Banda pentru căutarea formanţilor se consideră a fi [70-5000Hz]. Extragerea acestei benzi se realizează printr-un filtru digital trece bandă (FTB), iar eliminarea zgomotului uniform prin aplicarea unui filtru de mediere (FTJ). Pentru o filtrare bună a semnalului, fără a afecta mult semnalul original, am preferat alegerea unui ordin mai mic pentru filtrul de mediere şi aplicarea înseriată (repetată) a filtrului.

În această etapă se realizează şi segmentarea semnalului, încercându-se o cât mai bună separare a zonelor vocalice de zonele consonantice şi de zonele care sunt pauze între rostiri. Căutarea valorilor formantice se face doar pe zonele vocalice extrase. Realizarea unui instrument automat de analiză este dificilă din mai multe puncte de vedere. Înregistrările diferă prin energia semnalului (vorbitorul poate vorbi mai încet sau mai tare), prin raportul semnal / zgomot, prin spectrul de frecvenţe specific fiecărui vorbitor (se ştie că , adesea, altfel arată traseele formantice în cazul unui vorbitor masculin şi altfel pentru un vorbitor de sex feminin, în special pentru primii doi formanţi F_0 şi F_1). O segmentare cât mai precisă este esenţială pentru obţinerea de rezultate bune în final.

Separarea zonei vocalice de zona consonantică se realizează cu două metode:

a) Criteriu global: Se compară energia din fereastra curentă de analiză $E_W = \sum_{n=1}^{N} x_n^2$ cu energia maximă $E_{\max}$ calculată după ce se parcurge tot semnalul. Dacă $E_W > procent_1 \bullet E_{\max}$, atunci se consideră că semnalul din fereastra curentă este vocalic şi că se poate determina F_0. S-a folosit $procent_1 = 20\%$.

b) Criteriu local: Dacă energia spectrală în banda [70, 1000Hz] > procent2 din toată energia spectrală, se consideră că avem vocală. Premisa acestei segmentări este că energia spectrală este mare în zona formanţilor. Pragul folosit este de 50% ($procent_2$).

Primul criteriu este *global*, căci după ce se caută în tot semnalul fereastra cu energie maximă se consideră că doar secvenţele de semnal care au energia ferestrelor de analiză mai mare de 20% sunt de interes în analiză. Valorile de prag utilizate, de 0.2 şi 0.5, au fost estimate empiric, după încercări succesive de ajustare a acestora.

Segmentarea folosind doar cele două criterii nu este satisfăcătoare, motiv pentru care s-au extras vectori cu mai multe trăsături (descrişi mai jos) şi au fost introduşi într-un *arbore de decizie* (http://www.rulequest.com/see5-win.html), instrument capabil să furnizeze un set de reguli, care să minimizeze eroarea de clasificare.

Folosind *modulul de căutare fişiere* de adnotare s-au obţinut vectori de timp $[t_0, t_1, \ldots, t_n]$ care delimitează fiecare fonem din secvenţa rostită. Dintre nivelurile de structurare a informaţiei folosite la adnotare (foneme, silabe, cuvinte, propoziţii etc.) s-a utilizat doar segmentarea la nivel de fonem (nivel 1 în figura 1).

După citirea unui fişier adnotat *nume.TextGrid* se identifică fişierul de sunet *nume.wav* corespunzător, verificând dacă dimensiunea acestuia coincide cu limitele de timp [t_0, t_{end}].

$$\frac{dimensiune\_fisier - dimensiune\_header}{nr\_biti\_per\_esantion \cdot Frecv\_esantionare} = t_{end} - t_0$$

Parametrii de intrare în sistemul de clasificare automată sunt determinaţi pentru fiecare fonem din secvenţă rostită analizată şi sunt următorii:

- *nume fonem*;

- *id_class* identificator clasa (1-vocală, 2-consoană, 3-pauză rostiri), *identificator fişier*;

- *zcr* (rata trecerilor prin zero, valoare normalizată pentru durata de o secundă);

- *avg_e* energia medie a ferestrelor de analiză, din secvenţa de semnal ce delimitează fonemul curent;

- *std_e* deviaţia standard a energiei pe durata fonemului, de la valoarea medie *avg_e*;

- *B1* energia în banda [70, 500]Hz exprimată procentual faţă de toată energia spectrală;

- *B2*, *B3*, *B4* energii spectrale în benzile [500, 1000] Hz, [1000, 2000] Hz şi [2000, 5000] Hz

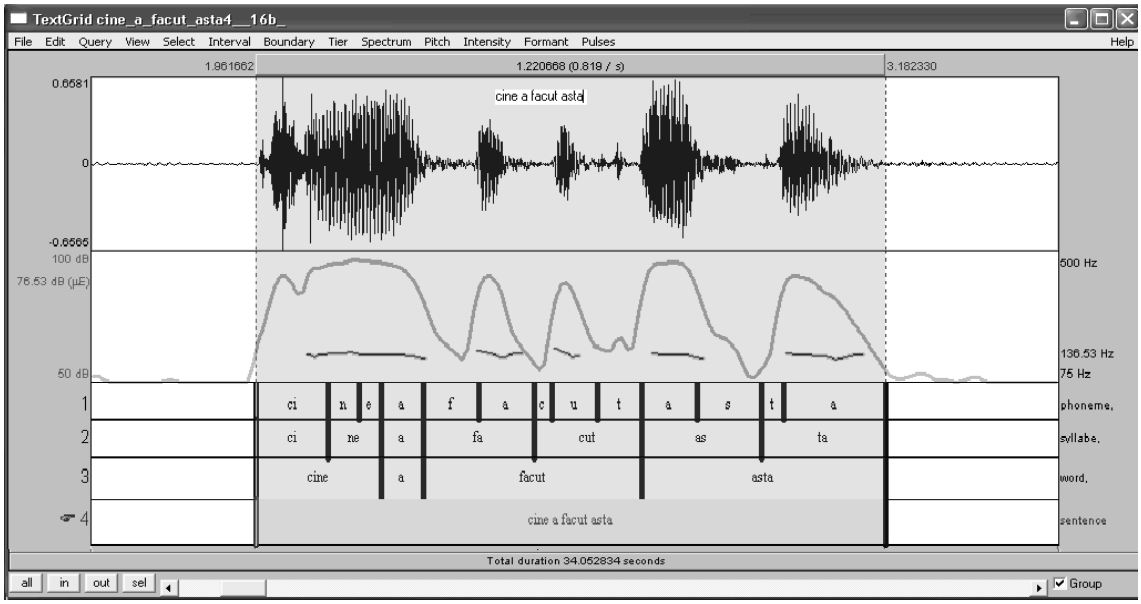

Figura 1. Adnotarea unui fişier de sunet pe mai multe niveluri folosind Praat™

Ferestrele de analiză utilizate au dimensiunea de W=1024 eşantioane, ceea ce corespunde, pentru o frecvenţă de eşantionare *Fs* de 22050 Hz, la o durată de 46,44 ms, respectiv pentru *Fs* de 16000 Hz la 64 ms. Pentru fonemele ale căror durate $[t_i, t_{i+1}]$ au fost mai mici decât W nu s-au extras vectori de trăsături. Într-o rafinare ulterioară a algoritmului de segmentare se vor asocia funcţii de apartenenţă (f.a.) fuzzy trapezoidale fiecărui fonem şi se vor pondera rezultatele ferestrelor de analiză cu gradul de apartenenţă descris de f.a.

Rezultatele prezentate mai jos sunt realizate pe un set de 12 fraze adnotate (”A trecut aşa un răstimp”, ”O şti el careva cum să rezolve asta”, „Mama vine şi ea mai târziu”, „Mama ştie ea ce face”, „Chiar ştie el ce face?”,” Vine ea mama!”).

```
See5 [Release 1.15]

Rule-based classifiers-50% data for training

Rule 1: (270/120, lift 1.4)
     E_MED > 0.000516
     B4 <= 0.068311
     ->  class 1  [0.555] | vocala
```

```
Rule 2: (55/2, lift 1.8)
      E_MED > 0.008076
      E_MED <= 0.095878
      B2 <= 0.053437
      B4 > 0.00501
      ->  class 2  [0.947] | consoana
Rule 9: (120/16, lift 1.6)
      E_STD <= 0.085847
      B2 <= 0.014152
      ->  class 2  [0.861] | consoana
Rule 10: (16, lift 11.1)
      E_MED <= 0.008076
      B2 > 0.014152
      B2 <= 0.07656
      B4 <= 0.068311
      ->  class 3  [0.944] | pauza rostiri
Rule 11: (36/5, lift 9.9)
      E_MED <= 0.000516
      ->  class 3  [0.842] | pauza rostiri
```

Tabel 1. Evaluarea matricei de confuzie pe un set date de antrenare de 400 cazuri: (eroare 7.8%)

clasa 1	**clasa 2**	**clasa 3**	**clasificate ca:**
133	21	0	(a): clasa 1 vocale
7	203	2	(b): clasa 2 consoane
0	1	33	(c): clasa 3 pauză rostiri

Tabel 2. Evaluarea matricei de confuzie pe un set date de test de 400 cazuri: (eroare 23.8%)

clasa 1	**clasa 2**	**clasa 3**	**clasificate ca:**
130	54	0	(a): clasa 1 vocale
31	157	7	(b): clasa 2 consoane
0	3	18	(c): clasa 3 pauză rostiri

Nu s-au luat în calcul diftongii, triftongii şi nici sunetele 'ghe', 'ghi', 'che', 'chi', 'ş', 'ţ'. Din cele 11 reguli furnizate de sistem, una a fost folosită pentru clasificarea vocalelor, 8 pentru clasificarea consoanelor şi 2 pentru pauze. Se justifică separarea consoanelor într-un studiu viitor în mai multe clase, funcţie de particularităţile acestora, de exemplu grupul consoanelor plozive, care se confundă uşor în segmentare cu zona de pauză, precum şi al consoanelor semivocalice ('l', 'm', 'n', 'r'), pentru care are sens să se extragă informaţie formantică şi care se confundă des în segmentare cu vocalele (vezi Tabel 1, 2).

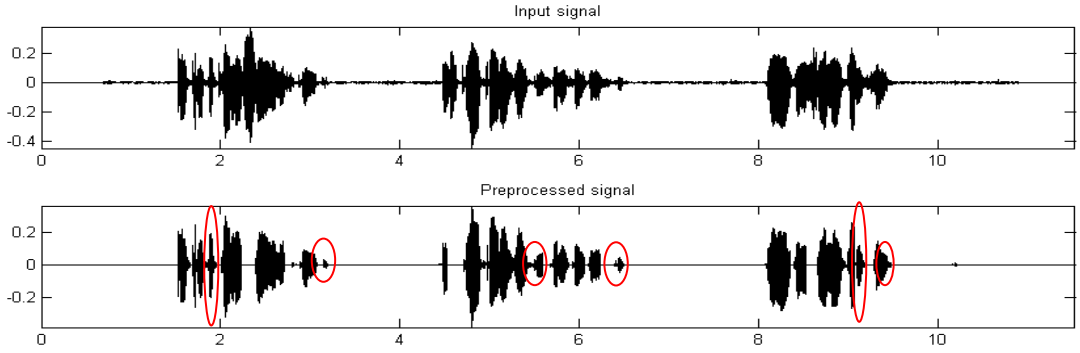

Figura 2. Selectarea zonelor vocalice de interes în urma aplicării noilor algoritmi de segmentare

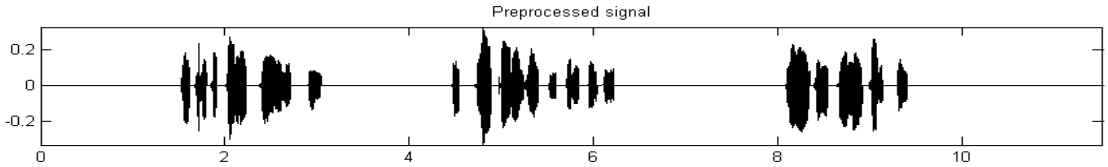

Figura 2. Segmentarea V/C folosind algoritmi de segmentare anteriori

În urma aplicării algoritmilor de segmentare folosind valorile prag determinate automat de regulile generate de arborele de decizie, zona de interes este mai precis selectată, apar zone vocalice, marcate pe figura 2, pe care algoritmii anteriori nu reuşeau să le identifice. S-au realizat extrageri de reguli, folosind instrumentul **See5** pe mai multe seturi de fişiere adnotate, o parte din regulile obţinute şi erorile de clasificare fiind date mai sus. În figura 2 s-a reprezentat segmentarea obţinută aplicând negata regulii R1, pentru eliminarea zonelor nevocalice:

```
IF (E_MED < 0.000516) OR (B4 > 0.068311) => nu este vocala
```

Pentru determinarea zonei de pauză între rostiri s-a folosit o clasificare, utilizând doar rata trecerilor prin zero - *zcr* şi energia medie în domeniul timp - *e_med*. S-au folosit de această dată 276 de fişiere adnotate de tipul „b_ba_aba" (consoană, consoană urmată de vocala 'a', consoană încadrată între două vocale 'a'). Au rezultat cca. 2000 de vectori de trăsături. Regulile generate pentru clasificarea pauzei dintre rostiri au permis identificarea a 119 secvenţe de pauză din 124. Eroarea globală, de 9,8%, este afectată de confuziile între regiunile vocalice şi cele consonantice. Ulterior, vom încerca minimizarea acestei erori, prin introducerea mai multor clase (categorii) de consoane şi introducerea în vectorii de trăsături a unor noi parametri de intrare.

4. Modulul de detecţie a valorilor formantice F_0,F_1,…,F_4

Metodele de extragere a informaţiei prozodice sunt clasificate, funcţie de domeniul de analiză a datelor, în două categorii:

Metode de analiză în domeniul timp: a) autocorelaţia; b) metoda diferenţelor AMDF;

Metode de analiză în domeniul frecvenţelor (informaţie spectrală): a) HPS - produsul spectrelor armonice; b) analiza cepstrală.

Toate metodele de analiză folosesc la rulare aceiaşi parametri (dimensiunea ferestrei de analiză, *W* şi valoarea pasului de deplasare a ferestrei) pentru ca vectorii de ieşire să aibă aceeaşi lungime şi să poată fi comparaţi în final în modulul decizional. Se recomandă ca durata ferestrei de analiză să fie de minim 4-5 ori perioada fundamentalei maxime pentru o determinare bună a F_0. Pentru Fs = 22050 Hz, considerând limita inferioară a domeniului de interes în detecţia frecvenţelor formantice de 80Hz, se obţine o fereastră de minim 1100 eşantioane, iar pentru Fs = 16000Hz, *W* trebuie să fie de 800 de eşantioane.

Metodele de analiză în domeniul frecvenţelor necesită ferestre de analiză de durate mai mici, dar chiar şi în cazul acestora, pentru metoda cepstrală durata minimă este de 615 eşantioane. Deci dimensiunea ferestrei de analiză va fi de minim 1024 de eşantioane (durată de 46 – 64 ms). Dacă *W* este prea mare, e posibil ca ea să includă mai multe foneme, sau tranziţii rapide în fluctuaţiile valorilor formantice şi deci există riscul ca

prin medierea rezultatelor detecţia sa aibă o imprecizie mare, inclusiv să redea informaţii asociate unei întregi grupări de foneme.

Metode de analiză în timp (autocorelaţia, AMDF)

Aceste metode sunt des întâlnite în literatura de specialitate şi sunt considerate metode care conduc la o bună detecţie a frecvenţei fundamentale. Pentru extractorii de frecvenţă fundamentală pe care i-am implementat, metodele de analiză în domeniul timp au dat mai puţine erori în detecţia F_0 decât cele în domeniul frecvenţelor.

Funcţia de corelaţie aplicată pe două semnale *x* şi *y* oferă informaţii legate de similitudinile dintre acestea. Această metodă de comparaţie este utilizată pentru detecţia unor regularităţi (legate de periodicitatea semnalelor, în cazul în care acestea pot fi considerate periodice sau cvasi-periodice).

$$R_{XY}[n]=\sum_{k=0}^{W-n} x[k]\cdot y[k+n]; \quad R_{YX}[n]=\sum_{k=0}^{W-n} y[k]\cdot x[k+n], \qquad n=\overline{0,W} \qquad (1)$$

$$C[n]=R_{XX}[n]=\sum_{k} x_k \cdot x_{n+k} \quad \text{(autocorelaţia)} \qquad (2)$$

Valorile de maxim local sunt date de periodicitatea semnalului. Valoarea maximă găsită în $R_{XY}[0]$ nu trebuie luată în consideraţie; ea reprezintă energia semnalului într-o fereastră de analiză de dimensiune fixată *W*. Următorul maxim local este asociat cu perioada fundamentală şi căutarea lui poate fi limitată într-un interval de valori dat de banda de căutare a frecvenţei fundamentale [70, 500] Hz. Pentru o frecvenţă de eşantionare de Fs=16000 Hz se va căuta în semnalul furnizat de funcţia de autocorelaţie între valorile R_{XX} [Fs/F_{0max}, Fs/F_{0min}], adică R_{XX} [16000/500, 16000/70]=R_{XX} [32, 229].

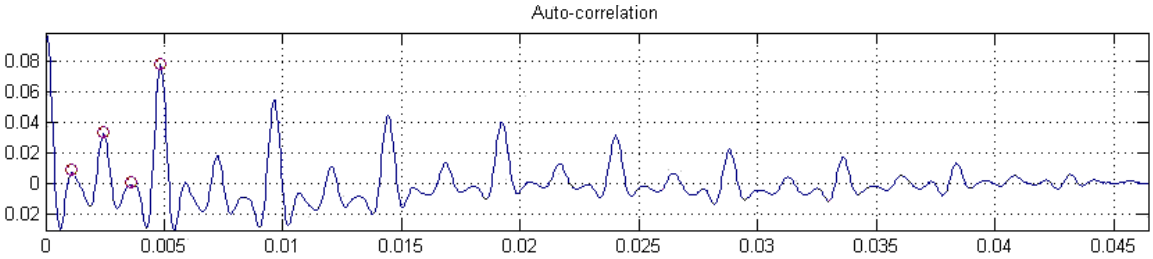

Figura 3. Detecţie valori formanţi prin metoda autocorelaţiei
a) extragere maxime corespunzătoare valori formantice F_0, F_1

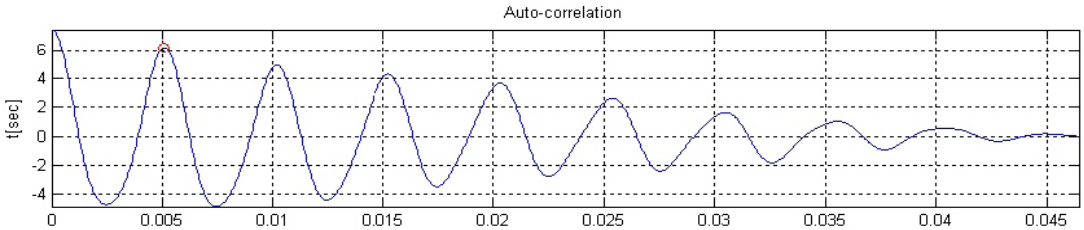

Figura 3. b) caz defavorabil - extragere valoare frecvenţă fundamentală F_0

Valorile frecvenţei fundamentale F_0 sunt uşor de extras folosind această metodă, dar există situaţii când informaţiile corelate cu ceilalţi formanţi F_1 şi F_2 nu sunt prezente în semnalul funcţiei de corelaţie. În fig. 3, pentru primul semnal există mai multe maxime locale cu periodicităţi diferite corespunzătoare primilor formanţi, pentru al doilea

semnal se observă doar distribuţia periodică a unui singur maxim local (corespunzător lui F_0).

Metoda AMDF – magnitudinea medie a funcţiei diferenţă (Average Magnitude Difference Function) se aseamănă ca algoritm cu funcţia de autocorelaţie; diferenţa între cele două constă în faptul că funcţia diferenţă nu necesită operaţii de înmulţire:

$$D_n = \sum_k \left(x_n - x_{n+k}\right), \quad n,k \in \overline{1,W} \qquad (3)$$

La metoda diferenţelor AMDF, minimele locale sunt cele care servesc la calcularea valorii perioadei fundamentale T_0, respectiv a primilor formanţi T_1, T_2, spre deosebire de metoda autocorelaţiei, unde maximele locale erau folosite în determinarea lui T_0.

Metoda permite o mai bună detecţie a formanţilor, în special a primului formant F_1, faţă de funcţia de autocorelaţie, pentru care detecţia este mai dificilă datorită estompării valorilor corespunzătoare formanţilor, prin funcţia de multiplicare.

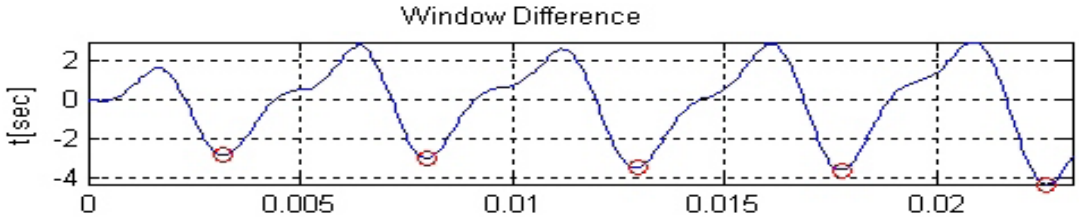

Figura 4. Detecţie valori formantice prin metoda funcţiei diferenţă AMDF

Aceleaşi dificultăţi întâlnite la metoda autocorelaţiei sunt valabile şi pentru metoda diferenţelor, vectorul pe care se caută valorile formantice fiind construit de aceasta dată din distanţele dintre minimele locale ale semnalului diferenţă.

Metode de analiză spectrale (HPS, metoda cepstrală)

Metoda cepstrală se bazează pe separarea componentelor spectrale care ţin de modul în care este generat sunetul H_g (depind de frecvenţa de rezonanţă a corzilor vocale, dimensiunea tubului generator şi pot oferi informaţii despre frecvenţa fundamentală), de cele care depind de modul de filtrare a semnalului vocal H_f (şi care descriu modelul rezonator al cavităţilor în care se formează sunetul vocal). În formula de calcul al cepstrului (spectrul spectrului logaritmat), operaţia de înmulţire dintre spectrul semnalul excitator şi spectrul funcţiei de transfer este transformată prin logaritmare într-o operaţie de adunare. Cele două componente sunt separabile; căutarea maximului corespunzător frecvenţei fundamentale se face în banda [70, 500] Hz.

$$H(\omega) = H_g(\omega) \cdot H_f(\omega), \quad cepstrum = IFFT\left(\log\left|FFT(s)\right|\right) \qquad (4)$$

$$cepstrum == \Im^{-1}\left(\log\left|H_g(\omega)\cdot H_f(\omega)\right|\right) = \Im^{-1}\left(\log\left|H_g(\omega)\right|\right) + \Im^{-1}\left(\log\left|H_f(\omega)\right|\right)$$

Deoarece calculul cepstrului implică trecerea în domeniul quefrenţelor (spectrul spectrului), algoritmul de calcul are cea mai mare complexitate dintre cele 4 metode de detecţie de informaţie prozodică. Se recomandă W de minim 1024 eşantioane, deoarece pentru o frecvenţă de eşantionare de Fs=22050 pentru banda de frecvenţe [70, 500]Hz avem nevoie ca vectorul cepstral (corespunzător frecvenţelor pozitive) să conţină minim Fs/F_{0min}=22050/70=315 eşantioane. Deoarece jumătate din spectrul unui semnal este

asociat frecvenţelor pozitive, iar cealaltă jumătate frecvenţelor negative, *W* minimă este de $2 \cdot Fs/F_{0min} = 650$ eşantioane.

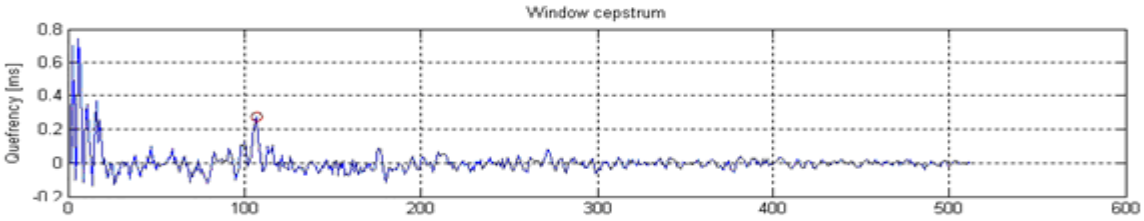

Figura 5. Detecţie valoare frecvenţa fundamentală F0 prin metoda cepstrală

Pentru determinarea formanţilor prin metoda cepstrală se calculează „spectrul netezit" (o anvelopă a spectrului) prin aplicarea unui FTJ în cepstru. Se păstrează din cepstru doar primele L valori (restul fiind anulate) şi se aplică transformata inversă, obţinând un semnal spectral cu tranziţii lente („componenta frecvenţelor înalte" fiind eliminată).

$$cepstrum^* = \begin{cases} cepstrum[k] & ,k \le L \quad \text{sau} \quad c > N_w - L \\ 0 & ,altfel \end{cases}$$

$$S^*(\varpi) = \exp\left(FFT\left(cepstrum^*\right)\right) \qquad (5)$$

Funcţie de dimensiunea ferestrei de liftare *L* avem mai multe sau mai puţine tranziţii în spectru (frecvenţa de tăiere a FTJ este mai mică sau mai mare). În anumite situaţii pentru valori mici ale lui *L*, unii formanţi pot fi greu de găsit sau nu apar în spectru, cum este şi cazul lui F_1 când *L*=30 (Figura 6.a). Pentru valori mai mari ale lui *L* sunt mai mulţi candidaţi, fiind necesară alegerea unei valori reprezentative.

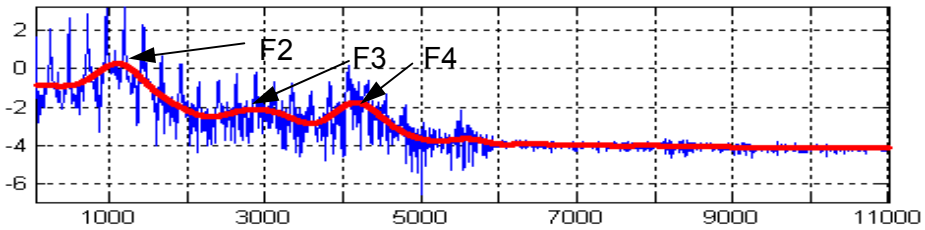

Figura.6a. Spectru + spectru „netezit" vocala 'a', pentru L=30

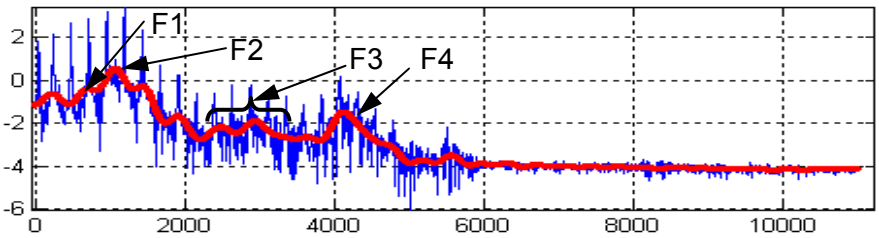

Figura.6b. Spectru + spectru „netezit" vocala 'a', pentru L=60

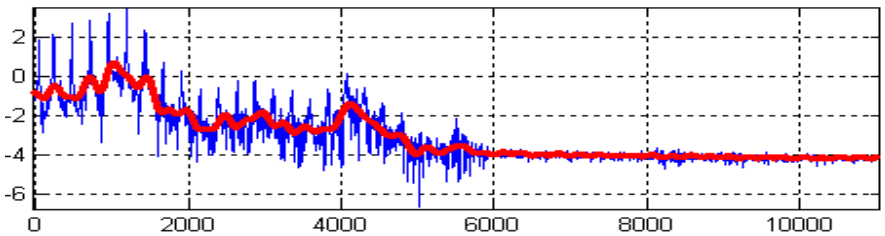

Figura 6c. Spectru + spectru „netezit" vocala 'a', pentru L=90

Ca direcţie viitoare de cercetare se va determina valoarea optimală a lui *L* (dimensiunea ferestrei de liftare), pentru o detecţie mai bună a formanţilor, printr-un algoritm care să permită asocierea de mulţimi fuzzy în care să fie căutaţi formanţii. Alegerea mulţimilor fuzzy va fi realizată funcţie de valoarea lui F_0, determinată anterior de extractorul de frecvenţă fundamentală. Al doilea autor a propus o metodă ce utilizează o funcţia diferenţă modificată combinată cu cea a spectrului produs de armonici pentru a realiza o detecţie mai bună a valorii frecvenţei fundamentale. Metoda este descrisă pe larg în (Teodorescu H.N., 2006).

Metoda HPS constă în determinarea spectrului semnalului, decimarea acestuia cu factori de decimare $1/2$, $1/3$, $1/4$,..., şi realizarea produsului între semnalele realizate. Decimarea se realizează printr-o parcurgere a semnalului şi selectarea eşantioanelor cu un pas (2,3,4,...).Metoda HPS – Harmonic Product Spectrum se bazează pe proprietatea că în spectrul unui semnal periodic cu frecvenţa fundamentală F_0, apar maxime la multiplii acestei frecvenţe $2\cdot F_0$, $3\cdot F_0$, $4\cdot F_0$,... (armonicele fundamentalei). Dacă semnalul este rescalat cu factori $1/2$, $1/3$, $1/4$,..., după operaţia de decimare, prin produsul semnalelor rezultate care au toate un maxim spectral în jurul frecvenţei fundamentale F_0, celelalte maxime vor dispărea sau vor fi puternic atenuate. Decimarea cu un factor $1/k$ a valorilor spectrale se poate face fie selectând o valoare (de obicei prima) dintr-un set de *k* valori consecutive, fie realizând media celor *k* valori,

$$H_n^{\,k} = H_{k\cdot n}^{0} \text{ (decimare) sau } H_n^{\,k} = \frac{1}{k}\sum_{i=0}^{k-1} H_{k\cdot n+i}^{0} \qquad (6)$$

unde H^0 este spectrul semnalului şi H^k semnalul rezultat după scalarea cu un factor $1/k$.

Metoda HPS ridică probleme atunci când avem subarmonici de amplitudine mare ale lui F_0. În special prima subarmonică este cea care conduce la detecţii eronate de F_0. Aplicarea metodei HPS pentru determinarea formanţilor este anevoioasă, chiar şi în lui F_1 şi F_2 deoarece la fiecare înjumătăţire a spectrului, banda de frecvenţe rămasă este tot mai îngustă. De exemplu, la un semnal achiziţionat la o frecvenţă de eşantionare de 16 kHz, spectrul util (al frecvenţelor pozitive) este [0-8000] Hz. După aplicarea algoritmului de decimare HPS de 3 ori, banda de frecvenţe rămasă este de [0-1000] Hz. La a patra aplicare, banda de frecvenţe rămasă nu mai include F_1! Metoda HPS este utilă doar atunci când se doreşte accentuarea valorilor formantice de frecvenţă joasă, fiind acceptabilă în cazul frecvenţei fundamentale.

5. Metoda hibridă comparativă

Pentru a putea realiza compararea rezultatelor, toate metodele de extragere F_0 sunt rulate cu aceiaşi parametri la intrare (dimensiunea ferestrei de analiză, pas de deplasare etc.). Algoritmul de selecţie se aplică pentru situaţiile în care diferenţa dintre valoarea detectată printr-o metoda *vF0_m1* este cu un procent de 20% mai mică sau mai mare decât valoarea furnizată de metoda altă metodă *vF0_m2*. În caz contrar, se consideră valoarea lui F0 ca fiind media celor două valori.

```
IF (vF0_m1 > vF0_m2*(1+percent)) OR (vF0_m1 < vF0_m2*(1-percent))
THEN compară cu N vecini la stânga şi [optional cu N vecini la dreapta]
```

```
        calculează media vF0 şi abaterea standard σ_F0
        IF |vF0_m1 – vF0 |< |vF0_m2 – vF0 | THEN vF0=vF0_m1
        ELSE                                    vF0=vF0_m2
ELSE    vF0 = (vF0_m1+vF0_m2)/2
```

Pentru a decide valoarea frecvenţei fundamentale F_0, se selectează dintre valorile comparate acea valoare care este mai apropiată de valorile determinate anterior. Numărul de vecini *N* folosiţi pentru această comparaţie depinde de dimensiunea pasului de deplasare a ferestrei de analiză, şi se alege astfel încât să nu comparăm valori ale F_0 aflate la o distanţă mai mare de 3-5 ms. Sunt comparate valori ale lui F_0 pe durate mici de timp, ca să nu apară fluctuaţii mari ale lui F_0. Stabilirea ponderilor ce sunt asociate fiecărui extractor este realizată statistic pe baza estimării raportului dintre detecţiile eronate de fundamentală şi detecţiilor corecte. Astfel, o metodă cu mai puţine erori are o influenţă mai mare asupra rezultatului final.

6. Concluzii

Validarea rezultatelor furnizate în final este realizată prin mai multe metode de detecţie de F0, respectiv pe baza comparaţiilor cu valorile anterioare ale lui F0 (pe intervale mici de timp nu pot avea loc variaţii bruşte). S-au comparat vizual valorile fundamentalei date de sistemul nostru hibrid cu rezultatele furnizate de alte instrumente de detecţie puse la dispoziţia utilizatorilor pe Internet: WASP (http://www.phon.ucl.ac.uk/resource/sfs/wasp.htm), Praat (http://www.praat.org) şi s-a constat că rezultatele noastre sunt mai bune. Studiul s-a realizat pentru un număr de 12 pronunţii de fraze aflate pe situl SROL la secţiunea „Fraze->Particularităţi lingvistice" [www.etc.tuiasi.ro/sibm/romanian_spoken_language/ro/fraze_sd_arhiva.htm]. Urmează ca pe viitor să se realizeze o statistică automată pe un număr mai mare de fişiere.

Din simulările efectuate s-a constatat că, din punct de vedere al erorilor de detecţie F0, cele mai robuste sunt metodele de analiză în domeniul timp. Dintre acestea, metoda autocorelaţiei este cea care oferă cele mai bune rezultate în detecţie. Dintre metodele de analiză în domeniul spectral, atât metoda HPS, cât şi metoda cepstrală ridică probleme mai ales la nivelul primei armonici, respectiv a primei subarmonici. O detecţie mai bună a frecvenţei fundamentale permite o extragere mai exactă a valorilor formantice (definirea intervalelor de căutare a formanţilor se face în funcţie de F0). Dintre metodele de extragere F1, F2, F3 metoda cepstrală s-a dovedit cea mai sigură. O problemă doar parţial rezolvată, de care urmează să ne ocupăm, o constituie segmentarea mai precisă a zonelor vocalice.

Mulţumiri. Autorii mulţumesc recenzorilor pentru observaţiile pertinente.

Referinţe bibliografice

Rowden C. (1991), *Speech Processing*, McGraw - Hill Book Company, Chapter 2, pp.35-74.

Rabiner L.R., Juang B.H. (1993), *Fundamentals of Speech Recognition* Englewood Cliffs, N.J.

Rabiner L.R. Schafer R. W. (1978), *Digital Processing of Speech Signal*, Prentice-Hall, Inc. Englewood Clifford, pp. 11-65

Calliope (1989), *La parole et son traitement automatique*, ISBN 2-225-81516-X, Masson, France

O'Shaughnessy, D.O. (1987), *Speech Communication Human and Machine*, INRS-Telecom.

Cristea P., Valsan, Z. (1999) *New Cepstrum Frequency Scale for Neural Network Speaker Verification*, Proc. of the VIth International Conference on Electronics, Circuits and Systems, ICECS, 5-8 sept. Cyprus.

Teodorescu H.N., (2006), *Aplicaţii ale analizei şi sintezei semnalului vocal*, Iaşi, Capitolul 2.

Teodorescu H.N., Trandabăţ D., Feraru M., Zbancioc M., Luca R., (2006a) "*A Corpus of the Sounds in the Romanian Spoken Language for Language-Related Education*", Ist International Conference on Human and Material Resources in Foreign Language Learning - HMRFLL, Murcia, Spain

Teodorescu H.N., Feraru M., Trandabat D., Zbancioc M. (2006b), "*Limba română vorbită*", Atelierul Resurse lingvistice şi instrumente pentru prelucrarea limbii române, ConsILR-06, 3-4, Iaşi, România, Editura Universitatii "Al.I. Cuza" Iasi

Teodorescu H.N., Zbancioc M., Mihailescu E. (2006c), "*Speech Technology and Bio-Medical Engineering Teaching Based on the Web – A New Tool and Case Study*", International Conference on Interactive Computed Aided Learning, ICL, September 27 - 29, Villach, Austria

Zbancioc M. (2006), *Tools for the Archive of the Romanian Language Sounds Project*, 4th European Conference on Intelligent Systems and Technologies, ECIT'2006, sept.20-23, Iaşi, Romania

Proiectul Sunetele Limbii Române, http://iit.iit.tuiasi.ro/romanain_spoken_language/index.htm

Praat, Boersma P., Weenink D., Institute of Phonetic Sciences, Amsterdam: http://www.praat.org

WASP web page, http://www.phon.ucl.ac.uk/resource/sfs/wasp.htm

Arbore de decizie *See5* http://www.rulequest.com/see5-win.html

CAPITOLUL 2

PLATFORME, DICŢIONARE ŞI CORPUSURI ADNOTATE PENTRU PRELUCRAREA TEXTELOR

LIMBA ROMÂNĂ ÎN PERSPECTIVA CLARIN

DAN CRISTEA[1,2], IONUŢ CRISTIAN PISTOL[1]

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

[2]*Institul de Informatică Teoretică, Academia Română, Filiala Iaşi*

{dcristea, ipistol}@info.uaic.ro

Rezumat

CLARIN[1] este un proiect PC7 care îşi propune dezvoltarea unei infrastructuri de resurse şi tehnologii lingvistice, reunind ultimele progrese în domeniul prelucrării limbajului natural, într-o formă accesibilă celor din afara domeniului, cum ar fi specialiştii în ştiinţele umane, ştiinţele sociale şi chiar publicului larg. Lucrarea propune o abordare teoretică şi aplicativă de integrare a instrumentelor de procesare lingvistică dedicate limbii române, care se încadrează spiritului CLARIN. La baza acestei abordări stă ALPE[2], un meta-sistem de configurare de soluţii în problemele de tratamente aplicate limbilor naturale. Se propune o ierarhie care reuneşte instrumentele cunoscute de procesare lingvistică disponibile pentru limba română. Beneficiile acestei abordări, odată demonstrate, pot fi extinse la nivelul comunităţii globale a "consumatorilor" de procesări lingvistice, deziderat central al proiectului CLARIN.

1. Introducere

Ultimii ani au fost martorii unui interes crescând pentru prelucrarea lingvistică a limbilor europene vorbite în ţările "noului val" al Uniunii Europene, interes ghidat în special prin proiecte europene de cercetare în domeniul prelucrării limbajului natural în diferite scopuri, de la sisteme de traducere automată la sisteme de e-learning. Printre alte manifestări de importanţă naţională ori internaţională, întâlnirile Consorţiului de Informatizare pentru Limba Română (ConsILR[3]), care de câţiva ani s-au transformat în Ateliere de lucru "Resurse lingvistice şi instrumente pentru prelucrarea limbii române", au relevat rezultate tot mai interesante în ceea ce priveşte procesarea limbii române cât şi în crearea de resurse lingvistice româneşti.

De mai mult timp ConsILR militează pentru colectarea informaţiilor de natură lingvistică şi a instrumentelor informatice capabile să proceseze texte în limba română. Aceste eforturi sunt în perfect acord cu preocupări similare manifestate în afara ţării care vizează crearea de infrastructuri la nivel internaţional pentru stocarea resurselor lingvistice şi procesarea limbajului natural. Un exemplu în această direcţie este recent lansatul proiect FP7 CLARIN, în care România este reprezentată prin două instituţii cu statut de partener şi alte două instituţii ca membri. De mare actualitate, în acest context, sunt şi preocupările de dezvoltare de meta-sisteme de procesare lingvistică. Exemple

[1] Common Language Resources and Technology Infrastructure Network
[2] Automated Linguistic Processing Architecture
[3] http://consilr.info.uaic.ro

sunt GATE (Cunningham et al., 2002), UIMA (Ferrucci şi Lally, 2004) şi ALPE (Cristea şi Pistol, 2008), care permit conceperea unor structuri de procesare complexe, prin integrarea de module existente, în vederea construirii de aplicaţii ce presupun procesări lingvistice. Dintre aceste meta-sisteme, ALPE promite inclusiv facilitarea interacţiunii utilizatorilor nespecialişti cu modulele de procesare.

Capitolul doi al acestei lucrări descrie pe scurt proiectul CLARIN, prezentând obiectivele şi avantajele includerii limbii române în această iniţiativă europeană. Capitolul trei prezintă principalele funcţionalităţi oferite de ALPE, iar capitolul patru propune o încercare de sistematizare într-o ierarhie a resurselor de procesare ce se cunosc pentru limba română. Capitolul cinci conţine comentarii referitoare la aspectele practice ale interacţiunii cu o ierarhie ALPE. Capitolul şase descrie planul de lucru şi obiectivele pe termen scurt şi lung, precum şi posibile direcţii noi de dezvoltare din perspectiva limbii române în CLARIN.

2. CLARIN

CLARIN (Váradi et al., 2008) este un proiect-program finanţat de Comisia Europeană, structurat în trei etape, care se desfăşoară pe parcursul anilor 2008-2018. Scopul acestuia este de a crea şi pune la dispoziţia celor interesaţi, cu precădere cercetătorilor din domeniul umanist şi al ştiinţelor sociale, resurse lingvistice şi tehnologii de prelucrare a limbajului, în toate formele lui de manifestare (textuală, vorbire, semne etc.). Ca purtător al conţinutului cultural şi al cunoaşterii civilizaţiilor, ca instrument de comunicare şi componentă a identităţii naţionale, precum şi ca obiect de studiu, limbajul invită acum, din ce în ce mai imperios, la o abordare care să beneficieze de suportul tehnologiilor informaţionale. CLARIN îşi propune crearea unei infrastructuri de cercetare care să facă posibilă partajarea şi reutilizarea resurselor precum şi prelucrarea lor prin instrumente specializate, la o scară care să justifice standardizarea. Totodată, CLARIN urmăreşte să transforme tehnologia actuală, extrem de fragmentată, precum şi resursele existente, ori ce vor fi create în viitor, în servicii interoperabile şi stabile, pe care utilizatorii să le poată accesa sau adapta după nevoi. Ambiţia proiectului este de a crea o arhitectură orientată spre servicii care să faciliteze comunităţii de cercetători umanişti sau din domeniul ştiinţelor sociale obţinerea de extensii şi adaptări în orice manieră imaginabilă, pentru accesul la resurse, pentru prelucrări asupra lor, ori pentru consultaţii. Serviciile vor avea la bază o reţea de centre de mărimi şi tipuri diferite, distribuite în Europa.

CLARIN speră să reunească într-o comunitate specializată toate instituţiile din Europa care, într-un fel ori altul, dispun de resurse lingvistice în format scris, vorbit ori multimodal, ori de tehnologii lingvistice. De asemenea, iniţiativa CLARIN nu poate fi considerată de succes dacă nu va reuşi să atragă masa mare de utilizatori care fac uz de astfel de resurse ori tehnologii în cercetarea lor. Comunitatea CLARIN include actualmente ca membri mai mult de 130 de instituţii din 32 de ţări europene. Dintre acestea, 24 de ţări au exprimat deja acordul de a cofinanţa proiectul.

Beneficiile principale aduse de includerea limbii române între limbile proiectului CLARIN constau, printre altele, în reconsiderarea eforturilor de tehnologizare a limbii române prin prisma standardelor ce vor fi adoptate la nivelul întregii Europe, accesul la tehnologii moderne, posibil de adaptat şi pentru prelucrarea limbii române, oferirea de

servicii de informare relative la tehnologii şi colecţii de resurse, mărirea gradului de utilizare a acestor resurse prin facilitarea accesului la ele a specialiştilor în ştiinţe umane şi sociale şi, nu în ultimul rând, mărirea vizibilităţii eforturilor de creare de tehnologii de procesare şi de resurse specifice limbii române, o dată cu includerea lor în colecţiile proiectului, ce se aşteaptă să fie larg accesate de cercetători. Se speră, totodată, ca feedback-ul oferit de utilizarea mai frecventă şi în situaţii noi a tehnologiilor şi resurselor să ducă şi la îmbunătăţirea calităţii acestora.

3. ALPE

ALPE este un meta-sistem care permite unui utilizator dispunând doar de minime abilităţi informatice să exploateze configuraţii de procesare a documentelor adnotate XML, deja create anterior, sau chiar să creeze altele noi. Generarea unei arhitecturi de procesare (workflow) se realizează ca un proces de navigare într-o ierarhie de scheme de adnotare XML (Cristea et al., 2006, Cristea şi Pistol, 2008). Ierarhia este un graf direcţionat în care nodurile reprezintă scheme de adnotare iar arcele sunt relaţii de subsumare. În acest context spunem că nodul A subsumă nodul B dacă:

- schema B conţine toate elementele schemei A;
- schema B include cel puţin un element (tag sau atribut) ce nu este cuprins în schema A.

Direcţia arcelor din graf este dată de relaţia de subsumare: de la nodul care subsumă către nodul subsumat. Un nod poate subsuma mai multe noduri şi poate fi subsumat de mai multe. Considerând elementul rădăcină ca fiind adnotarea XML vidă (cuprinzând numai identificatorul de format XML), el subsumă toate celelalte noduri, ceea ce înseamnă că nu există noduri izolate.

Arcelor grafului li se pot ataşa module de procesare lingvistică capabile să transforme un fişier ce corespunde formatului de intrare (nodului origine al arcului) într-un fişier corespunzând formatului de ieşire (nodului destinaţie al arcului). De notat că nu întotdeauna un arc, căruia îi corespunde aşadar o relaţie de subsumare de scheme, are ataşat un modul de procesare. Arcele cărora le sunt ataşate minimum un modul de procesare se numesc arce de procesare (processing edges), iar cele care nu au nici un modul ataşat se numesc arce purtătoare (carrier edges).

Pe un graf de acest tip, ALPE defineşte o serie de operaţii ce permit calculul automat al unor lanţuri de procesare. Aceste lanţuri de procesare (processing flows) sunt capabile să transforme automat un document dintr-un format în altul, dacă formatul de intrare şi cel de ieşire corespund la două noduri ale ierarhiei şi dacă modulele corespunzătoare arcelor de procesare sunt disponibile. Cât despre arcele purtătoare, există două moduri în care ele pot interveni într-un lanţ: dacă, incorporate unui lanţ de procesare, ele se combină cu alte arce în nodul de ieşire, atunci ele mixează informaţia din intrare cu cea provenită din celelalte arce, altfel ele blochează lanţul.

ALPE oferă două funcţionalităţi de bază:

- Un utilizator poate îmbogăţi o ierarhie ALPE deja existentă prin:

- oferirea unui nou format de adnotare, ceea ce duce la includerea unui nod nou în ierarhie şi legarea lui automată de nodurile existente, astfel încât relaţia de subsumare să fie respectată;

- oferirea unui nou modul de procesare şi a informaţiilor privind formatele sale de intrare şi ieşire (modalitatea de apel, resursele adiţionale necesare, condiţii de acces, etc.). Acest modul de procesare va fi automat integrat în ierarhia existentă ca arc de procesare.

- Un utilizator poate procesa un document, transmiţându-l ierarhiei ca document de intrare şi indicând un format de ieşire. ALPE calculează toate lanţurile de procesare posibile. Din acest evantai de soluţii, utilizatorul poate alege pentru rulare efectivă pe acelea care corespund propriilor lui criterii de cost/eficienţă.

O primă variantă a sistemului este deja implementată, fiind folosită pentru configurarea unui sistem de tip Întrebare-Răspuns şi pentru procesarea lingvistică a documentelor într-un sistem de e-Learning. ALPE va fi disponibil ca serviciu web, prin intermediul lui utilizatorii urmând a fi capabili să creeze şi utilizeze propriile ierarhii, sau să contribuie la dezvoltarea unei ierarhii globale. Se are în vedere şi promovarea ALPE pentru a fi adoptat ca help-desk interactiv în CLARIN. O cale de dezvoltare atractivă o constituie adaptarea ALPE pentru a lucra în reţele de tip GRID, lucru care ar aduce importante îmbunătăţiri relativ la viteza de procesare şi operabilitate în condiţiile măririi numărului de utilizatori şi a dimensiunii ierarhiilor.

4. Ierarhia de resurse de procesare pentru limba română

Întâlnirile ConsILR precedente (Forăscu et al., 2006; Pistol et al., 2007), participările cu succes la competiţii dedicate limbii române (Iftene et al., 2008; Orasan et al., 2008) şi includerea limbii române în proiecte de cercetare europene (Tufiş et al., 2004; Lemnitzer et al., 2007) indică atât mărirea interesului arătat limbii române de un grup tot mai numeros de cercetători, cât şi existenţa deja a unui set important de instrumente de prelucrare dedicate limbii române. Colectarea acestor instrumente, facilitarea accesului la ele şi a utilizării lor în diverse aplicaţii noi se integrează atât obiectivelor ConsILR cât şi ale CLARIN. În această secţiune încercăm să dăm o caracterizare ca o ierarhie ALPE a unui set de formate XML dedicate aplicaţiilor de prelucrare a limbajului natural, pe care le-am putut identifica în literatura dedicată limbii române.

Nodurile din Figura 1, cu excepţia nodului rădăcină, trebuie înţelese ca reprezentând clase de formate de adnotare cu acelaşi conţinut semantic. De exemplu, diferitele marcaje utilizate pentru elementele lexicale identifică o clasă de formate de adnotare denumită TOK. Definiţiile formatelor din noduri sunt, pe scurt, următoarele:

- BASE: format cu marcaje minimale de început şi sfârşit de document XML;
- TOK: clasă de formate ce marchează elementele lexicale de bază (cuvinte, unităţi de punctuaţie);
- NP: clasă de formate ce marchează grupurile nominale;
- VP: clasă de formate ce marchează grupurile verbale;

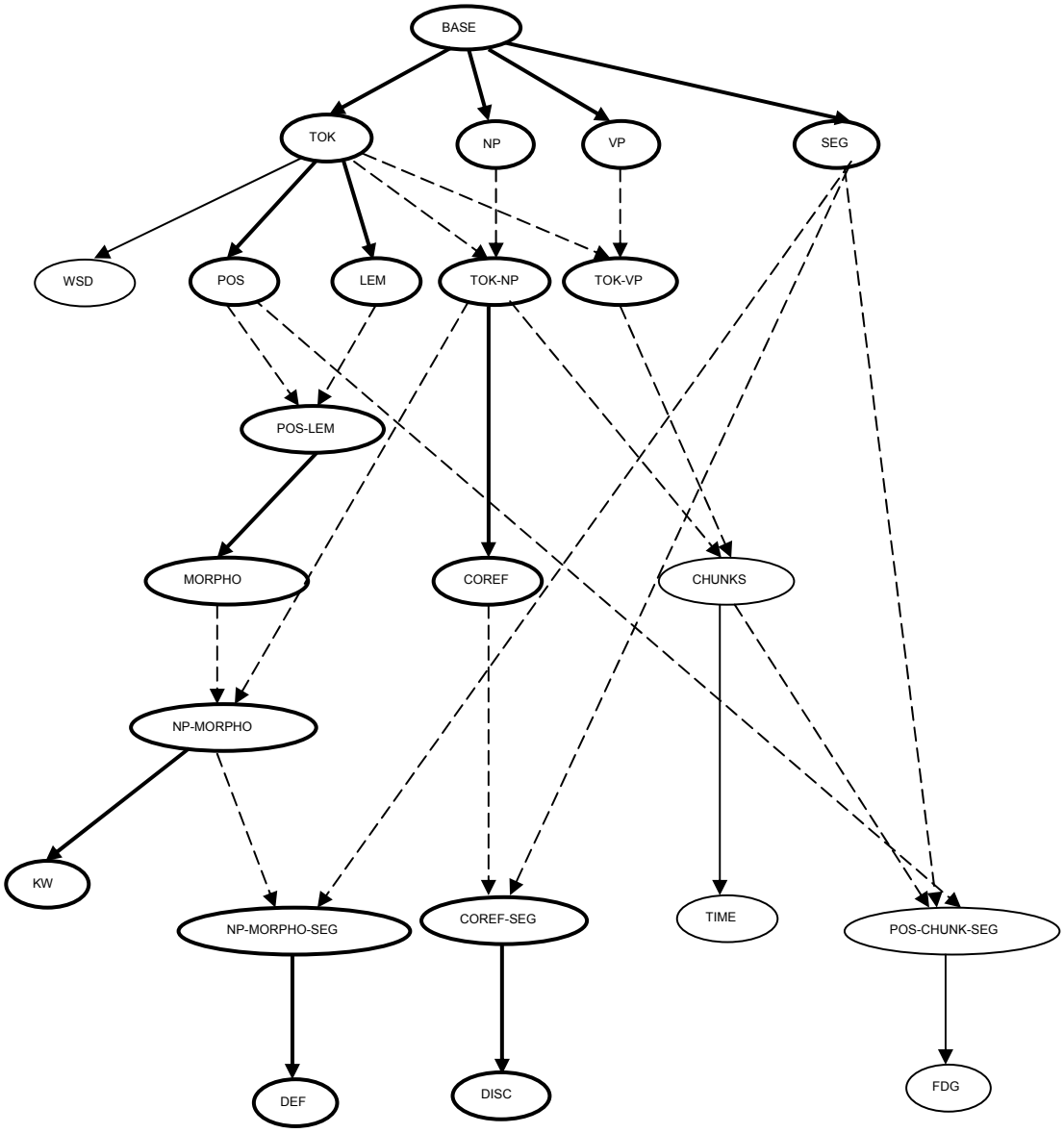

Figura 1: Ierarhia ALPE pentru limba română

- SEG: clasă de formate ce marchează fraze sau unităţi elementare de discurs (propoziţii);
- WSD: clasă de formate ce marchează sensuri ale unităţilor lexicale;
- POS: clasă de formate ce marchează părţile de vorbire ale unităţilor lexicale;
- LEM: clasă de formate ce marchează formele de bază ale unităţilor lexicale (leme);
- TOK-NP: clasă de formate ce reuneşte TOK şi NP;
- TOK-VP: clasă de formate ce reuneşte TOK şi VP;
- POS-LEM: clasă de formate ce reuneşte POS şi LEM;

- MORPHO: clasă de formate ce reuneşte informaţiile morfo-sintactice;
- COREF: clasă de formate ce marchează lanţuri coreferenţiale;
- CHUNKS: clasă de formate ce reuneşte NP şi VP;
- NP-MORPHO: clasă de formate ce reuneşte TOK, NP şi MORPHO;
- KW: clasă de formate ce marchează termeni (cuvinte cheie);
- NP-MORPHO-SEG: clasă de formate ce adaugă la NP-MORPHO şi informaţii de segmentare;
- COREF-SEG: clasă de formate ce reuneşte COREF şi SEG;
- TIME: clasă de formate ce adaugă marcaje pentru adnotarea temporală;
- POS-CHUNK-SEG: clasă de formate ce reuneşte POS, CHUNK şi SEG;
- DEF: clasă de formate ce adaugă marcaje pentru definiţii;
- DISC: clasă de formate ce adaugă marcaje pentru structura de discurs;
- FDG: clasă de formate ce adaugă marcaje pentru dependenţele funcţionale.

Arcele îngroşate indică existenţa unuia sau a mai multor module de procesare corespunzătoare. O parte din modulele considerate în această ierarhie provin din:

- Serviciile web ale ICIA (Tufiş et al., 2007);
- Adnotatorul de expresii temporale (Forăscu şi Solomon, 2004);
- Rezolvitoarele de anaforă AR-Engine şi RARE (Cristea et al., 2002; Pavel et al., 2007);
- Parserul de discurs (Cristea et al., 2005);
- POS taggere (Tufiş şi Dragomirescu, 2004);
- Dezambiguitorul semantic (Ion şi Tufiş, 2004).

Alte detalii privind resursele de procesare disponibile pentru limba română pot fi găsite în (Cristea şi Tufiş, 2002; Forăscu et al. 2007; Pistol et al., 2008).

Arcele subţiri indică posibila existenţă a unor module de procesare corespunzătoare, dar indisponibile integrării în ierarhie în perspectiva imediată. Arcele marcate cu linii întrerupte sunt arce *carrier*. Nodurile îngroşate sunt noduri ce pot fi atinse de lanţuri de procesare din orice alt nod al ierarhiei. Nodurile marcate cu linie subţire nu pot fi atinse de lanţuri de procesare decât plecând din noduri subsumate lor.

În forma din Figura 1, ierarhia ar permite numeroase prelucrări semnificative, cum ar fi: adnotarea automată a structurii de discurs, marcarea definiţiilor, marcarea unui text cu informaţie temporală, informaţie sintactică, precum şi operaţii de combinare şi simplificare a adnotărilor.

5. Interacţiunea cu o ierarhie ALPE

Premizele care stau la baza interacţionării cu o ierarhie ALPE sunt:

a. pe parcursul procesărilor, documentul de bază (*hub*) este neschimbat şi doar adnotările XML aplicate lui suferă transformări;

b. o schemă XML (un nod al ierarhiei) trebuie înţeleasă ca purtătoarea unui mesaj de adnotare, cu semnificaţie cunoscută, care complementează un document.

Premiza *a* vede spaţiul nodurilor ierarhiei ca adnotări diferite aplicate aceluiaşi document. Această restricţie, desigur, nu este compatibilă cu multe aplicaţii din NLP.

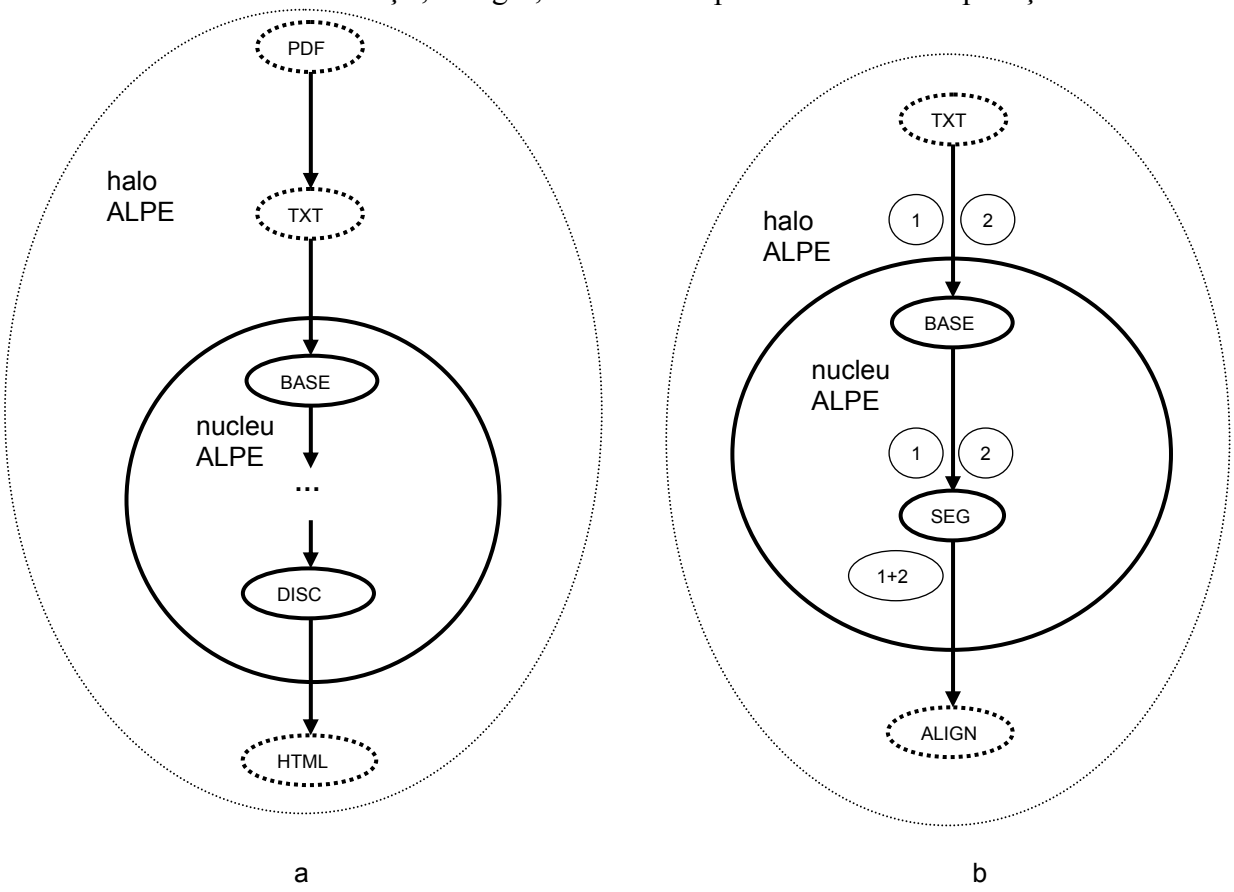

Figura 2: Conectarea nucleului ALPE cu haloul

Exemplele în care documentul de bază se modifică pe parcursul procesării sunt numeroase: transformarea textelor din formate diferite de XML în formate XML (de exemplu, convertoare pdf-txt, html-xml), programe care „impurifică" textul iniţial (de exemplu, prin adăugarea de blancuri ori newline-uri), POS-taggere ori lematizatoare care primesc un text în intrare şi întorc ieşiri tabulare (câte un element lexical pe linie, de exemplu), programe care modifică substanţial textul din intrare (rezumatoare, traducătoare automate etc.), ori care convertesc informaţia dintr-un mediu în altul (convertoare *text-to-speech*, *speech-to-text* etc.).

Acomodarea diversităţii imense de aplicaţii în universul modelului nostru presupune percepţia ierarhiei altfel decât unicul spaţiu al prelucrărilor. În jurul acestuia trebuie înţeles că gravitează, ca un halo, o serie întreagă de standarde care nu reprezintă notaţii XML aplicate unui document, ori dacă sunt adnotări XML, atunci ele nu se aplică asupra aceluiaşi document. Dacă nodurile din afara nucleului ALPE sunt legate de

noduri ale nucleului, înseamnă că există procese capabile de astfel de transformări între nodurile corespunzătoare.

Spre exemplu, Figura 2a arată situaţia unei aplicaţii de rezumare în care textul de intrare are formatul pdf, iar rezumatul este postat ca document html pe Web.

În cazul unui aliniator de texte (Figura 2b), situaţia este diferită: două documente separate sunt prelucrate (în paralel ori serial) de la un format TXT (exterior nucleului ALPE) până la un format XML în care sunt marcate frazele (segment), după care ele sunt date unui aliniator care, ieşind din nou în afara nucleului ALPE, produce alinierea. Ambele fişiere de intrare sunt prelucrate de acelaşi lanţ de procesare ALPE, rulat pe fiecare fişier în parte. Rezultatul celor două lanţuri de procesare este combinat în afara ierarhiei ALPE de către aliniator. De notat că aceasta poate fi tot un fişier XML.

Alte exemple de aplicaţii ce presupune o migrare între nucleul şi haloul ALPE sunt date de generatorul text-voce şi analizorul voce-text. Înregistrările sonore ce au rol de ieşire, respectiv intrare în aceste aplicaţii se regăsesc în noduri din afara nucleului ALPE, pentru că se referă la formate diferite de cele prelucrate în nucleu (textuale). Ele diferă însă atât prin format (ce nu mai este XML) cât şi ca tip de document, trecând de la document scris la înregistrare audio. ALPE poate astfel funcţiona ca suport pentru aplicaţii multimodale ce includ etape de procesare pe text.

Premiza *b* prevede că unei ierarhii ALPE îi corespunde o colecţie de standarde de notaţii XML. Diversitatea extravagantă dată de un spaţiu al numelor neconstrâns de un standard la care să adere majoritatea utilizatorilor de XML în aplicaţiile de prelucrări lingvistice trebuie însă să găsească o expresie în model. Soluţia constă în a vedea un nod ALPE ca pe un nor de notaţii care poartă aceeaşi semnificaţie semantică. De exemplu, un cuvânt (*token*) poate fi notat ca un element XML `TOK`, sau `WRD`, sau `W`. O marcare morfologică se poate exprima printr-o diversitate de notaţii, de la includerea tuturor informaţiilor într-un singur atribut (ca în cazul tag-setului MULTEXT) până la separarea atributelor morfologice într-un set de etichete (ca în cazul tagsetului XCES-EAGLES). Este clar că o procesare similară, indiferent de formatul de intrare a documentelor, duce la economisirea de resurse de calcul şi permite uniformizarea lanţurilor de prelucrări. Este, de aceea, de dorit ca nucleul ALPE să integreze notaţii universal acceptate, dacă se poate, consolidate ca standarde, fără însă ca celelalte notaţii să fie interzise uzului utilizatorilor. Programe de conversie (*wrappere*) vor asigura compatibilitatea intrărilor şi ieşirilor din sistem cu cerinţele utilizatorilor, prelucrarea în nucleul sistemului realizându-se în conformitate cu standardele acceptate.

6. Concluzii

Integrarea resurselor dedicate procesării limbii române într-o ierarhie ALPE promite atât sporirea vizibilităţii şi utilizării acestor resurse cât şi mărirea eforturilor dedicate prelucrării electronice a limbii române. Posibilitatea de a utiliza şi compara resursele existente poate duce la crearea de sisteme tot mai complexe de prelucrare şi la ideea îmbunătăţirii modulelor existente ce au performanţe scăzute.

Dezvoltarea ierarhiei ALPE urmează să fie făcută, prin colaborare, în colectivele implicate în procesarea limbii române, în marea lor majoritate agreând întâlnirile ConsILR. Succesul acestui efort va contribui atât la materializarea unuia din scopurile

originare ale ConsILR-ului, dar şi la integrarea cercetărilor româneşti dedicate domeniului ingineriei lingvistice în consorţiul exigent şi elevat al CLARIN. Cum independenţa de limbă este un deziderat care se apropie tot mai mult de realitate în realizarea de aplicaţii în zilele noastre, devine posibil ca propunerea unei ierarhii ALPE în contextul unei anumite limbi, cum e româna, să servească drept iteraţia zero într-un ciclu de dezvoltare a unei ierarhii generale de instrumente de procesare, care să servească toate limbile implicate în proiect.

Referinţe bibliografice

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

Cristea, D., Pistol, I. (2008). Managing Language Resources and Tools Using a Hierarchy of Annotation Schemas. In *Proceedings of the Workshop on Sustainability of Language Resources*, LREC-2008, Marakesh.

Cristea, D., Postolache, O., Dima, G.E., Barbu, C. (2002). AR-Engine – a framework for unrestricted coreference resolution. *Proceedings of Language Resources and Evaluation Conference - LREC 2002*, Las Palmas, vol. VI, 2000-2007.

Cristea D., Postolache O., Pistol I. (2005). Summarisation through Discourse Structure, In Alexander Gelbukh (Ed.): *Computational Linguistics and Intelligent Text Processing*, 6th International Conference CICLing 2005, Mexico City, February 2005, Springer LNCS, vol. 3406, ISBN 3-540-24523-5, pp. 632-644.

Cristea, D.; Tufiş, D. (2002): Resurse lingvistice româneşti şi tehnologii informatice aplicate limbii române. In Ichim, O şi Olariu F.-T. (eds.): *Identitatea limbii şi literaturii române în perspectiva globalizării*, Academia Română, Institutul de Filologie Română „A. Philippide", Editura Trinitas, Iaşi, pp. 211-234.

Cunningham H., D. Maynard, K. Bontcheva, V. Tablan. (2002): GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the ACL (ACL'02)*. Philadelphia, US.

Ferrucci D. şi Lally A. (2004): UIMA: an architectural approach to unstructured information processing in the corporate research environment, *Natural Language Engineering 10*, No. 3-4, 327-348.

Forăscu C., Cristea D., Tufiş D. (eds) (2007) *Lucrările atelierului Resurse lingvistice şi instrumente pentru prelucrarea limbii române* Iaşi, Editura Universitatii "Al.I. Cuza" Iasi, România, ISBN 978-973-703-208-9.

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

Iftene A., Pistol I., Trandabăţ D. (2008). UAIC Participation at QA@CLEF2008. In *Proceedings of the CLEF 2008 Workshop*. 17-19 September. Aarhus, Denmark.

Ion R., Tufiş D. (2004). Multilingual Word Sense Disambiguation Using Aligned Wordnets. In *Romanian Journal on Information Science and Technology*, Dan Tufiş (ed.) Special Issue on BalkaNet, Romanian Academy, vol. 7, no. 2-3, pp. 198-214, ISSN 1453-8245.

Lemnitzer, L., Vertan, C., Killing, A., Ivanov Simov, K., Evans, D., Cristea, D., Monachesi, P. (2007): Improving the Search for Learning Objects with Keywords and Ontologies. In *Creating New Learning Experiences on a Global Scale*, EC-TEL 2007, Lecture Notes in Computer Science, vol. 4753/2007, pp. 202-216, ISBN 978-3-540-75194-6.

Orasan, C., Cristea, D., Mitkov, R., Branco, A. (2008). Anaphora Resolution Exercise - An Overview. In *Proceedings of LREC-2008*. Marakesh.

Pavel, G., Postolache, O., Pistol, I., Cristea, D. (2007): Rezolutia anaforei pentru limba română. In C. Forăscu, D. Tufiş, D. Cristea (eds.): *Resurse lingvistice şi instrumente pentru prelucrarea limbii române*, Editura Universitatii "Al.I. Cuza" Iasi, România, ISBN 978-973-703-208-9.

Pistol I.C., Cristea D., Tufiş D. (eds) (2008) *Lucrările atelierului Resurse lingvistice şi instrumente pentru prelucrarea limbii române* Iaşi, 14-15 decembrie 2007, Editura Universitatii "Al.I. Cuza" Iasi, România, ISBN 978-973-703-208-9.

Tufis, D., Cristea, D., Stamou, S. (2004): BalkaNet: Aims, Methods, Results and Perspectives. A General Overview. In *Romanian Journal of Information Science and Technology*, Romanian Academy, Bucharest, Romania, special issue on Balkanet, July, pp. 9–43, ISSN 1453-8245.

Tufis D., Dragomirescu L. (2004). Tiered Tagging Revisited. In *Proceedings of the 4th LREC Conference*, Lisabona.

Tufiş D., Ion R., Ceauşu A., Ştefănescu D. (2007). Servicii web lingvistice ale ICIA. *Lucrările atelierului "Resurse lingvistice şi instrumente pentru prelucrarea limbii române"* Iaşi, 14-15 decembrie 2007, Editura Universitatii "Al.I. Cuza" Iasi, România, ISBN 978-973-703-208-9.

Váradi T., Krauwer S., Wittenburg P., Wynne M. and Koskenniemi K. (2008): CLARIN: Common Language Resources and Technology Infrastructure. In *Proceedings of Language Resources and Evaluation Conference - LREC 2008*. Marakesh.

PARSAREA ARBORILOR DE SENSURI ŞI SEGMENTAREA LA DEFINIŢII ÎN DICŢIONARUL TEZAUR eDTLR

NECULAI CURTEANU[1], ALEX MORUZ[1,2], DIANA TRANDABĂŢ[1,2], CECILIA BOLEA[1], MĂDĂLINA SPĂTARU[1,2], MARIA HUSARCIUC[1,2]

[1] *Institutul de Informatică Teoretică Iaşi, Academia Română*
[2] *Facultatea de Informatică, Universitatea "Al. I. Cuza", Iaşi;*

curteanu@iit.tuiasi.ro, mmoruz@iit.tuiasi.ro, dtrandabat@info.uaic.ro

Rezumat

Lucrarea de faţă prezintă o nouă metodă de parsare a textului de dicţionar, bazată pe *configuraţii* de tip SCD (Segmentare-Coeziune-Dependenţă), utilizate pentru transformarea eficientă a unui tezaur într-un lexicon structurat. Astfel, strategia dezvoltată reuneşte două configuraţii diferite de parsare: una care identifică şi extrage, pentru fiecare intrare din dicţionar, arborele specific de sensuri (ierarhia sensurilor principale şi secundare), şi o altă configuraţie care parsează fiecare nod din arborele de sensuri cu scopul de a clasifica *definiţiile* acelui sens din dicţionar. Spre deosebire de metodele standard de parsare a textului de dicţionar, în care *toate* câmpurile unei intrări de dicţionar sunt analizate *secvenţial*, noua metodă reuşeşte detaşarea procesului de construire a arborelui de sensuri (*prima configuraţie* SCD) de procesul parsării la definiţiile sensurilor (*a doua configuraţie* SCD). Separarea celor două procese se face în principal prin selectarea *breadth-first* a tuturor *marcherilor* la sensuri, urmată de analiza lor *depth-first*, în fiecare intrare de dicţionar. Pentru clasificarea tipurilor de definiţii la sensurile din Dicţionarul Tezaur al Limbii Române şi parsarea lor a fost realizată o modelare lexical-semantică a definiţiilor, iar strategia de parsare propusă se aplică în cei doi paşi mai sus menţionaţi. Sunt discutate analiza erorilor şi rezultatele parsării la sensuri şi definiţii, precum şi posibilitatea aplicării parserului pe un dicţionar tezaur dintr-o altă limbă.

1. Introducere

Parsarea unui dicţionar presupune transformarea intrărilor ce conţin text sub formă de glosă, într-un format indexabil. Astfel, fiecare intrare de dicţionar este transpusă într-o structură complexă, care conţine atât sensurile definite, precum şi descrieri detaliate ale formei intrării, cu referire la ortografie, morfologie, fonetică, etimologie, uz etc. Scopul acestei lucrări este introducerea unei metode noi de parsare a unui tezaur, bazată pe configuraţii de tip Segmentare-Coeziune-Dependenţă (SCD), (Curteanu, 2006). Spre deosebire de orientările standard în parsarea intrărilor de dicţionar (Neff and Boguraev, 1989), de exemplu sistemul *LexParse* (Hauser and Storrer, 1993; Kammerer, 2000; Lemnitzer and Kunze, 2005) sau gramaticile lexicografice (Curteanu and Amihăesei, 2004; Tufiş *et al.*, 1999), metoda folosită de noi detaşează complet procesul de *construire a arborilor de sensuri* de procesul *parsării definiţiilor* la sensuri.

O *configuraţie* SCD are următoarele componente:

- Un set de clase de marcheri: un *marcher* reprezintă o graniţă pentru o categorie lingvistică specifică;
- O ierarhie de tip arbore, care stabileşte dependenţele dintre clasele de marcheri;
- Un algoritm de parsare, care execută următorii paşi: recunoaşterea marcherilor, identificarea structurilor dintre doi marcheri şi clasificarea acestor structuri ţinând cont de ierarhia claselor de marcheri. Algoritmul poate fi aplicat pe diferite clase sau ierarhii de marcheri, depinzând strict de *semantica* textului ce urmează a fi parsat.

Prima expunere a ideilor de bază ale parsării *Dicţionarului Tezaur al Limbii Române* (referit în continuare prin DTLR) cu o metodă derivată din strategia SCD a fost făcută în (Curteanu *et al.*, 2007), unde este schiţat algoritmul de parsare DSSD (Dictionary Sense Segmentation and Dependency) cu extragerea *ab initio* a marcherilor de sensuri din intrare, în prima etapă, şi parsarea definiţiilor la sensuri într-o a doua etapă. Este inclusă o primă formă a grafului de dependenţe între clasele de marcheri la sensuri. În (Curteanu *et al.*; 2008) sunt prezentate rezultatele teoretice şi de implementare ale parsării DTLR. Lucrarea de faţă aduce următoarele noutăţi: **(1)** Folosim o *primă configuraţie* SCD (SCD-*config*1) pentru a extrage arborele de sensuri din DTLR. SCD-*config*1 corespunde algoritmului de parsare DSSD (Curteanu *et al.*; 2008), care obţine arborii de sensuri din DTLR cu o precizie de 91.18%. **(2)** Pentru a rafina conţinutul lexical-semantic al sensurilor din DTLR, este necesară coborârea analizei în sensurile secundare, la nivelul *definiţiilor* DTLR, care constituie întinderea de text situată între două noduri consecutive din arborele de sensuri al unei intrări. Lucrarea de faţă prezintă şi modelarea definiţiilor din DTLR. Această a doua etapă a parsării DTLR reprezintă o nouă configuraţie SCD, notată SCD-*config*2, ce constă dintr-un set specific de clase de marcheri pentru *segmentarea la definiţii*. Rezultatul final al parsării va fi aplicarea în cascadă a SCD-*config*1 şi SCD-*config*2, în această ordine.

Configuraţiile SCD propuse de noi sunt prezentate în următoarele secţiuni, şi aplicate pentru parsarea DTLR. Astfel, Secţiunea 2 prezintă parsarea la arborii de sensuri (Sense Tree Parsing), în timp ce Secţiunea 3 expune modelarea lexical-semantică pentru diferitele tipuri de definiţii care pot fi găsite într-un nod din arborele de sensuri. Secţiunea 4 analizează rezultatele parsării utilizând cele două configuraţii SCD, urmând apoi prezentarea concluziilor şi dezvoltările pe care le avem în vedere.

2. Parsarea arborilor de sensuri din intrările de dicţionar

Parsarea arborelui de sensuri folosind configuraţii SCD a fost inspirată de comparaţia între clasele de marcheri de sens din DTLR şi clasele de marcheri SCD pentru parsarea textului general (Curteanu, 2006).

Clasele de marcheri folosite în procesul de *parsare a arborilor de sensuri* din intrările DTLR sunt descrise mai jos:

Clasa de marcheri ce conţine *majuscule* (**A.**, **B.**, etc.) reprezintă *nivelul cel mai de sus* al ierarhiei de sensuri a marcherilor DTLR (Fig. 1) pentru orice intrare de dicţionar dată. Când apare, acest marcher desemnează înţelesul cel mai general şi desemnează un *sens principal* al cuvântului definit. Dacă acest nivel are doar un element de acest tip, atunci marcherul corespunzător lipseşte, el fiind înlocuit de un marcher de nivel inferior.

Clasa de marcheri ce conţine *cifre romane* (**I.**, **II.**, etc.) reprezintă *al doilea nivel* de sens pentru o intrare DTLR. Acest nivel este subsumat de un marcher de tip *majusculă*, dacă acesta există; dacă majuscula nu există (sau nu este reprezentată în mod explicit), marcherul de tip cifră romană apare pe nivelul cel mai de sus al arborelui de sensuri. Dacă intrarea lexicală are doar un sens pentru acest nivel al analizei, marcherul nu este reprezentat în mod explicit, el fiind înlocuit de un marcher de nivel inferior.

Clasa de marcheri ce conţine *cifre arabe* (**1.**, **2.**, etc.) reprezintă al *treilea nivel* de sens pentru o intrare DTLR. Acest nivel este subsumat de un marcher de tip cifră romană, dacă există; dacă acesta nu este reprezentat în mod explicit, este subsumat de primul marcher explicit de nivel superior. Dacă intrarea are doar un sens pentru acest nivel al analizei, marcherul nu este reprezentat în mod explicit.

Aceste prime *trei niveluri* codifică *sensurile principale* ale unei intrări DTLR.

Rombul plin reprezintă *al patrulea nivel* de sens şi este folosit pentru a enumera *sensurile secundare* ale unei intrări din DTLR. Acest nivel este subsumat de orice marcher de sens de nivel superior (oricare dintre marcherii unui sens principal).

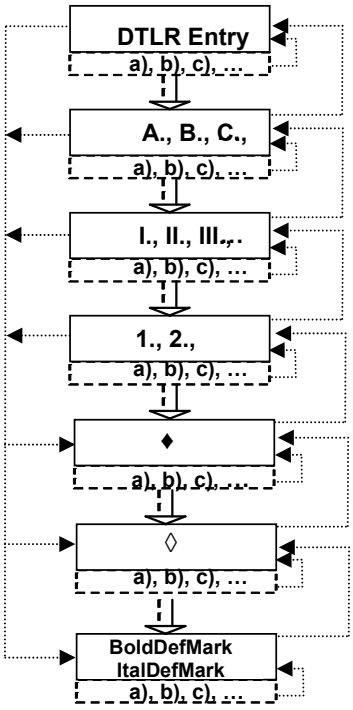

Figura 1: Ierarhia marcherilor DTLR

Marcherul *romb gol* reprezintă *al cincilea nivel* al analizei sensurilor şi este folosită pentru a enumera expresii pentru un *sub-sens secundar* dat. Acest nivel este subsumat de un marcher de tip *romb plin* sau de orice marcher de sens principal.

Marcherii de tip *BoldDefMark* şi *ItalDefMark* sunt delimitatorii definiţiilor de tip *BoldDef,* respectiv *ItalDef* (v. Secţiunea 3).

Marcherii de tip *litere mici* (**a)**, **b)**, etc.) nu reprezintă, în realitate, o clasă distinctă de marcheri de sens, ci mai curând o *procedură* folosită pentru rafinarea, prin *enumerare*

literală, a unui sens sau sub-sens. Un marcher de acest tip nu reprezintă un nivel specific în ierarhia claselor de marcheri, deoarece aparţine nivelului de sens al părintelui. Regulile de bază ale procedurii de *enumerare literală* în DTLR sunt următoarele: **(*i*)** se asociază cu nivelul ierarhiei claselor de marcheri în cadrul căreia apare şi **(*ii*)** poate îngloba sub-sensuri de nivel mai mic (decât nivelul nodului părinte).

Fig. 1 prezintă ierarhia claselor de marcheri din DTLR. *Săgeţile* cu linie întreruptă indică faptul că orice nivel de sens este opţional; *săgeţile continue* indică ierarhia claselor de marcheri. Datorită caracteristicilor sale specifice, *enumerarea literală* este ilustrată pe un nivel *ataşat* nivelului căruia îi este asociat în ierarhie.

Exemplul de mai jos prezintă rezultatul parsării arborelui de sensuri al unei intrări DTLR. Se observă că exemplul de intrare prezentat (***VENIT***^{2}) reprezintă numai secvenţele marcherilor de sens din DTLR (în dicţionar această intrare întinzându-se pe mai mult de două pagini):

```
<entry>
   <hw><VENÍT², -Ă </hw>
   <pos>adj. </pos>
   <senses>
      <definition>…</definition>
      <marker>1.
      <definition>…</definition>
      </marker>
      <marker>2.
      <definition>…</definition>
         <marker>◊
            <marker> a)
            <definition>…</definition>
            </marker>
            <marker> b)
            <definition>…</definition>
            </marker>
            <marker> c)
            <definition>…</definition>
            </marker>
         </marker>
         <marker>◊
            <marker> a)
            <definition>…</definition>
            </marker>
            <marker> b)
            <definition>…</definition>
            </marker>
         </marker>
      </marker>
   </senses>
</entry>
```

3. Modelarea sensurilor din dicţionar cu tipuri specifice de definiţii

Analiza conţinutului semantic al sensurilor în DTLR este realizată cu o *a doua configuraţie* SCD, bazată pe un set de *marcheri de definiţii*. Marcherii sunt folosiţi pentru a crea un număr finit de şabloane care delimitează subsensuri prin informaţii lingvistice distinct detaliate, prin trăsături morfologice, sintactice, semantice sau

pragmatice şi prin aspecte lexicale, etimologice, diacronice şi prozodice. Analizând intrările DTLR, au fost găsite următoarele tipuri de definiţii:

1. *MorfDefs* – definiţii morfologice;
2. *RegDefs* – definiţii scrise cu font *regular*;
3. *BoldDefs* – definiţii scrise cu *bold*;
4. *ItalDefs* – definiţii scrise cu *italic*;
5. *SpecDefs* – definiţii ce conţin specificaţii;
6. *SpSpecDefs* - definiţii scrise cu litere spaţiate, ce conţin *anumite* specificaţii;
7. *DefExems* – exemple la definiţii, cu rolul de a întregi înţelesurile unei definiţii.

Tipurile de definiţii propuse aici primesc roluri funcţionale specifice în descrierea sensurilor principale, secundare, sau de granularitate semantică mai fină (Curteanu *et al.*; 2008). Se pot distinge *două taxonomii* ale definiţiilor din DTLR. Prima conţine următoarele clase:

(obli) *definiţii obligatorii*, care conţine, de exemplu, *MorfDef*-uri şi, pentru fiecare sens din DTLR, *una* din următoarele trei definiţii: *RegDef*, *BoldDef* sau *ItalDef*. Nu există nici o intrare de dicţionar care să nu conţină *MorfDef* şi (cel puţin) *una* dintre definiţiile ce aparţin mulţimii {*RegDef*, *BoldDef*, *ItalDef*}.

(opti) *definiţii opţionale*, de exemplu *SpecDef*s, *SpSpecDef*s şi *DefExem*s, care pot să apară ca *modificatori* sau *specificatori* în faţa unei definiţii obligatorii.

Cealaltă taxonomie împarte definiţiile în:

(auto) *definiţii autonome*, care sunt *RegDef*, *BoldDef* şi *ItalDef*; aceste definiţii având un rol de sine stătător în introducerea sensurilor din DTLR;

(cont) *definiţii contingente*, de exemplu *MorfDef*s, *SpecDef*s, *SpSpecDef*s şi *DefExem*s, care nu pot fi folosite independent, având înţeles doar în *contextul* altor definiţii DTLR.

MorfDef apare obligatoriu în rădăcina oricărei intrări din DTLR, fiind moştenită la nivelurile inferioare ale arborelui de sensuri. *SpecDef*s, *SpSpecDef*s şi *DefExem*s sunt definiţii *contingente*, deoarece ele nu pot defini un (sub)sens în mod autonom, ci numai ca instrumente auxiliare de modificare a altor definiţii autonome sau contingente.

3.1. Definiţiile morfologice – MorfDefs

Definiţiile morfologice (*MorfDef*s) sunt formate din una sau mai multe etichete care descriu categorii morfologice la diferite niveluri ale arborelui de sensuri. Primul element într-o intrare de dicţionar, după cuvântul titlu, este un *MorfDef* complex, o listă de *MorfDef*-uri care detaliază toate categoriile morfologice posibile pentru cuvântul de intrare. Cu cât sensurile devin mai rafinate, cu atât *MorfDef*-urile ulterioare devin mai specifice (sub-liste ale *MorfDef*-ului complex), până când ajung să desemneze o singură categorie morfologică. Dacă definiţia unui sens nou nu conţine şi un *MorfDef*, atunci definiţia morfologică se moşteneşte de la primul sens regent care are un *MorfDef*. În cele ce urmează, exemplele de un anumit tip de definiţii DTLR sunt evidenţiate cu gri:

VERZIŞÓR, -OÁRĂ adj., subst. **I.** Adj. Diminutiv al lui v e r d e (**I 1**)... **II.** Subst. **1.** S. m. (La pl.) Corp de trupă al cavaleriei... **2.** S. m. şi f. (Iht.; prin Munt.) Boiştean... **3.** S. n. (Prin Mold.; în forma *verdişor*) Rachiu cu mentă... **4.** S. f. (Regional) Varietate de struguri... **5.** S. n. (Familiar) Bancnotă de culoare verde...

Expresia regulată care descrie *MorfDef*-urile este:

```
(x)+, x ∈ {"subst.", "adj.", etc.}
```
[1]

3.2. Definiţii de tip regular – RegDefs

Definiţiile scrise cu font *regular* (*RegDef*s) reprezintă cel mai frecvent instrument lingvistic folosit în DTLR pentru a descrie sensuri. *RegDef* corespunde glosei cuvântului de intrare sau unor sintagme care îl conţin, reprezentând descrierea standard a sensurilor în majoritatea dicţionarelor. Expresia regulată de mai jos descrie cea mai generală formă a unei definiţii *RegDef*.

```
(([A-Z]|[a-z])+(\(.+\))*)+
```

O definiţie de tip *RegDef* poate să apară în rădăcina intrării, în sensurile principale sau secundare, poate fi *moştenită* în aceste sensuri (cum ar fi sensul **I.** din **VENÍRE**) sau poate face parte, ca *explicaţie*, din corpul altor două tipuri de definiţii autonome din DTLR: *BoldDef* şi *ItalDef*, ca în exemplul de *BoldDef* "**Bun venit**" sau de *ItalDef* "*Venit naţional*" de mai jos.

VENÍRE s. f. Acţiunea de a v e n i şi rezultatul ei. **I. 1.** Deplasare către cineva sau către ceva; parcurgere a unui traseu pentru a ajunge la un anumit loc,…
◊ E x p r. **Bun venit =** formulă de salut prin care se exprimă mulţumirea în legătură cu sosirea, cu prezenţa cuiva.
◊ *Venit naţional* = parte a produsului economiei naţionale dintr-o perioadă de timp, care rămîne după…

3.3. Definiţii de tip bold – BoldDefs

O definiţie de tip *BoldDef* este folosită cu scopul de a explica sensul unei *sintagme* sau al unei *exprimări* specifice; expresia este scrisă cu litere îngroşate, urmată de un *separator BoldDef* (în general "="[2]) şi de un *RegDef*. De obicei, *BoldDef*-urile rafinează subsensuri specifice, cum ar fi sensurile secundare introduse în DTLR prin ♦ şi ◊. Expresia regulată care descrie forma generală a unui *BoldDef* este dată mai jos:

```
(bold(.+)*)+(separator)(RegDef)
```

◊ **A semăna în verde** = a semăna imediat după arat, cînd arătura este încă proaspătă. … **A ara în verde =** a ara un pămînt care este încă jilav. ...

Există situaţii în care *BoldDef*-urile pot să apară în sensurile principale, inclusiv pe *nivelul-rădăcină* al unei intrări lexicale din DTLR. Un *BoldDef* poate fi foarte complex, conţinând numeroase variante ale expresiei marcate cu caractere bold.

3. A se duce (sau **a merge, a se lăţi,** învechit şi regional, **a ieşi) vestea** (cuiva, a ceva, de ceva etc.) sau **a i se duce** (ori **a-i merge, a i se lăţi,** învechit, rar, **a i se ridica,** regional, **a-i ieşi** cuiva) **vestea, a-i merge** (sau **a i se duce** cuiva) **vestea** şi **povestea,** (învechit şi regional) **a ieşi veste** (de cineva sau de ceva) = a deveni foarte bine cunoscut, a i se duce faima;…,

3.4. Definiţiile de tip italic – ItalDefs

ItalDef-urile sunt similare din punct de vedere sintactic *BoldDef*-urilor, dar sunt diferite din punct de vedere semantic, deoarece ele descriu în general colocaţii, spre deosebire

[1] Lista posibilelor elemente morfologice este evident finită.

2 Uneori, separatorul "=" este înlocuit cu expresii echivalente, cum ar fi "vezi", "v.", "se spune", etc., şi introduce o relaţie de echivalenţă semantică între expresia din stânga şi secvenţa din dreapta.

de *BoldDef*-uri care descriu expresii. Partea de definiţie ce conţine colocaţia este codificată cu caractere cursive (italice).

```
(italic(.+)*)+ (separator)(RegDef)
```

5. *Verde antic* = matostat. ...

VERZÉR subst. (Regional; în sintagma) *Verzerul tilegii* = schimbătoare la roţile plugului...

3.5. Definiţii de specificare – SpecDefs

SpecDef-urile sunt definiţii *contingente* scrise cu font *regular* şi cuprinse, în general, între paranteze. Multe dintre ele sunt abrevieri, cuvinte sau expresii rezervate care denotă diferite contexte de utilizare ale intrării DTLR, cum ar fi: "(Regional)", "(Argou)", "(Fam.[iliar])" etc. Uneori *SpecDef*-urile nu apar între paranteze, dar acest lucru se întâmplă numai în cazurile în care acestea reprezintă cuvinte rezervate sau abrevieri. Expresia regulată care recunoaşte acest tip de definiţie este:

```
\( ([a-z]|[A-Z])+ \) | x; x ∈ {abrevieri}
```

SpecDef-urile sunt folosite la orice nivel în arborele de sensuri şi au ca scop specificarea, 'modificarea' definiţiilor, ca în exemplele ce urmează.

(1) În rădăcina intrării de dicţionar, imediat după *MorfDef*:

VENIÁL, -Ă adj. (Livresc; despre păcate[2], greşeli etc.) Care poate fi iertat (de Biserică); uşor, fără importanţă…

(2) În rădăcina unui sens principal:

2. (Învechit şi regional; despre lichide, substanţe etc.) Veninos **(2).** …

(3) În rădăcina unui sens secundar:

♦ F i g. (Despre oameni) Rău **(A I 1)**; duşmănos; (despre manifestări, stări, acţiuni etc. ale oamenilor) care trădează răutate **(I 1),**...

3.6. Definiţii de specificare scrise spaţiat – SpSpecDefs

Un alt tip de definiţie contingentă este *SpSpecDef*, care precizează mai multe trăsături standard. *SpSpecDef* se scrie cu litere spaţiate şi conţine elemente dintr-o listă prestabilită de abrevieri, având următoarea formă:

```
(([A-Z]|[a-z]) )+
```

SpSpecDef poate să apară la toate nivelurile de sensuri din DTLR, uneori împreună cu alte definiţii contingente, ca în exemplul de mai jos:

2. T r a n z. şi r e f l. F i g. A (se) amărî, a (se) supăra, a (se) necăji, a (se) mînia. *A sa prea iubită inimă ş-a veninat.* PANN, E. II, 94/18.

Unele *referinţe externe* (către sensuri din alte intrări DTLR) sunt scrise tot cu font spaţiat, din acest motiv trebuie verificat întotdeauna dacă un cuvânt scris spaţiat face parte din lista prestabilită de abrevieri care formează *SpSpecDef*-uri sau nu.

Plantă erbacee din familia scrofulariacee, cu florile albe sau trandafirii, care creşte în locuri umede sau mlăştinoase şi care este folosită în medicină pentru proprietăţile ei iritante şi purgative; avrămeasă, (regional) milostivă (v. m i l o s t i v **III 2**), potroacă1 (4), mila-Domnului (v. m i l ă1 **I 6**) (Gratiola officinalis). Cf. hem 2182, conv. lit. xxiii, 1060, brandza, fl. 349, damé, t. 188, barcianu, jahresber. viii, 101,...

3.7. Exemplificări de definiţii – DefExems

Definiţiile *autonome* pot primi unul sau mai multe *exemple* din surse bibliografice referite prin *sigle* sau create de către autorii dicţionarului. *DefExem*-urile au rolul de a rafina sensul definiţiilor autonome şi a tuturor sensurilor mai generale decât ele (sensuri secundare şi principale). O secvenţă de *DefExem*-uri, fiecare fiind urmat de o *siglă*, este următoarea:

A intra în viaţă = **a)** (despre oameni; şi în forma **a păşi în viaţă**) a începe să se confrunte cu realitatea. *Cum a intrat el în viaţă? Cît amor de drept şi bine, Cîtă sinceră frăţie adusese el cu sine?* EMINESCU, O. I, 53. …; **b)** (rar) a începe să activeze, să funcţioneze. *Guvernul cel nou... va intra în luna lui marţiu în viaţă.* VASICI, ap. BARIŢIU, C. II, 47.

4. Modelarea sensurilor din dicţionar cu tipuri specifice de definiţii

4.1. Analiza arborelui de parsare

SCD-*config*1 a fost testată pe mai mult de 500 intrări din dicţionar, de dimensiuni medii şi mari. Rata de succes a fost de 91.18%, fiind calculată prin compararea fişierului de ieşire din program cu fişierul adnotat manual la arbori de sens. Cauzele erorilor găsite în urma parsării intrărilor de dicţionar pot fi grupate în două clase de bază:

I. Inconsecvenţe în scrierea articolului în DTLR

O primă sursă de erori de parsare este lipsa monotoniei valorilor marcherilor la acelaşi nivel din ierarhia marcherilor de sens:

Ex.1. **A.** [**B.** lipseşte] … **C.** etc.;
Ex.2. **2.** [în loc de **1.**]... **2.** etc.;
Ex.3. **a)**… **b)** … **c)** … **b)** [în loc de **d)**]etc.

O soluţie este verificarea *monotoniei stricte* a valorilor marcherilor. Astfel, înainte de definitivarea arborelui de sensuri, este necesară verificarea validităţii succesiunilor de marcheri de pe fiecare nivel de sens.

II. Ambiguităţi în stabilirea regentului şi a subordonatului unui sens

În următoarea secvenţă de marcheri de sens apare o ambiguitate inerentă:

Ex.4. **1. a) b) c)** ◊ [◊]

Problema apare atunci când nu se poate stabili dacă romburile "◊" trebuie considerate ca depinzând de **c)** sau de marcherul de un nivel superior (**1.**). Rezolvarea acestei ambiguităţi depinde de contextul semantic al perechilor de marcheri implicate.

4.2. Analiza tipurilor de definiţii din DTLR

Până în prezent a fost realizată segmentarea elementelor dintre doi marcheri de sens succesivi, ţinând cont de marcherii de definiţii DTLR şi de expresiile regulate cu care pot fi recunoscute. Evaluarea segmentării la definiţii a fost abordată folosind două metrici: *potrivire exactă* şi *suprapunere*. Potrivirea exactă (*exact-match* metric) reprezintă numărul de segmente corect extrase (folosind *precizia*, *recall*-ul şi *F-measure*); suprapunerea (*overlap* metric) reprezintă procentul de cuvinte clasificate

corect (folosind, de asemenea, *precizia*, *recall*-ul şi *F-measure*). Deoarece există situaţii în care segmente de acelaşi tip sunt consecutive, primul şi ultimul cuvânt al fiecărui segment sunt marcate, pentru a putea penaliza clasificarea cuvintelor într-un segment mare în locul unei succesiuni de segmente mai mici.

Pentru evaluare au fost utilizate 52 de intrări din dicţionar, de dimensiuni diferite, ca standard-*gold*, însumând un număr de aproximativ 2000 de segmente şi 22.000 de cuvinte. Rezultatele sunt prezentate în Tabelul 1.

Tabel 1: Evaluarea pentru segmentarea la nivel de definiţii DTLR

Tipul evaluării	Precizie	Recall	F-measure
Potrivire exactă	93.24%	85.41%	89.15%
Suprapunere	97.86%	97.80%	97.83%

După analiza rezultatelor evaluării, am observat că cele mai frecvente erori se datorează segmentării greşite a siglelor. Tabelul 2 prezintă cele mai frecvente zece tipuri de erori.

Tabel 2: Cele mai frecvente zece erori în segmentarea definiţiilor din DTLR

Rezultatul parsării	Parsare gold	% erori introduse
Sigle	Început de Siglă	29.45%
Sigle	Sfârşit de Siglă	28.08%
RegDef	SpecDef	6.39%
RegDef	Sfârşit de RegDef	3.88%
RegDef	Început de RegDef	2.96%
DefExem	ItalMarker	2.73%
RegDef	Început de SpecDef	2.51%
Sigle	RegDef	2.28%
RegDef	Sigle	2.05%
RegDef	SpSpecDef	2.05%

Corectarea segmentării siglelor duce la o *F-measure* de 94.43% pentru metrica de potrivirea-exactă şi de 98.01% pentru metrica de suprapunere.

5. Concluzii

Această lucrare a prezentat o metodă nouă de parsare a intrărilor de dicţionar, în mod concret a tezaurului DTLR, bazată pe configuraţii SCD. Prima configuraţie a exploatat setul de marcheri de sensuri DTLR pentru construirea arborelui de sensuri, obţinând o acurateţe de 91.18%. A doua configuraţie SCD este folosită pentru a parsa definiţiile din DTLR cuprinse în fiecare nod din arborele de sensuri. Acurateţea pentru clasificarea definiţiilor depăşeşte 93%.

Parserul bazat pe configuraţii SCD are avantajul că, odată stabilite în mod adecvat clasele de marcheri şi ierarhia lor pentru un anumit dicţionar tezaur, oricât ar fi acesta de complex (cum este cazul DTLR), programul poate parsa foarte eficient acel tezaur.

Mulţumiri. Rezultatele din această lucrare au fost obţinute în cadrul cercetărilor la grantul eDTLR – PNCDI 2, No. 91_013/18.09.2007. Mulţumiri speciale sunt datorate cercetătoarelor Gabriela Haja şi Elena Dănilă, Institutul de Filologie Română ”Al.

Philippide" Iaşi, pentru discuţiile consistente privind funcţionarea, dependenţele şi moştenirea definiţiilor la sensuri în DTLR (Secţiunea 3).

Referinţe bibliografice

Curteanu, N., E. Amihăesei (2004): Grammar-based Java Parsers for DEX and DTLR Romanian Dictionaries. *Proc.of ECIT-2004 Conference*, Iasi, Romania.

Curteanu, N. (2006): Local and Global Parsing with Functional (F)X bar Theory and SCD Linguistic Strategy. (I.+II.), *Computer Science Journal of Moldova*, Academy of Science of Moldova, Vol. 14 no. 1 (40):74-102 and no. 2 (41):155-182.

Curteanu, N., G. Pavel, C. Vereştiuc, D. Trandabăţ (2007). Parsarea eDTLR cu gramatici în mediul JavaCC. Stadiul actual, probleme şi soluţii de dezvoltare. (Ed. I. Pistol, D. Cristea, D. Tufiş) *Resurse lingvistice şi instrumente pentru prelucrarea limbii române*, ConsILR-2007, Ed. Univ. "Al. I Cuza" Iaşi, p. 87-96.

Curteanu, N., Moruz, A., Trandabăţ, D. (2008). Extracting Sense Trees from the Romanian Thesaurus by Sense Segmentation & Dependency Parsing, *Proceedings of CogAlex Workshop, COLING 2008*, pp. 55-63, ISBN 978-1-905593-56-9.

Hauser, R., Storrer, A. (1993). Dictionary Entry Parsing Using the LexParse System. *Lexikographica 9* (1993), 174-219

Kammerer, M. (2000): *Wöterbuchparsing Grundsätzliche Überlegungen und ein Kurzbericht über praktische Erfahrungen*, http://www.matthias-kammerer.de/content/WBParsing.pdf

Lemnitzer, L., Kunze, C. (2005): *Dictionary Entry Parsing*, ESSLLI 2005 Tutorial.

Neff, M., Boguraev, B. (1989) Dictionaries, Diction-ary Grammars and Dictionary Entry Parsing, *Proc. of the 27th ACL* Vancouver, British Columbia, Canada Pages: 91 - 101

Tufiş, D., Rotaru, G., Barbu, A.M. (1999). Data Sampling, Lemma Selection and a Core Explanatory Dictionary of Romanian. *Proc. of the 5th International Workshop on Computational Lexicography COMPLEX*, Pecs, Hungary, pp. 219-228, 1999

SEGMENTAREA ÎN UNITĂŢI TEXTUALE ATOMICE A INTRĂRILOR DIN DICŢIONARUL LIMBII ROMÂNE ÎN VEDEREA ANALIZEI STRUCTURALE

RADU ION

Institutul de Cercetări pentru Inteligenţă Artificială, Academia Română, Bucureşti – România

radu@racai.ro

Rezumat

Lucrarea de faţă prezintă un algoritm de adnotare a unităţilor textuale atomice care compun definiţiile intrărilor de dicţionar din Dicţionarul Limbii Române (DLR) al Academiei Române. Algoritmul care va fi prezentat se bazează pe colecţii de expresii regulate care sunt aplicate succesiv (atât colecţiile cât şi expresiile din fiecare colecţie) pe intrarea de dicţionar. Ca rezultat, fiecare expresie regulată va „recunoaşte" secvenţe continue de text care au anumite semnificaţii în cadrul definiţiei. Această fază de procesare poate fi folosită ulterior de un parser al cărei gramatici va utiliza adnotările pe post de simboluri terminale. În acest fel, se va simplifica scrierea gramaticii care acceptă o intrare de dicţionar.

1. Introducere

Dicţionarul Limbii Române (DLR) este continuarea Dicţionarului Academiei[1] (DA) a cărui construcţie a început în 1913. El reia enumerarea minuţioasă a fondului lexical de la intrarea *Lojniţă* cu scopul declarat de a inventaria tezaurul lexical al limbii române. Cele două lucrări sunt astfel colectiv cunoscute sub denumirea de Dicţionarul Tezaur al Limbii Române (DTLR) care este „cea mai amplă lucrare lexicografică românească, considerat nu o dată o operă de importanţă naţională" (Sala, 1996).

Proiectul eDTLR[2], început în anul 2007, are drept scop transpunerea DTLR în format electronic cu urmări benefice evidente pentru comunitatea lexicografică românească implicată în dezvoltarea lui dar şi pentru comunitatea lingvisticii computaţionale româneşti (Cristea et al., 2007). În ce priveşte lucrul la DTLR, formatul electronic permite operaţii ca interogarea (pe diverse criterii) şi vizualizarea intrărilor cu o uşurinţă de neimaginat pentru lexicografii secolului trecut. Lingvistica computaţională românească are însă, probabil, cel mai mult de câştigat de pe urma unei astfel de resurse lexicografice monumentale. De la analizele morfologice până la diversitatea enormă de sensuri inventariate, eDTLR este util pentru o multitudine de probleme precum dezambiguizarea semantică automată (engl. *Word Sense Disambiguation*), analiză şi generare morfologică, adnotarea morfosintactică (engl. *Part Of Speech Tagging*). De asemenea, eDTLR este o sursă nepreţuită de validare semantică şi extindere a ontologiei

[1] Române.

[2] https://consilr.info.uaic.ro/edtlr/wiki/index.php?title=Despre_proiect

lexicale pentru limba română, RoWN (Tufiş et al., 2008) prin diversitatea sensurilor inventariate pentru fiecare cuvânt.

Formatul electronic al unui dicţionar este valoros în măsura în care evidenţiază prin adnotări structura intrărilor. Astfel, de cele mai multe ori, o intrare de dicţionar este divizată logic într-o mulţime de sensuri. Fiecare sens conţine o definiţie, exemple de utilizare a cuvântului în sensul respectiv, sensuri secundare etc. Structura intrărilor de dicţionar poate fi utilizată de aplicaţii de Prelucrare Automată a Limbajului Natural (PLN) sau poate fi utilă diferitelor tipuri de interogări care se pot imagina (de exemplu, definiţia primului sens al celui de al doilea omonim al cuvântului „mină"). În general, formatul electronic se obţine printr-o analiză structurală (engl. *Parsing*) a unei intrări de dicţionar furnizată sub formă de text electronic[3].

În cele ce urmează vom descrie pe scurt câteva metode de generare a formatelor electronice ale dicţionarelor şi apoi vom prezenta colecţia noastră de expresii regulate care segmentează o intrare din DLR în unităţi textuale atomice în vederea analizei structurale.

2. *Analiza automată a structurii unei intrări de dicţionar*

Transformarea intrărilor de dicţionar din format text (care este un format electronic nestructurat) în format electronic care evidenţiază structura este o problemă care interesează comunitatea lexicografiei computaţionale în măsura în care se pot crea automat resurse lexicografice computaţionale din diversele formate text ale dicţionarelor.

„Dictionary Parsing Project (DPP)[4]" derulat de grupul de PLN din cadrul USC Information Sciences Institute îşi propune să extragă relaţii semantice (hipernimie, holonimie, relaţii sintagmatice de tipul şofer–vehicul etc.) din „Noah Webster's 1913 Dictionary of the English Language[5]". Pentru aceasta, structura unei intrări în formă text a dicţionarului este iniţial procesată pentru a se obţine forma descrisă cu expresii regulate din Figura 1. Din această formă, intrarea de dicţionar este mai departe prelucrată în direcţia depistării unor unităţi de text denumite „fraze" care au semnificaţii bine-stabilite pentru compoziţia intrării de dicţionar. De exemplu, fraza care delimitează începutul intrării de dicţionar şi care conţine cuvântul-titlu, partea de vorbire şi numărul sensului, se defineşte cu următoarea expresie regulată:

```
HEADWORDLINE := <hw>({(WORD{WORD}*)}+)</hw> <pos>POS</pos>
<sn>NUMBER</sn>
```

în care simbolurile „{,},*,+" fac parte din limbajul de specificare a expresiilor regulate (vezi Figura 1) iar WORD, POS şi NUMBER sunt definiţiile altor expresii regulate care descriu un cuvânt, partea sa de vorbire şi, respectiv, un număr de sens.

Cea mai mare parte a literaturii care se referă la achiziţia şi prelucrarea de MRD (engl. *Machine Readable Dictionaries*) descrie metodologii de a transforma MRD (formă electronică nestructurată sau text în accepţiunea noastră) în LDB (engl. *Lexical*

[3] Acesta obţinându-se la rândul său prin transformarea textului tipărit în imagine electronică cu recunoaşterea automată a caracterelor şi generarea textului electronic corespunzător celui tipărit.
[4] http://www.isi.edu/natural-language/dpp/
[5] http://humanities.uchicago.edu/orgs/ARTFL/forms_unrest/webster.form.html

DataBases – formatul electronic în care structura unei intrări de dicţionar este evidenţiată prin adnotări specifice).

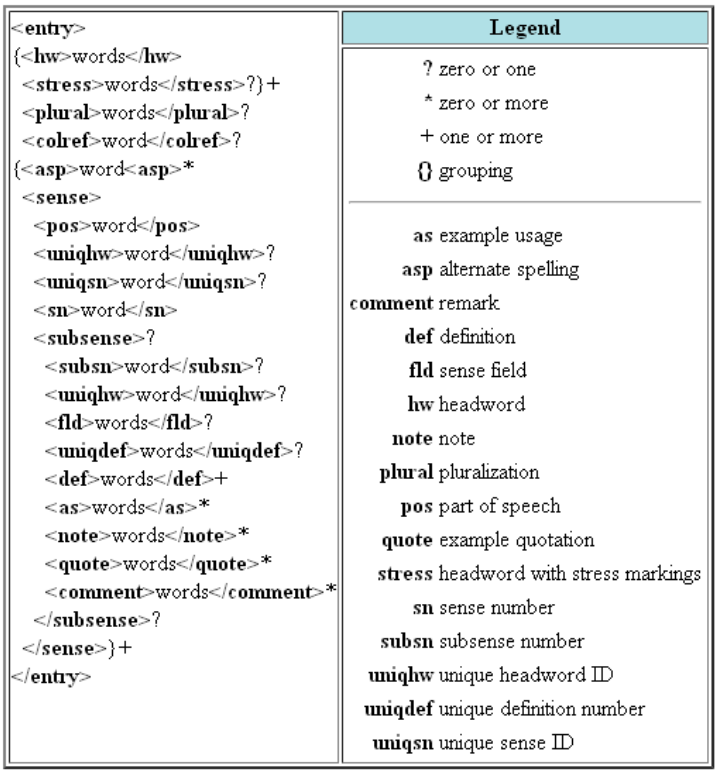

Figura 1: Structura unei intrări din dicţionarul Webster 1913[6].

Neff şi Boguraev (1989) disting două tipuri de sisteme de analiză structurală a intrărilor de dicţionar:

1. sistemele monolit în care regulile de analiză sunt conţinute în aplicaţie şi care, din acest motiv, nu pot fi adaptate să funcţioneze pe alte dicţionare;
2. sistemele bazate pe gramatici independente de context (GIC) în care sistemul este format dintr-un parser şi o gramatică. Gramatica are producţii cu care recunoaşte intrări de dicţionar, „marele avantaj" fiind acela că sistemul se poate adapta când avem de-a face cu un alt dicţionar, prin scrierea altei gramatici.

Bineînţeles că „scrierea unei alte gramatici" echivalează practic cu scrierea unui nou sistem monolit de analiză întrucât intrări din dicţionare diferite, diferă substanţial în termenii convenţiilor de alcătuire a unei intrări.

În ce priveşte tratamentul aplicat dicţionarelor în limba română, putem exemplifica transformarea Dicţionarului Explicativ al Limbii Române (DEX), (Tufiş et al., 1999) din format text în format bază de date XML prin utilizarea unei GIC special dezvoltată pentru DEX. Experienţa autorilor demonstrează faptul că adnotarea conformă cu TEI[7] nu a fost posibilă în cazul DEX fără a sacrifica din informaţia lexicală (TEI nu are

[6] Captură de imagine de la http://www.isi.edu/natural-language/dpp/
[7] Text Encoding Initiative, http://www.tei-c.org/index.xml

elemente care să descrie informaţia lexicală din DEX) şi/sau cea editorială (ordinea în care elementele unei intrări sunt date şi diversele notaţii care se pierd – devin redundante – prin adnotare). Acest lucru ne îndreptăţeşte să credem că nici DTLR nu va putea fi 100% reprezentat în această codificare.

3. Segmentarea unei intrări din DLR

Dacă ar fi să adoptăm sistemul de analiză structurală care foloseşte GIC, putem afirma că o metodă evidentă de a obţine gramatici mai simple (şi astfel, probabil, mai uşor de generalizat) ar fi *să simplificăm limbajul pe care gramatica trebuie să-l accepte*. Altfel spus, fiecare parser consumă un şir de simboluri care îi este furnizat la intrare pentru ca la final să raporteze dacă a acceptat şirul sau nu (dacă l-a acceptat, poate de asemenea să prezinte structura arborescentă a şirului de simboluri). Dacă simplificăm limbajul (reducem numărul de simboluri posibile) vom simplifica implicit gramatica care îl acceptă, rezultând firesc o structură mai simplă. *Ideea principală* este următoarea: structura simplificată ar trebuie să fie cât mai apropiată de structura logică pe care o putem vedea într-un dicţionar: un cuvânt are mai multe sensuri, fiecare sens are o definiţie şi exemple de utilizare, etc.

Pentru o intrare DLR în format text codificat UTF-8 (deci fără marcajele bold, italic, superscript, etc.), am imaginat această simplificare printr-o operaţie de segmentare: identificarea pasajelor de text continue care reprezintă unităţi atomice (pe care le vom numi în continuare „*tokeni*”) în alcătuirea unei intrări (de exemplu sursa unui citat, anul atestării documentare a unui citat, un identificator de sens, o parte de vorbire, cuvântul titlu, etc.). Pentru segmentarea unei intrări DLR, am utilizat o listă ordonată de colecţii de expresii regulate. Fiecare expresie regulată dintr-o colecţie este menită a identifica un tip de token care se realizează sub o anume formă în text. De exemplu, pentru a identifica anul atestării unui exemplu de utilizare (care este citatul) a sensului cuvântului titlu, textul ne poate oferi tokeni cum ar fi: „(cca. 1550)”, „(a. 1742)”, „(cca 1569–1575)”, etc. Când sunt întâlniţi, aceşti tokeni sunt recunoscuţi de diversele expresii regulate din colecţia dedicată acestui tip de token: anul atestării documentare.

Prima problemă cu un astfel de tip de abordare este că diversele expresii regulate dintr-o colecţie pot recunoaşte *tokeni care se suprapun*. Pentru a elimina acest inconvenient, fiecare expresie regulată din colecţie are asociată o prioritate (un număr natural) iar expresiile sunt „încercate” în ordinea crescătoare a acestor priorităţi. Vom prefera bineînţeles expresiile regulate care recunosc *tokeni cât mai lungi*. Alături de mecanismul priorităţilor, o a doua metodă, cea a verificării argumentelor, este folosită pentru a ne asigura că o anumită expresie regulată recunoaşte un token întreg şi nu unul parţial. Fiecare expresie regulată conţine o serie de capturi (secvenţe incluse între paranteze rotunde „(” şi „)” în cadrul expresiei regulate) pe care le numim „*atributele*” tokenului. De exemplu, tokenul „(cca. 1550)” are ca atribut anul în care s-a facut atestarea, anume 1550. Aceste atribute sunt verificate automat cu ajutorul unor liste de atribute posibile în momentul în care expresia regulată a recunoscut un token. Pentru exemplificare, fie două expresii regulate în Perl care recunosc tokeni de tipul surse de citat:

```
#Recunoaste o citare de tipul BELEA, P. A. 148
"author_2" => qr/((${RXAUTH}),\s*(${RXWORK})\s+(${RXNOINT}))/,
```

```
  "author_2_args" => { "_text" => 1, "author" => 2, "source" =>
3, "pages" => 4 },
  "author_2_check" => { "author" => \%DLRAUTHORS, "source" =>
\%DLRSOURCES },
  "author_2_rank" => 400
  #Recunoaste o citare de tipul MARCOVICI, D. 154/13
  "author_3" =>
qr/((${RXAUTH}),\s*(${RXWORK})\s+(${RXNOFRAC}))/,
  "author_3_args" => { "_text" => 1, "author" => 2, "source" =>
3, "pages" => 4 },
  "author_3_check" => { "author" => \%DLRAUTHORS, "source" =>
\%DLRSOURCES },
  "author_3_rank" => 200
```

Expresia regulată „author_2” recunoaşte tokenul „BELEA, P. A. 148” în care „BELEA” este autorul, „P.A.” este abrevierea lucrării iar „148” este numărul de pagină în lucrarea respectivă. Aceste 3 atribute sunt delimitate cu „()” în expresia regulată[8]. În momentul în care motorul de aplicare a expresiilor regulate a recunoscut tokenul „BELEA, P. A. 148” cu expresia „author_2”, se extrag atributele tokenului

author=”BELEA”, **source**=”P.A.”, **pages**=”148”, **_text**=”BELEA, P. A. 148”

prin inspecţia listei de atribute corespunzătoare (author_2_args) iar valorile atributelor author şi source se verifică căutându-se în listele DLRAUTHORS şi DLRSOURCES[9] (author_2_check). *Numai în cazul în care valorile au fost validate*, se acceptă tokenul şi se continuă procesul de recunoaştere. Recunoaşterea unui token înseamnă marcarea lui în textul intrării de dicţionar cu o notaţie de tip XML care specifică atât tokenul cât şi atributele sale. Pentru exemplul nostru, tokenul va fi adnotat ca

```
<AUTHCITE source="P.A." author="BELEA" pages="148">BELEA, P.
A. 148</AUTHCITE>
```

Deocamdată, segmentatorul nostru recunoaşte 6 tipuri de tokeni:

- cuvinte-titlu împreună cu terminaţii şi părţi de vorbire (DLRENTRY, colecţie cu 3 expresii regulate);
- atestare documentară (ATTESTED, 6 expresii regulate);
- trimiteri la sensurile altor cuvinte (ALSOSEE, 7 expresii regulate);
- citare cu autor (AUTHCITE, 14 expresii regulate);
- marcaje de sens (SENSE, 3 expresii regulate);

[8] $RXAUTH, $RXWORK şi $RXNOINT sunt variabile ale căror valori sunt alte expresii regulate care recunosc un nume de autor, un titlu de lucrare şi respectiv un număr de pagină. De exemplu, valoarea variabilei $RXNOINT este „qr/(?:[0-9]+)/”.

[9] La momentul scrierii acestor rânduri lista autorilor are aprox. 500 de nume iar cea de lucrări, cca. 400 de intrări.

- citare fără autor (SRCCITE, 19 expresii regulate).

În general în fiecare colecţie există câte o expresie regulată pentru fiecare formă a tipului de token întâlnită în practică. Evident, aceste colecţii nu sunt complete şi vor trebui îmbogăţite cu expresii regulate pentru fiecare formă necunoscută de token. Plătim astfel preţul unei gramatici de analiză a unei intrări de dicţionar mai simplă care altfel ar fi trebuit să conţină reguli de producţie pentru astfel de tokeni.

Pentru a exemplifica ieşirea segmentatorului cu adnotarea tokenilor de tipurile descrise mai sus, dăm începutul intrării REVĂRSÁT

```
<DLRENTRY suffix="-Ă" pos="adj." hword="REVĂRSÁT"
note="2">REVĂRSÁT2, -Ă adj.</DLRENTRY> <SENSE ind="Despre ape
curgătoare" subsense="1">1. (Despre ape curgătoare)</SENSE>

Care s-a vărsat peste margini, care a ieşit din albie; care a
inundat.

Cf. <ALSOSEE word="revărsa" subsense="1">revărsa (1)</ALSOSEE>.

Agiunsă de a trece apele, revărsate.

<AUTHCITE source="S. L." volume="II" author="ASACHI"
pages="19">ASACHI, S. L. II, 19</AUTHCITE>, ...
```

4. Concluzii

Despre evaluarea corectitudinii segmentării putem spune doar că exemplele de segmentare verificate de noi au fost corecte (când am întâlnit erori, am ajustat priorităţile de aplicare şi/sau am modificat expresiile astfel încât să obţinem rezultatele scontate). În momentul în care parserul va putea folosi această segmentare, vom putea da un procent de intrări analizate corect. Trebuie să spunem că acest algoritm de segmentare este destul de lent întrucât fiecare expresie regulată este încercată pe fiecare intrare de dicţionar. Timpul mediu de segmentare a unei intrări din DLR folosind cele 52 de expresii regulate existente acum este de aproximativ 1.1 secunde dar va creşte cu creşterea numărului de expresii regulate.

Acum dispunem de o listă de aproximativ 3800 de sigle bibliografice[10] şi de o listă cu toate abrevierile folosite în DLR, resurse care vor mări considerabil recall-ul segmentatorului. Următorul pas în dezvoltarea unui parser DLR este să ne concentrăm pe segmentarea completă a 100 de intrări extrase aleatoriu din DLR urmată de scrierea unei gramatici-nucleu care să accepte aceste intrări. Apoi, vom analiza DLR intrare cu intrare şi vom modifica segmentatorul/gramatica astfel încât noile intrări să fie acceptate. Metoda de analiză structurală alternativă dezvoltată de colegii noştri (Curteanu et al., 2008) va produce analize care vor fi de referinţă pentru parserul nostru. Analize identice pentru o aceeaşi intrare vor putea fi considerate corecte întrucât au fost generate independent de două metode diferite.

[10] Nu am ştiut de existenţa acestei liste când am lucrat la segmentator. A fost creată odată cu DLR de autorii dicţionarului. Mulţumim domnului Victor Celac pentru o copie a acestei liste.

Referinţe bibliografice

Cristea, D., Răschip, M., Forăscu, C., Haja, G., Florescu, C., Aldea, B., Dănilă, E. (2007). The Digital Form of the Thesaurus Dictionary of the Romanian Language. In C. Burileanu, H.N. Teodorescu (eds.), *Proceedings of the 4th International IEEE Conference SpeD 2007, "Advances in Spoken Language Technology"*, Iaşi, 10-12 Mai 2007. Editura Academiei Române, ISBN 978-973-27-1516-1.

Curteanu, N., Moruz, A.-M., Trandabăţ, D. (2008). Extracting Sense Trees from the Romanian Thesaurus by Sense Segmentation & Dependency Parsing. In *Proceedings of the COLING 2008 Workshop on Cognitive Aspects of the Lexicon (COGALEX 2008)*, pp. 55–63, Manchester, UK, 24 August, 2008. ISBN: 978-1-905593-56-9.

Neff, M.S., Boguraev, B.K. (1989). Dictionaries, Dictionary Grammars and Dictionary Entry Parsing. In *Proceedings of the 27rd Annual Conference of the Association for Computational Linguistics*, pp. 91-101.

Sala, M. (1996). Dicţionarul limbii române (DLR). In Dan Tufiş (ed.), *Limbaj şi Tehnologie*. Editura Academiei Române, Bucureşti, 1996. ISBN 973-27-0542-6. 270 p.

Tufiş, D., Rotariu, G., Barbu, A.-M. (1999). TEI-Encoding of a Core Explanatory Dictionary of Romanian. In Ferenc Kiefer, Gábor Kiss, and Júlia Pajzs (eds.), *Proceedings of the 5th International Workshop on Computational Lexicography (COMPLEX 1999)*, pp. 219-228, Pecs, Hungary, May 1999. Linguistics Institute, Hungarian Academy of Sciences.

Tufiş, D., Ion, R., Bozianu, L., Ceauşu, A., Ştefănescu, D. (2008). Romanian Wordnet: Current State, New Applications and Prospects. In Attila Tanacs, Dora Csendes, Veronika Vincze, Christiane Fellbaum, and Piek Vossen (eds.), *Proceedings of 4th Global WordNet Conference, GWC-2008*, pp. 441-452, Szeged, Hungary, January 2008. University of Szeged, Hungary. ISBN 978-963-482-854-9.

PLATFORMĂ PLURILINGVĂ DE FORMARE ŞI AUTOFORMARE ÎN DOMENIUL LIMBILOR ROMANICE

DOINA SPIŢĂ ŞI CLAUDIA BÎZDÎGĂ

Universitatea "Al.I.Cuza", Facultatea de Litere, Iaşi - România

doinaspita@yahoo.com, claudia_bizdiga@yahoo.es

Rezumat

"**GALAPRO**"[1] este un program transversal *Langues* "Formation de formateurs à l'intercompréhension en Langues Romanes", componentă a Programului pentru educaţie şi formare pe tot parcursul vieţii, Nº 135470-LLP-1-2007-1-PT-KA2-KA2MP, 2008-2010. El reuneşte universităţi din Portugalia – Universitatea din Aveiro fiind coordonatorul de proiect, Belgia, Franţa, Italia, Spania, precum şi Universitatea "Al.I.Cuza" din Iaşi. Proiectul îşi propune să dezvolte o reţea de formare specializată în domeniul cunoscut de mulţi ani în cercetarea lingvistică aplicată sub denumirea de *intercompréhension des langues romanes,* limbi între care, până în 2008, româna era cvasi ignorată ca referinţă. Obiectivul este acela de a crea şi de a experimenta, cu ajutorul unei platforme digitale, scenarii colaborative de formare centrate pe sarcini capabile să răspundă nevoilor şi aşteptărilor (în termeni de competenţe profesionale şi profil lingvistic şi comunicativ) diverselor categorii de public ţintă. Pe termen lung, proiectul vizează resurse conceptuale şi practice (glosare plurilingve, publicaţii, bază de date) transferabile către alte familii de limbi şi către alte discipline.

1. O soluţie de comunicare lingvistică pentru comunităţile plurilingve

Demersul propus de "Galapro" se înscrie în perspectiva plurilingvismului şi a didacticii acţionale, aşa cum sunt ele definite în documentele programatice ale Consiliului Europei în domeniul politicilor lingvistice. Libertatea de circulaţie şi mobilitatea pe piaţa muncii, armonizarea sistemelor de învăţământ şi intensificarea schimburilor economice, culturale şi ştiinţifice în Uniunea Europeană impun cu evidenţă cunoaşterea mai multor limbi ca pe o soluţie strategică prioritară pentru prezervarea diversităţii lingvistice şi culturale.

Patru orientări posibile sunt recomandate de către instituţiile europene reţelelor de cercetare care servesc această prioritate: difuzarea bunelor practici cu privire la învăţarea limbilor de către adulţi; inventarierea nevoilor actuale şi identificarea celor ale viitorului, legate de cooperarea europeană în domeniul învăţării limbilor de către adulţi, prin metode formale, nonformale şi informale; elaborarea de strategii care să acopere lipsurile existente în acest domeniu în care oferta nu mai corespunde exigenţelor actuale şi mai ales ale celor de perspectivă; în sfârşit, difuzarea programelor şi instrumentelor

[1] **Membrii echipei româneşti de cercetare**: Doina Spiţă – coordonator; Claudia Tărnăuceanu, Mihaela Lupu, Dana Nica, Maria Husarciuc, Paula Onofrei, Claudia Bîzdîgă.

permiţând formarea persoanelor implicate în învăţarea limbilor de către adulţi. Aşa cum vom vedea, programul "Galapro" răspunde tuturor acestor obiective specifice.

2. Istoricul cercetării

Primele proiecte vizând elaborarea unor metode de dezvoltare a competenţelor de înţelegere în scris în domeniul limbilor romanice au apărut la sfârşitul anilor '80, imediat după intrarea Spaniei şi Portugaliei în Uniunea Europeană. Reţinem trei dintre acestea: *EuRom4,* coordonat de Claire Blanche Benveniste, la Aix-en-Provence; *EuroComRom,* coordonat de Horst G. Klein, la Frankfurt; în fine *Galateea*, coordonat de Louise Dabène, la Grenoble, ale cărei cercetări au condus la *Galanet,* proiect Socrates-Lingua coordonat de Christian Degache, de la Universitatea Stendhal Grenoble 3, în anii 2001-2004. Nici unul dintre aceste proiecte nu a vizat în mod explicit limba română.

Platforma de comunicare plurilingvă a fost iniţiată în cadrul proiectului *Galanet*. Obiectivul era acela de a pune la dispoziţia vorbitorilor de portugheză, spaniolă, italiană şi franceză un instrument de formare la distanţă pe Internet, care să le ofere posibilitatea de a comunica. Originalitatea platformei consta în oportunitatea oferită vorbitorilor de diferite limbi romanice de a practica *intercomprehensiunea*, înţeleasă ca o formă de comunicare plurilingvă în care fiecare înţelege limbile vorbite de ceilalţi, dar se exprimă în limba / limbile romanice pe care el însuşi le cunoaşte, dezvoltând în acest fel competenţe de nivel diferit de cunoaştere a diverselor limbi. Pentru stimularea comunicării, s-a recurs la formularea unor sarcini comune de lucru, participanţii fiind puşi în situaţia de a interacţiona pentru elaborarea unui proiect colectiv. Intervenţia directă a "tutorilor", ca şi numeroasele "resurse" puse la dispoziţie pe platformă aveau rolul de a facilita comunicarea şi de a-i permite fluidizarea. Publicul ţintă vizat era constituit din studenţi în învăţământul superior sau în centre de limbi, liceeni şi adulţi cunoscători a cel puţin o limbă romanică de referinţă ca limbă maternă sau străină, fără însă a fi în mod necesar cunoscători, fie şi la nivel de debutant, al celorlalte trei limbi.

De la începutul anului 2008, partenerii s-au angajat într-un nou proiect (2008-2010, LLP KA2), numit *Galapro,* vizând de această dată formarea de formatori în domeniul intercomprehensiunii. Coordonatorul de proiect este Maria Helena de ARAUJO e SA, de la Universitatea din Aveiro[2]. Pentru a servi noii finalităţi, platforma este în curs de revizuire, cu atât mai mult cu cât două alte limbi romanice au fost invitate în echipă: româna şi catalana.

2.1. Ce este deci intercomprehensiunea?

Aşa cum puncta Jean-Pierre Chavagne de la Universitatea Lumière Lyon 2, parteneră în proiect, *intercomprehensiunea* înseamnă înţelegere încrucişată, înţelegere reciprocă, faptul că, în situaţie de dialog, fiecare se poate exprima în limba sa înţelegând-o, în acelaşi timp, pe a celorlalţi, ceea ce este mult mai avantajos, cel puţin în termeni de

[2] În perioada cuprinsă între 2 – 4 octombrie 2008, a avut loc la Iaşi a doua întâlnire a echipei internaţionale. Au participat reprezentanţii Universităţii din Aveiro – coordonatorul proiectului, Stendhal Grenoble 3, Lumière Lyon 2, ai Universităţii Autonome din Barcelona, Universităţii Complutense din Madrid, Universităţii din Cassino şi Universităţii din Mons-Hainaut - Departamentul de Tehnologie a Educaţiei.

randament lingvistic, decât a încerca să te exprimi într-o limbă care nu este a ta, cu riscul de a nu te face înţeles decât rudimentar. În sprijinul utilizării acestei strategii plurilingve de comunicare ce exploatează proximitatea lingvistică pot fi evocate şi alte argumente. Mai întâi, acela că este mai uşor şi mai rapid să înveţi să înţelegi o limbă decât să o vorbeşti. Apoi, conversaţia este mai echilibrată şi mai eficientă: cele două persoane sunt în poziţie de egalitate, fiecare se poate exprima cu un plus de claritate şi fineţe, căci se exprimă în limba pe care o cunoaşte cel mai bine. Situaţia creată prezintă, în acelaşi timp, o importantă dimensiune de convivialitate, interlocutorii apreciindu-şi reciproc efortul investit în a-l înţelege pe celălalt.

2.2. Platforma colaborativă – un concept spaţial

Vă invităm să o vizitaţi la adresa www.galanet.eu (Figura 1), iar noi vă vom fi ghizi.

Figura 1: Pagina principală a platformei

Disponibilă în toate limbile proiectului, platforma este un *concept în acelaşi timp spaţial şi temporal* – cum afirmă autoarele *Manualului de instrucţiuni*, pentru că ea presupune, pe de o parte, un scenariu cronologic – *sesiunile de formare* şi, pe de altă parte, un *spaţiu de învăţare virtual*, cu săli şi instrumente de exersare. Astfel:

Zona A este rezervată scenariului cronologic al sesiunilor. Este un scenariu conceput în patru faze (vezi cele patru butoane), corespunzătoare etapelor de derulare a unei sesiuni, respectiv unui anumit interval de timp şi unui forum. Concepută după acest model, formarea urmăreşte un proces gradual, care îi conduce pe participanţi spre sarcini de lucru din ce în ce mai complexe.

Zona B, numită şi „Ochiul”, îndeplineşte o dublă funcţiune: permite să ştii cine mai este conectat şi, dacă doreşti, să angajezi o comunicare tip *chat*.

Zona C este zona barelor de opţiuni: cea din dreapta sus permite alegerea limbii de lucru (catalana, franceza, italiana, portugheza, româna sau spaniola), accesul la mesagerie şi

la anunţurile de pe panoul de afişaj; cea din centru jos permite compunerea echipelor, schimbarea statutului de participare şi modificarea fazei sesiunii în curs de desfăşurare.

Zona D este spaţiul propriu-zis de lucru şi permite accesul la o serie de „instrumente" având funcţii bine precizate: unele servesc comunicării între participanţi (chatul, mesageria, forumul); altele ajută la arhivare (a chaturilor sau a ultimelor conexiuni); altele permit organizarea sesiunilor („Cine este cine?", „Profilul meu", „Profilul echipei mele", „Preferinţele mele"); în fine, altele au funcţia de „facilitator" al auto-formării (modulele şi resursele).

Platforma dispune de mai multe săli polivalente:

Forumul (Sala a) este spaţiul central al interacţiunilor, locul unde se desfăşoară scenariul pedagogic al sesiunii; aici aflăm date despre participanţi, despre profilul echipei, de aici se accesează mesageriile personale.

Biroul meu (Sala b) este locul de unde îmi pot trimite mesajele, unde pot fi cunoscut după „Profilul" şi „Preferinţele mele": limbi de referinţă, parolă, documentele din forum pe care doresc să le primesc.

Biroul echipei mele (Sala c) este spaţiul în care sunt propuse şi votate temele proiectelor comune („Alegerea temei") şi din care poţi avea acces la „Profilul" şi la „Chatul echipei".

Sala de redactare (Sala d) este locul unde se concepe şi se editează „Dosarul de presă", cu acces la chat pentru echipa redacţională.

Sala de reuniune (Sala e) este un *chat* rezervat întâlnirilor coordonatorilor diferitelor echipe.

Biblioteca (Sala f) este locul în care pot fi consultate fişierele.

În *Sala tehnică* (Sala g) descoperim „Profilul echipei tehnice".

Sălile pentru chat (Sala h) sunt trei saloane diferit colorate (albastru, galben şi roşu) în care participanţii care intervin în diverse echipe îşi pot da întâlnire pe chat. Aceste întâlniri sunt automat arhivate (cu excepţia celor private) în spaţiul denumit *Arhiva chaturilor* (Sala i). Pentru conversaţii care nu se doresc a fi arhivate, platforma ne invită în spaţiul numit *Bar* (Sala j). În fine, pentru a avea acces la compoziţia şi profilul diferitelor echipe poţi merge în Sala k („Cine este cine?"), pentru a te documenta poţi consulta spaţiul de „Resurse", unde vei găsi, de exemplu, compendii de gramatică şi de fonetică în limbile proiectului sau abordări comparative (Sala n), iar ca să fii mereu informat asupra actualităţilor din proiect, poţi consulta *Panoul de afişaj* din holul central.

2.3. Platforma colaborativă – un concept temporal

Aşa cum am menţionat mai sus, platforma „Galanet" prezintă oportunitatea de a-ţi permite accesul la ceea ce se numesc *sesiuni de formare*, construite pe bază de scenarii cronologice de activităţi care se succed în patru faze, controlate cu ajutorul celor patru butoane din Zona A. Astfel:

Faza 1 permite participanţilor să se cunoască, pentru ca în final să poată propune şi apoi alege, prin vot, o temă de lucru comună. Ea presupune desfăşurarea mai multor

activităţi, după cum urmează: înscrierea participanţilor – îi vom numi „stagiari" -, de către coordonatori, în diverse echipe (vezi bara de opţiuni de jos); elaborarea profilului echipei şi a fiecăruia dintre stagiari (are loc în spaţiul denumit *Biroul meu,* care cuprinde „profilul" şi „preferinţele mele"); cunoaşterea profilului celorlalţi („Cine este cine?"); pregătirea temelor de discuţie (chaturi, bar, forumuri); votarea temei (se desfăşoară în *Biroul echipei mele*); afişarea temei alese de către responsabilul de sesiune (în *Biroul echipei mele* şi în forum); desemnarea, de către responsabilul de sesiune, a echipei redacţionale (în cadrul forumului şi prin curier electronic).

Faza 2 permite schimburile de opinii între stagiari pe tema aleasă, via chaturi şi forumuri. Ea presupune următoarele activităţi succesive, monitorizate de către animatori: discuţii privind alegerea subtemelor şi fixarea acestora de către echipa redacţională, care crează şi forumurile de discuţii; redactarea sintezelor diferitelor discuţii desfăşurate pe subteme; redactarea liniei editoriale şi repartizarea sarcinilor între membrii echipei redacţionale (se realizează în sala de reuniune, bar, forumuri şi chaturi).

În *faza 3* au loc schimburile de opinii între stagiari şi colectarea documentelor, în funcţie de rubricile definite de către comitetul de redacţie. Ea se desfăşoară în trei timpi: adunarea şi redactarea documentelor de către stagiari, în funcţie de linia editorială hotărâtă (în spaţiul numit forum); cunoaşterea şi discutarea documentelor propuse de către celelalte echipe (chat, forum, bar) şi redactarea sintezelor de discuţii.

Faza a 4-a este rezervată pregătirii şi publicării „Dosarului de presă". Acesta se constituie din sintezele obţinute în finalul fazei precedente şi poate cuprinde extrase din schimburi plurilingve, precum şi documente ilustrative. Activităţile succesive vizează: sintetizarea forumurilor pe echipe (forumuri, bar şi chaturi); redactarea sintezei propriilor discuţii de către fiecare echipă (sala de redactare); discutarea sintezelor de către toţi participanţii (forumuri, bar şi chaturi); bilanţul final.

2.4. Dimensiunea colaborativă

Aşa cum sperăm că a reieşit din cele prezentate anterior, organizarea şi desfăşurarea unei sesiuni de formare este rezultatul unui efort de concertare între echipele diferitelor limbi prezente pe platformă şi a căror funcţionare, pentru o etapă de timp dată, se bazează pe acordul negociat şi consimţit de către participanţi. Aceştia pot avea statute diverse:

Responsabilul de sesiune este cel care deschide o nouă sesiune. El este cel care decide asupra diverselor responsabilităţi de coordonare, precum aceea de înscriere a stagiarilor şi de repartizare a lor în echipe, asupra datelor de deschidere şi de încheiere a diverselor faze, precum şi asupra termenului până la care se pot accepta cererile spontane de înscriere la o sesiune.

Coordonatorul local constituie echipele (una sau mai multe), înscrie stagiarii, poate invita animatori şi poate primi cererile de înscriere depuse pe pagina principală.

Animatorul este responsabil de dinamica grupului. El facilitează identificarea, de către stagiari, a celor mai bune strategii de autoformare şi răspunde la întrebări.

Stagiarul participă la toate activităţile propuse în sesiune, încearcă să citească în toate limbile în care participanţii intervin, îi interpelează pe ceilalţi stagiari şi pe animatori în legătură cu diverse probleme de formă sau de fond. El este dator să manifeste aceeaşi consideraţie pentru toate limbile de comunicare folosite pe platformă.

În fine, *Vizitatorul* poate să intre în forumuri şi în spaţiile de autoformare, poate să consulte diferitele arhive, dar nu poate interveni.

3. De la Galanet la Galapro

Obiectivul principal al proiectului *Galapro* este acela de a difuza şi valorifica achiziţiile *Galanetului*, exprimate în informaţii, strategii şi instrumente de formare deja experimentate, prin formarea unor agenţi educativi specializaţi în tehnicile de sensibilizare lingvistică prin *intercomprehensiune*.

Cercetările recente din domeniul didacticii limbilor străine abordează conceptul de *intercomprehensiune* fie în raport cu politicile de constituire a unei Europe unite şi coerente, fie în raport cu nevoile de comunicare ale diverselor comunităţi sociale sau profesionale. În această optică, modulele de formare pentru intercomprehensiune propuse de *Galapro* vor viza cu prioritate sensibilizarea mediilor educative faţă de nevoia pregnantă de dezvoltare a unor competenţe plurilingve şi pluriculturale. Principiile conducătoare vor fi acela de *diversificare* (reflectată în formularea sarcinilor de lucru şi activităţilor, metodologiei şi instrumentelor didactice) şi de *flexibilitate* (prin crearea unui cadru funcţional de dezvoltare autonomă, a unei gestionări curriculare adaptate, a conceperii programelor şi sistemului de evaluare în manieră suplă etc.). În acest spirit, se preconizează realizarea unei ample anchete vizând identificarea nevoilor şi aşteptărilor diferitelor tipuri de public ţintă (a se consulta pagina principală a sitului, pe care se regăsesc Chestionarele 1 şi 2, redactate în cele şase limbi ale proiectului).

Cui i se adresează *Galapro* ? Mai întâi, profesorilor de limbi, în formare iniţială, debutanţi sau experimentaţi. Apoi profesorilor de alte discipline, cum ar fi istorie şi geografie, arte sau turism, interesaţi în descoperirea şi dezvoltarea conceptului de Învăţământ Disciplinar Integrat printr-o Logică a Intercomprehensiunii (*Enseignement de Matières Intégré par une Logique d'Intercompréhension*). Apoi tutorilor şi animatorilor de formări la distanţă, în exerciţiu sau potenţiali, orientaţi către intercomprehensiune. În sfârşit, studenţilor, specialişti sau nu în domeniul lingvistic, precum masteranzii în ştiinţele limbajului, limbi şi culturi străine, psiho-pedagogie, ştiinţele educaţiei, ştiinţele comunicării.

Acestui public, *Galapro* îi propune participarea la o serie de sesiuni prototipice de formare de formatori în limbi romanice (catalana, franceza, italiana, portugheza, româna şi spaniola), concepute pe baza a două principii integrate:

- formarea în didactica intercomprehensiunii prin practicarea intercomprehensiunii;
- difuzarea intercomprehensiunii prin formarea de agenţi sau de viitori agenţi educativi, pregătiţi să acţioneze în contexte diverse.

Dincolo de competenţe didactice, *Galapro* ambiţionează diseminarea principiilor ce fundamentează însuşi conceptul de *intercomprehensiune*, adică *plurilingvism*,

diversificare şi *flexibilitate*, prin inducerea unui comportament lingvistic şi social adecvat nevoilor unui public specific, acela care are apetitul experienţelor de mobilitate geografică şi / sau virtuală.

Fiecare itinerar – sau sesiune – de formare (prevăzut a priori a se desfăşura pe o durată cuprinsă între 4 şi 15 săptămâni, în funcţie de particularităţile contextuale) va viza îndeplinirea uneia sau a mai multor sarcini finale, ale căror produse colaborative vor fi publicate pe site. Diversele sesiuni succesive vor contribui astfel la constituirea progresivă, prin capitalizare, a unui glosar plurilingv al principalelor concepte ce fundamentează *Galapro*. Această bază de resurse va fi deschisă „vizitatorilor", atât pentru raţiuni de perfectibilitate – prin înregistrarea reacţiilor, cât şi pentru difuzare.

Perspectiva partenerială va permite echipei internaţionale de proiect construirea, până la finele anului 2010, a unui scenariu de formare modular şi flexibil, adaptabil nevoilor specifice ale publicului propriu fiecărei sesiuni, un demers în acelaşi timp colectiv, colaborativ şi coerent.

4. Limba română în dispozitivul propus de Galapro

Devenind, în 2008, membră a unei echipe de proiect ce funcţionează de mai bine de zece ani, echipa românească este preocupată ca, într-o primă etapă, să recupereze, în plan personal, informaţia şi competenţele dezvoltate de Galanet, iar la nivel de echipă să contribuie la integrarea limbii române în diversele spaţii şi etape de lucru propuse de platformă. Vor fi create, după modelul activităţilor şi traseelor deja experimentate pentru celelalte limbi, module de auto-formare în limba română. Echipa românească va contribui apoi la efortul colaborativ de reconstruire şi readaptare a platformei faţă de exigenţe derivate din formularea noului obiectiv, formarea de formatori.

Principalele sarcini de lucru, în curs de realizare în perioada actuală, vizează traducerea întregii platforme în limba română, construirea modulelor de învăţare specifice şi constituirea bazei de documentare asupra acestei limbi neacoperită de câmpul de cercetare Galanet.

Un prim exerciţiu de iniţiere s-a desfăşurat în cadrul reuniunii de proiect pe care am găzduit-o la Iaşi, la începutul lunii octombrie. S-a vorbit atunci despre urmele modelului latin la nivel de lexic, morfologie, sintaxă, fonetică şi fonologie, s-a vorbit despre evoluţia istorică specifică limbii române, despre procesul de *reromanizare* sau *relatinizare* sau *occidentalizare* din secolele al XVIII-lea şi al XIX-lea, despre „miracolul" existenţei acestei „insule de latinitate în mijlocul unei mări slave". Evocarea rădăcinilor adânci ale latinităţii noastre, pe bază de exemple edificatoare, a provocat discuţii şi a determinat întărirea sentimentului de coeziune a grupului, prin conştientizarea, o dată în plus, a apartenenţei la o matcă comună.

Au urmat apoi două secvenţe interactive, concepute în spiritul pedagogiei intercomprehensiunii: prima secvenţă a fost propusă de echipa ieşeană, cealaltă de echipa universităţii din Barcelona. Aceasta din urmă a propus un model de didacticizare a unei înregistrări audio-video de română vorbită. Secvenţa poate fi vizionată accesând platforma conform traseului: pasul 1 – „Session en préparation"; pasul 2 – click pe una dintre sesiuni; pasul 3 – se merge la Salonul cu 16 scaune (spaţiul de autoformare);

pasul 4 – se ajunge la meniu (dreapta sus), de unde se optează pentru FR>RO, apoi se alege modulul pentru limba română.

Secvenţa interactivă a echipei româneşti a ales ca suport un citat din Lucian Blaga:

> *"După ce am descoperit că viaţa nu are nici un sens, nu ne rămâne altceva de făcut decât să-i dăm un sens".*

Participanţii, vorbitori de limbi romanice altele decât limba română, au fost invitaţi să reconstituie originea latină a termenilor româneşti şi să găsească echivalenţele în limbile lor respective. S-au obţinut rezultatele de mai jos:

după < lat. *de post* *după ce* + indicativ ~ lat. *postquam* + indicativ *ce* < lat. *quid* (pronume interogativ)	**RO:** după ce **ES:** después de **GA:** despois de **CAT:** després d' **PT:** depois de **FR:** après **IT:** dopo
a descoperi (IV) < lat. *disco(o)perio, -ire,-operui, -opertum* (IV) *am descoperit* (indicativ, perfect compus, pers. I pl.) < lat. *habemus* + participiul trecut, la pasiv	**RO:** am descoperit **ES:** haber descubierto **GA:** descubrir (*limba galiciană nu prezintă timpuri compuse*) **CAT:** haver descobert **PT:** haver descoberto **FR:** avoir découvert **IT:** avvere scoperto
că < lat. *quod*	**RO:** că **ES:** que **GA:** que **CAT:** que **PT:** que **FR:** que **IT:** che
viaţa < **vivitia* < *vivus, -a, -um* (< *viu* + *-eaţă)* *-a* (articol hotărât enclitic, feminin, singular, nominativ) < *illa* (pronume demonstrativ), feminin, singular	**RO:** viaţa (*articol enclitic*) **ES:** la vida *(articol hotărât proclitic în cazul celorlalte limbi)* **GA:** a vida **CAT:** la vida **PT:** a vida **FR:** la vie **IT:** la vita
nu < *non* *a avea (II)* < *habere (II)* *are – indicativ prezent, pers.a III-a, sg.* < *haberet* (conjunctiv imperfect) sau *habuerit* (conjunctiv perfect)	**RO:** nu are **ES:** no tiene **GA:** non ten **CAT:** no té **PT:** não tem **FR:** n'a pas **IT:** non ha
nici < lat. *neque* *un* < lat. *unus* (pronume nehotărât) → latina vulgară: valoare apropiată de cea a articolului din limbile romanice *nu … nici* (dubla negaţie) *vs.* lat. *duplex negatio est affirmatio*	**RO:** nici un **ES:** ningún **GA:** ningún **CAT:** cap **PT:** nenhum **FR:** n'a pas de = aucun **IT:** nessun

sens < fr. *sens* < lat. *sensus*	**RO:** sens **ES:** sentido **GA:** sentido **CAT:** sentit **PT:** sentido **FR:** sens **IT:** senso
ne (pronume personal, pers. I, pl., dativ) < *nă* < *nobis* (pronume personal, pers. I, pl., dativ)	**RO:** ne **ES:** nos **GA:** nos **CAT:** ens **PT:** nos **FR:** nous **IT:** ci
a rămâne (III) < *remaneo, -ēre* (II)	**RO:** rămâne **ES:** queda **GA:** queda **CAT:** queda **PT:** fica **FR:** reste **IT:** rimane
altceva < *alt (<alter)* + *ceva* (*ce* <*quid* + *va* < *vare* <*vare* <**voare* <*volet*) (variante pentru *va* < *vra* < *vrea*)	**RO: altceva** **ES: otra** cosa **GA: outra** cousa **CAT: una altra** cosa **PT: outra** coisa **FR: autre** chose **IT: altro**
a face (III) < *facere* (III) *de făcut* (supin) < prep. *de* + participiu trecut *făcut* < **facutus*, participiu trecut, forma pasivă, de la *facere*	**RO:** de făcut **ES:** por hacer **GA:** por facer **CAT:** per fer **PT:** por fazer **FR:** à faire **IT:** da fare
decât < *de* + *quantum*	**RO:** decât **ES:** que **GA:** que **CAT:** que **PT:** que **FR:** que **IT:** che
să < lat. *si*, **se* (it. *se)*; devine morfem al conjunctivului *a da* (I) < *dare* (I) *dăm* (indicativ prezent, pers. I pl.) < *damus*	**RO:** să(-i) dăm **ES:** dar(le) **GA:** dar(lle) **CAT:** donar(-li) **PT:** dar(-lhe) **FR:** (lui) donner **IT:** dar(le)
-i < *îi* (pronume personal, pers. a III-a, sg., dativ) < lat. *illi* (pronume demonstrativ, dativ)	**RO:** (să)-i (dăm) **ES:** (dar)le **GA:** (dar)lle **CAT:** (donar)-li **PT:** (dar)-lhe **FR:** lui (donner) **IT:** (dar)le

un < lat. *unus* (pronume nehotarât) → lat. vulgară: valoare apropiată de cea a articolului din limbile romanice *sens* < fr. *sens* < lat. *sensus*	**RO:** un sens **ES:** un sentido **GA:** un sentido **CAT:** un sentit **PT:** um sentido **FR:** un sens **IT:** un senso

RO: După ce am descoperit că viaţa nu are nici un sens, nu ne rămâne altceva de făcut decât să-i dăm un sens.
ES: Después de haber descubierto que la vida no tiene ningún sentido, no nos queda otra cosa por hacer que darle un sentido.
GA: Despois de descubrir que a vida non ten ningún sentido, non nos queda outra cousa por facer que darlle un sentido.
CAT: Després d'haver descobert que la vida no té cap sentit, no ens queda una altra cosa per fer que donar-li un sentit.
PT: Depois de haver descoberto que a vida não tem nenhum sentido, não nos fica outra coisa por fazer que dar-lhe um sentido.
FR: Après avoir découvert que la vie n'a aucun sens, il ne nous reste autre chose à faire que lui donner un sens.
IT: Dopo avvere scoperto che la vita non ha nessun senso, non ci rimane altro da fare che darle un senso.

Listă abrevieri: RO: română, ES: spaniolă, GA: galiciană, CAT: catalană, PT: portugheză, FR: franceză, IT: italiană

Referinţe bibliografice

Andrade, Ana Isabel, Maria Helena de Araujo e Sa, Covadonga Lopez Alonso, Silvia Melo, Arlette Séré. (2005). *Manuel d'Instructions,* Projecto Socrates Lingua 2.

Chavagne, Jean-Pierre. (2008). *L'intercompréhension en langue romanes* – la plate-forme Galanet (ppt).

Pagina oficială a platformei *Galapro*: www.galanet.eu

CONSIDERAŢII TEORETICE ASUPRA APLICABILITĂŢII UNEI BAZE DE DATE CU EXEMPLE DE TRADUCERE

NADIA LUIZA DINCĂ

Institutul de Cercetare pentru Inteligenţa Artificială

hnadia_luiza@hotmail.com

Rezumat

Una dintre regulile interne ale traducerii bazate pe exemple este dependenţa calităţii traducerii de lungimea şi modul de reprezentare a exemplelor de traducere. La rândul lor, acestea sunt gestionate de către o bază de exemple, în proiectarea căreia lingvistul este obligat să răspundă la două întrebări cheie:

- ce mod de reprezentare va alege pentru exemplul de traducere?

- care sunt posibilităţile de generalizare a exemplului de traducere stocat în baza de date?

În acest articol propun două posibile răspunsuri ale acestor întrebări, orientându-mă, pentru limbile română şi engleză, spre reprezentarea exemplelor de traducere ca arbori de dependenţă şi, respectiv, spre generalizarea lor prin informaţia semantică introdusă de clasele verbale descrise de Levin.

1. Introducere

În momentul în care a fost introdusă, (Nagao, 1984), traducerea bazată pe exemple era definită ca o traducere prin analogie, care utilizează o bază neadnotată de exemple, colectată, de regulă, dintr-un dicţionar bilingv. Echivalenţele erau exprimate sub forma perechilor de cuvinte, exceptând echivalenţele verbale formalizate prin cadre cazuale.

Ulterior, în categoria sistemelor structurale de traducere bazată pe exemple, se introduce reprezentarea exemplelor de traducere ca arbori de dependenţă cu legături explicite între subarbori (incluzând **nodurile frunză** care corespund unităţilor lexicale). Aceste legături permit folosirea **fragmentelor de exemplu** sau **subarborilor** pentru recunoaşterea corespondenţelor exacte cu segmente sau structuri ale intrării în limba sursă, şi pentru identificarea şi combinarea unităţilor de traducere echivalente în limba ţintă.

Sistemul de traducere automată MBT2, dezvoltat de S. Sato şi M. Nagao în 1990, utilizează arborii de dependenţă pentru reprezentarea exemplelor de traducere şi consideră trei operaţii de bază aplicabile subarborilor de dependenţă existenţi în baza de date:

a. operaţia de ştergere a unui subarbore;

b. operaţia de înlocuire a unui subarbore cu o expresie corespondentă intrării de traducere;

c. operaţia de adăugare a unei expresii corespondente intrării de traducere ca fiică pentru nodul rădăcină al unui subarbore.

Existenţa mai multor unităţi de traducere candidate şi generate prin backtracking solicită o selecţie, al cărei principal criteriu este mărimea unităţii de traducere: se preferă, astfel, o unitate de traducere mai mare, accepţia conceptului de mărime fiind numărul de noduri din unitatea de traducere. În principiu, sistemul MBT2 rezolvă compromisul dintre lungimea şi similaritatea corespondenţelor, considerat ca fiind baza creării de traduceri corecte. Problemele specifice acestei metode de lucru se referă la necesitatea unor calcule laborioase pentru determinarea scorului unităţii de traducere optimale şi la obligativitatea existenţei unui tezaur care să determine corect valorile de similitudine dintre cuvinte.

Sistemul de traducere automată bazată pe exemple propus de Kaji în 1992 se particularizează prin două subsisteme: învăţarea de modele de traducere şi, respectiv, traducerea bazată pe confruntarea modelelor cu datele.

Un model de traducere este o pereche de propoziţii bilingve, în care unităţile echivalente (cuvinte şi sintagme) sunt înlocuite prin variabile cărora le sunt asociate restricţii sintactice şi semantice. Se caută mai întâi acel model care unifică partea de limbă sursă cu o propoziţie de intrare, apoi se înlocuiesc cuvintele şi sintagmele prin variabilele modelului de traducere. Se face transferul pe limba ţintă, iar cuvintele şi sintagmele legate de variabile sunt traduse folosindu-se o metodă convenţională.

Procedura de învăţare a modelelor de traducere cuprinde etapa de antrenare a corpusului (pentru fiecare pereche de propoziţii se construiesc modelele de traducere potrivite) şi etapa de generare (se rafinează mulţimea de modele de traducere pentru a rezolva conflictele apărute în situaţii în care un model în limba sursă ar avea mai mulţi candidaţi de traducere).

Kaji exemplifică rafinarea prin două şabloane generate pentru exemplele de traducere „play baseball” şi „play the piano”. Cele două structuri generalizate sunt marcate prin categoriile semantice „sport” şi „instrument”, astfel: (1) play X [NP / sport]; (2) play X [NP / instrument].

Combinarea unei metode de lucru bazate pe exemple cu o analiză a textului sursă formalizată prin reguli gramaticale reprezintă alternativa propusă de O. Furuse şi H. Iida în 1992, respectiv 1994. Cunoştinţele de transfer se aplică asupra şirului de intrare, executându-se simultan parsarea structurală şi confruntarea şabloanelor cu datele.

Din perspectiva celor doi cercetători, un şablon elementar este o secvenţă formată din variabile şi simboluri pentru fixarea graniţelor constituenţilor lingvistici. Nu există simboluri gramaticale de tipul grupului nominal sau grupului verbal, ci un set de părţi de vorbire specificate potrivit rolului gramatical (substantivul comun, substantivul propriu, verbul-a fi, verbul auxiliar, etc.). Adiţional cuvintelor funcţionale, un şablon de traducere foloseşte şi bigramul parte de vorbire, exemplificat în expresia „I sing” prin construcţia echivalentă „pron-verb” : I sing -> şablon::= **I pron-verb sing**.

Agoritmul propus de Furuse şi Iida pentru identificarea şabloanelor de traducere cuprind, în esenţă, următorii paşi:

a. asignarea de informaţii morfologice fiecărui cuvânt din propoziţie;

b. introducerea marcatorilor de graniţă pentru constituenţii parsaţi morfologic;

c. derivarea structurilor posibile prin confruntarea şabloanelor cu datele, de la nivelul cel mai înalt de descriere lingvistică (propoziţia introductivă), până la nivelurile inferioare: propoziţia compusă, propoziţia simplă, sintagma verbală, sintagma nominală, cuvântul compus.

Dincolo de diferenţele dintre cele trei metode de reprezentare a exemplelor de traducere, se constată, ca invariantă, descompunerea propoziţiilor în constituenţi pentru a realiza o corespondenţă parţială în care părţile ce diferă se substituie prin variabile. Un avantaj al constituenţilor astfel obţinuţi îl constituie posibilitatea de prelucrarea a acestora în manieră independentă şi, în consecinţă, flexibilitatea traducerii.

În acest articol, proiectarea unei baze de exemple de traducere, pentru limbile română şi engleză, se realizează în maniera următoare:

- Exemplul de traducere este reprezentat prin arbori de dependenţă între care se stabilesc legături de corespondenţă. Sunt identificate, totodată, tipurile de relaţii sintactice de dependenţă dintre unităţile constituente ale unui grup verbal.

- În scopul generalizării, verbul primeşte o clasă semantică după tipologia creată de Levin. În situaţia în care există o breşă în găsirea unei corespondenţe între şirul de intrare şi subşirurile din baza de exemple, aceasta este rezolvată prin apelarea clasei semantice a verbului şi, implicit, a listei de verbe cu care verbul căutat contractează o relaţie sinonimică.

2. Reprezentarea exemplelor de traducere

2.1. Descrierea exemplului de traducere

Exemplul de traducere este un grup de cuvinte, uneori cu un înţeles diferit decât cel rezultat din însumarea înţelesurilor fiecărui cuvânt, căruia i se atribuie în limba ţintă o traducere şi un înţeles exact, favorizând o echivalenţă cu un nivel calitativ al traducerii ridicat.

Ca formă de reprezentare, un exemplu de traducere este compus din trei părţi:

- un arbore de dependenţă în limba sursă (în acest articol, limba sursă este limba română);

- un arbore de dependenţă în limba ţintă (limba engleză, în articol);

- legături de corespondenţă.

Cele trei părţi sunt evidenţiate în grupul verbal următor, extras din romanul lui G. Orwell, „1984”, subiect al unui amplu proiect lingvistic, **Multext-East**:

îşi imaginase orice ↔ had imagined everything
ro_e ([ro1, [imagina, v],
[ro1.1, [îşi, pron]],
[ro2, [orice, pron]]])

en_e ([en1, [have, aux],
[en1.2, [imagine, v],
[en2, [everything, pron]]]])
clinks ([[ro1, en1], [ro2, en2]])

Arborii de dependenţă ro_e şi en_e afişează, pe fiecare linie, după prefixul de limbă, un număr (ro 1-2), (en 1-2), acesta etichetând un **nod** din subarbore ce conţine forma bază a cuvântului şi categoria sintactică asociată. Drumurile, în fiecare din cei doi subarbori construiesc, pe fragmentele traductibile posibile, unităţi de traducere (ro1- ro1.1, ro1-ro2, en1-en1.1, en1-en1.1-en2).

Fluxul procesului de traducere pleacă de la un arbore de dependenţă în limba sursă, pe care îl descompune, stabileşte corespondenţele sursă, realizează transferul către corespondenţele ţintă, după care le combină pentru a obţine arborele de dependenţă echivalent în limba ţintă.

2.2. Tipurile de relaţii de dependenţă

Toate unităţile constituente dintr-un enunţ sunt aranjate de către vorbitor în construcţii bine formate, pe baza dependenţelor create între acestea: un cuvânt depinde de un altul prin poziţia sa lineară şi prin forma gramaticală.

Structura sintactică de suprafaţă, cea care interesează în lucrare, este un arbore ale cărui noduri sunt etichetate cu lexemele din propoziţie, iar **arcele**, denumite şi ramuri, primesc numele unei relaţii sintactice specifice, exemplificate mai jos.

Cele trei clase mari de dependenţe sintactice, şi anume: complementaritate, modificare şi coordonare, organizează, la rândul lor, un număr mare de relaţii sintactice, dispuse la nivelul grupului verbal astfel:

I. Relaţia de subordonare
a. obiect direct:
*(cumpărase – **ob-dir → cartea**) ↔ (bought – **ob-dir → [the] book**)*
*(luă – **ob-dir → [o] ţigară**) ↔ (took – **ob-dir → [a] cigarette**)*
*(**o ← ob-dir** – ura) ↔ (hated – **ob-dir → her**)*
b. obiect indirect în Dativ
*([să] spună – **ob-indir → i**) ↔ ([should] tell – **ob-indir → him**)*
c. obiect prepoziţional în Acuzativ
*([se simţea] atras– **ob-prep → de** [el]) ↔ ([felt] drawn– **ob-prep → to** [him])*
*(vorbea – **ob-prep → despre** [ea]) ↔ (referred – **ob-prep→ to** [it])*
d. obiect infinitival
*([le] putea – **ob-inf → vedea**) ↔ (could– **ob-inf → see**)*
II. Relaţia de coordonare
*scoase – **ob-dir→** [un] **toc** –**coord→** [o] **sticlă** [de cerneală] –**coord→ şi**[un volum] ↔ took down – **ob-dir→** [a] **penholder** –**coord→** [a] **bottle** [of ink] –**coord→ and** [a book]*

2.3. Arborii de dependenţă şi legăturile de corespondenţă

În crearea arborilor de dependenţă pentru grupurile verbale în limbile română şi engleză am urmărit trei criterii de existenţă a unei relaţii sintactice de dependenţă între două cuvinte dintr-o propoziţie:

- criteriul de conectivitate dintre două forme lexicale;

- criteriul de dominanţă între două cuvinte;

- criteriul tipului specific de dependenţă sintactică dintre două lexeme.

Criteriile sunt dependente de limbă, ceea ce face uneori dificilă stabilirea unei corespondenţe. Este cazul pronumelor reflexive din limba română, de exemplu, nerealizate, în planul enunţului, în limba engleză:

îşi turnă o ceaşcă de ceai ↔ poured out a teacupful
ro_e ([ro1, [turna, v],
[ro2, ob-indir, [îşi, pron]],
[ro3, ob-dir,
[ro3.1, [ceaşcă, n],
[ro3.2, [o, art]],
[ro3.3, [de, prep],
[ro3.4, [ceai, n]]]]])
en_e ([en1,
[en1.1., [pour, v],
[en1.2, jonctiv, [out, prep]],
[en2, ob-dir,
[en2.1, [teacupful, n],
[en2.2, [a, art]]]]])
clinks ([[ro1,en1], [ro3, en2]])

Tipul de dependenţă specifică obiectului indirect este preluat, în limba engleză, de către subiect, ca agent al acţiunii descrise de verb. În acelaşi timp, criteriul de conectivitate pentru limba română admite o abatere de la regula generală a dispunerii lineare a cuvintelor, de aceea verbul, ca nucleu al sintagmei sintactice, va impune pronumelui relaţia de obiect indirect. Nodul 3 din limba română, extins în nodurile: 3.1.- „ceaşcă”, 3.2.- „o”, 3.3.- „de”, 3.4.- „ceai”, are legături de corespondenţă cu nodul 2 din limba engleză, dezvoltat într-un nod părinte şi un altul fiică.

În exemplul de traducere următor, pronumele reflexiv din română este realizat în engleză prin forma pronumelui personal, o explicaţie găsindu-se în structura *dativ+verb+substantiv*, unde formele neaccentuate de dativ, ale pronumelui reflexiv sau personal, exprimă ideea de posesie:

îşi întinsese braţele către ecran ↔ extended her arms towards the screen

ro_e, ([ro1, [întinde, v],
[ro2, posesie, [îşi, pron]],
[ro3, ob-dir, [braţe, n]],
[ro4, direcţie,
[ro4.1, [către, prep],
[ro4.2, [ecran, n]]]]])
en_e ([en1, [extend, v],
[en2, ob-dir,
[en2.1., posesie, [her, pron]],
[en2.2., [arms, n]]],
[en3, direcţie,
[en3.1, [towards, prep],
[en3.2, [screen, n],
[en3.3, [the, art]]]]])
clinks ([[ro1, en1], [ro3, en2.2],[ro4, en3]])

Nodul 2 din arborele în limba engleză cumulează două relaţii de dependenţă de rang diferit. Cea dominantă este de obiect direct, aplicată numelui comun, care impune o relaţie de posesie pronumelui personal în genitiv, întreaga structură construind, împreună cu verbul, unitatea traductibilă "extended her arms". Subarborele în engleză respectă criteriul de aranjare lineară a formelor lexicale în scopul stabilirii legăturii de dependenţă sintactică. În schimb, unitatea traductibilă corespondentă în română "îşi întinsese braţele" consideră criteriul de dominanţă sintactică pentru a identifica orientarea relaţiei de dependenţă.

Generalizând structurile în care dispunerea lineară a cuvintelor cedează în faţa dominanţei sintactice, se observă două principale relaţii de dependenţă pe care verbul tranzitiv le impune formelor lexicale predecesoare: obiectul direct, respectiv obiectul indirect.

O altă situaţie gramaticală cu un regim aparte de echivalenţă între limbile română şi engleză o constituie anticiparea sau reluarea obiectului direct sau indirect prin forme pronominale personale. În exemplul de traducere următor, anticiparea se realizează prin pronumele personal neaccentuat "îl", fără un corespondent lexical în limba engleză:

îl văzuse pe O'Brien ↔ had seen O'Brien
ro_e ([ro1, [vedea, v],
[ro2, ob-dir,
[ro2.1., binar, [îl, pron]],
[ro2.2., [pe, prep],
[ro2.3., [O'Brien, n]]]])
en_e ([en1, [had, aux]],
[en2, [see, v],

[en3, [O'Brien, n]]])
clinks ([[ro1,en2],[ro2, en3])

Introducerea relaţiei de binaritate pentru anticiparea obiectului direct este de natură să rezolve problema satisfacerii valenţei verbale, deoarece verbul tranzitiv „a vedea" nu permite două complemente directe. Pronumele „îl" şi constructul „pe O'Brien", cu dominanţă pe lexemul prepoziţional, au acelaşi rang sintactic şi contribuie la complinirea verbului. În limba engleză, dependenţa binară nu mai există, verbul dominând un obiect direct realizat nominal.

3. Generalizarea exemplelor de traducere

3.1. Preliminarii

Una dintre principalele probleme pe care trebuie să le înfrunte traducerea bazată pe exemple este necesitatea de a folosi un exemplu de traducere pentru mai mult de o situaţie de intrare. În mod obişnuit, primul pas îl constituie identificarea unei corespondenţe între şirul lexical de intrare sau subşiruri ale acestuia şi exemplele de traducere stocate în baza de date. Aceasta poate cauza uneori frustrări în ce priveşte calitatea traducerii, deoarece baza de exemple, oricât de complexă ar ajunge la un moment dat, nu reuşeşte să acopere flexibilitatea lingvistică.

O posibilă soluţie pentru acest inconvenient rezidă în combinarea relaţiilor semantice şi structurilor sintactice, astfel încât un lexem din exemplul de traducere să deschidă posibile instanţieri pentru lexemele din seriile semantice, urmând generarea relaţiei de sinonimie.

Fie următoarea structură sintactică de tradus:

Ceruse libertatea cuvântului.

După deflexionare şi dezambiguizare, etape ce nu formează subiectul propriu-zis al acestei lucrări, algoritmul trebuie să treacă la căutarea corespondenţelor. În baza de exemple însă, verbul „a cere" nu intră în nicio combinaţie cu grupul nominal „libertatea cuvântului". În schimb, structura nominală este identificată ca fiind în relaţie de dependenţă faţă de un alt verb, „a solicita", în construcţia sintactică: „solicitase libertatea cuvântului", cu echivalent de traducere: „was advocating freedom of speech". De aceea, la pasul următor se verifică existenţa unei relaţii semantice între verbele „a cere" şi „a solicita". Se identifică astfel clasa semantică a *verbelor de transfer al unui mesaj*[1], clasă ce instanţiază, pentru limba română, lexemul verbal solicita:5, iar pentru limba engleză preach:2, advocate:2. Se opreşte căutarea şi se validează corespondenţa dintre „ceruse libertatea cuvântului", respectiv „solicitase libertatea cuvântului".

Trebuie precizat faptul că relaţiile semantice pentru limba română sunt identificate cu ajutorul dicţionarului de sinonime, iar pentru limba engleză prin intermediul ontologiei lexicale WordNet, interogată cu ajutorul editorului Visdic, versiunea 1.3.50.

[1] În taxonomia descrisă de Beth Levin, este vorba de clasa 37.1- *Verbs of Transfer of a Message*.

3.2. Rolul introducerii relaţiilor semantice la nivelul sintagmei verbale

Există o relaţie profundă între proprietăţile semantice ale unui verb şi cele sintactice, astfel explicându-se de ce, dezambiguizând înţelesul unui lexem, vorbitorii completează modelul comunicaţional intuindu-i tiparul sintactic.

Aceasta este, în esenţă, şi motivaţia introducerii relaţiilor semantice la nivelul sintagmei verbale. Verbul ce guvernează o relaţie de dependenţă nu este izolat în mulţimea tuturor verbelor, ci este actant al unei relaţii de sinonimie cu alte lexeme verbale. Nu toate sensurile verbelor participă la crearea sinsetului, ci doar acelea care sunt ordonate în jurul unui înţeles comun.

Precizarea tipurilor de relaţii de dependenţă sintactică şi crearea sinseturilor grupate în clase verbale sunt de natură să stimuleze calitatea traducerii prin mai buna adaptare la flexibilitatea limbii şi la condiţiile de bună formare a unei propoziţii.

În exemplul de traducere următor sunt puse în evidenţă proprietăţile sintactice şi semantice ale centrului verbal:

dădea o muzică stridentă, militărească. ↔ had played a strident military music
ro_e ([ro1, [Verbe de Reprezentaţie-> da:9, transmite:13], [da, v],
[ro2, ob-dir,
[ro2.1, [muzică, n],
[ro2.2, [o, art]],
[ro2.3, [stridentă, adj]],
[ro2.4, [militărească, adj]]]]])
en_e ([en1, [have, aux],
[en1.1, [Performance Verbs -> play:7, perform:3], [play, v],
[en2, ob-dir,
[en2.1, [music, n],
[en2.2, [a, art]],
[en2.3, [strident, adj]],
[en2.4, [military, adj]]]]]])
clinks([[ro1, en1], [ro2, en2]])

Crearea arborilor de dependenţă se realizează, după cum se observă mai sus, prin descrierea, pentru verbul principal a clasei semantice, a sinsetului asociat şi a tipurilor de dependenţă guvernate de verb. Se generalizează astfel posibilităţile de corespondenţă între şirul de intrare şi exemplele din baza de date, dar se impune, totodată, un filtru de validare a acestora din perspectiva respectării tipurilor de relaţii de dependenţă. Dintr-o mulţime de candidaţi potenţiali la stabilirea corespondenţei, sunt selectaţi doar cei care domină aceleaşi tipuri de dependenţă sintactică. Împreună, cele două descrieri - sintactică şi respectiv, semantică- au rol în dezambiguizarea traducerii.

4. Concluzii

Acest articol prezintă câteva consideraţii teoretice despre aplicabilitatea unei baze de exemple de traducere. În dezvoltarea lui, am plecat de la premisa relaţiilor sintactico-semantice dintre cuvinte urmărind două idei esenţiale, şi anume utilitate şi generalizare. Avem, astfel, pe de o parte, relaţiile de dependenţă sintactică dintre verb şi celelalte valori morfo-sintactice dominate de el, iar pe de altă parte, relaţia de sinonimie dintre verb şi lexemele din acelaşi sinset, respectiv clasă verbală.

La prima vedere, o bază de exemple de traducere încărcată cu toate aceste informaţii poate fi dificil de manipulat, prin mărimea numărului de căutări. În vederea diminuarii acestui inconvenient propun trei criterii de selecţie a exemplelor pentru a construi o bază de date. În primul rând trebuie selectate expresiile şi structurile considerate a fi cele mai frecvente în limbile sursă şi ţintă. Odată constituit acest nucleu, urmează, cu prioritate de rang doi, secvenţele propoziţionale al căror înţeles este diferit de compunerea sensurilor unităţilor constituente. Cu prioritate de rang 3 pentru completarea bazei de date sunt grupurile verbale extrase din corpusul „1984”.

În condiţiile asigurării unei bune acoperiri lexicale şi a unor structuri ordonate semantic, respectiv sintactic, se poate considera că şi numărul de căutări în baza de exemple este mai mic. În acelaşi timp, creşte posibilitatea identificării printre lexemele verbale cele mai frecvente, a acelora care sunt în relaţie de sinonimie cu verbul de intrare.

Un alt avantaj al proiectării bazei de exemple prin relaţii de sinonimie, filtrate de relaţii de dependenţă sintactică, se găseşte în dezambiguizare, semantică şi sintactică. Există verbe care au, fără îndoială, mai multe sensuri, unele dintre ele particularizând relaţii de dependenţă diferite. În momentul în care sinonimia dintre sensurile a două verbe este evaluată ca având aceleaşi tipuri de dependenţă sintactică, se reţine un anume sens din mulţimea de candidaţi ai verbului de intrare. Operaţia este validă şi pentru argumentele selectate de verb: dacă verbele aparţin unui sinset, iar unul are o linie sintagmatică deja cunoscută, celălalt îi imită comportamentul sintactic.

Referinţe bibliografice

Levin, B. (1993). *English Verb Classes and Alternations- A Preliminary Investigation*, The University of Chicago Press.

Mel’čuk, I. (2003). Levels of Dependency in Linguistic Description: Concepts and Problems. In V. Agel, L. Eichinnger, H.-W. Eroms, P. Hellwig, H. J. Herringer, H. Lobin (eds): *Dependency and Valency. An International Handbook of Contemporary Research*, vol. 1, Berlin - New York, W. de Gruyter, 188-229.

Multext-East Home Page: http://nl.ijs.si/ME/

Nagao, M. (1984), A framework of a mechanical translation between Japanese and English by analogy principle, *Proceedings of the international NATO symposium on Artificial and human intelligence*, Lyon, France, 173 – 180.

Furuse, O, Iida, H. (1992), Cooperation between transfer and analysis in example-based framework. *Proceedings of the 14th International Conference on Computational Linguistics,* Nantes, France. ACL*,* 645-651.

Furuse, O., Iida, H. (1994), Constituent boundary parsing for example-based machine translation, *Proceedings of the 15th conference on Computational linguistics*, vol. 1, Kyoto, Japan, 105-111.

Kaji H., Kida, Y., Morimoto, Y. (1992), Learning translation templates from bilingual text, *Proceedings of the Fourteenth International Conference on Computational Linguistics*, Nantes, 672-678,.

Sato, S., Nagao, M. (1990), Towards memory-based translation. *Proceedings of COLING-90*, Helsinki, Finland, vol. 3, 247-252.

Seche, M., Seche, L. (2002), *Dicţionar de sinonime*, Editura Litera Internaţional

Visdic: http://nlp.fi.muni.cz/projekty/visdic/

CAPITOLUL 3

APLICAŢII ALE TEHNOLOGIILOR LINGVISTICE

EVALUAREA RĂSPUNSURILOR OFERITE DE UN SISTEM DE TIP ÎNTREBARE RĂSPUNS PENTRU LIMBA ROMÂNĂ

ADRIAN IFTENE, ANCUŢA ROTARU, DANA-ALINA MARCU

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

{adiftene, ancuta.rotaru, dana.marcu}@info.uaic.ro

Rezumat

În cadrul competiţiei QA@CLEF2008[1] una din principalele provocări a fost exerciţiul de validare a răspunsurilor AVE (Answer Validation Exercise). Lansat cu 3 ani în urmă acesta are ca scop evaluarea răspunsurilor oferite de către un sistem de tip întrebare răspuns, dorind astfel să mărească şi calitatea acestora. Anul acesta, ca şi în ediţia din 2007, printre cele 5 limbi participante a fost prezentă şi limba română, iar noi am participat pentru prima dată cu un sistem dedicat acesteia.

Articolul de faţă prezintă structura exerciţiilor, principalele componente ale sistemului construit de noi pentru competiţia din acest an, precum şi rezultatele acestei ediţii. Important de remarcat este faptul că pe 3 limbi participante (engleză, germană şi română) sistemele AVE au obţinut rezultate mai bune decât sistemele de tip Întrebare-Răspuns, în ordonarea răspunsurilor oferite de un sistem de tip Întrebare-Răspuns.

1. Introducere

AVE[2] a avut loc pentru prima oară în 2006 (Peñas et al., 2007) din nevoia de a promova dezvoltarea şi evaluarea sub-sistemelor care aveau ca scop validarea corectitudinii răspunsurilor oferite de sistemele de tip Întrebare-Răspuns (ÎR). Din start s-a dorit ca AVE să îmbunătăţească si calitatea sistemelor de tip ÎR, dar în primul rând să verifice dacă răspunsurile alese corespund sau nu fragmentelor de texte ajutătoare existente.

De la an la an metodologia de evaluare s-a modificat în încercarea de a surprinde cât mai bine factorii care ar duce efectiv la îmbunătăţirile sistemelor de tip ÎR. Astfel, în 2007 sistemele trebuiau să selecteze doar un singur răspuns valid pentru fiecare întrebare dintr-o mulţime de răspunsuri posibile, spre deosebire de ediţia din 2006 când era posibil să se aleagă mai multe răspunsuri valide. În 2008, s-a observat că această metodologie are o problemă: nu se ştia cum se vor comporta sistemele în cazul în care toate răspunsurile posibile ar fi fost incorecte. Se dorea ca ele să poată cere alte răspunsuri de la sistemele de tip ÎR, în speranţa că vor putea obţine măcar un răspuns corect. Ediţia de anul acesta a avut ca obiectiv eliminarea acestor neajunsuri descoperite în ediţiile precedente.

În continuare vom prezenta caracteristicile competiţiei de anul acesta şi modul în care am construit sistemul folosit de noi pe limba română. În partea de final vom prezenta rezultatele şi concluziile.

[1] CLEF: http://www.clef-campaign.org/2008.html
[2] AVE: http://nlp.uned.es/clef-qa/ave/

2. Descrierea Exerciţiului

2.1. Formatul datelor de intrare

Urmărind tiparul propus în ediţia din 2007, în ediţia din 2008 (Rodrigo et al., 2008) sistemele trebuiau să ia în considerare triplete de forma (*Întrebare*, *Răspuns*, *Fragment de Text*) şi să hotărască dacă răspunsul la întrebare este corect şi poate fi dedus din fragmentul de text ataşat. Astfel, pentru fiecare tripletă de această formă, participanţii trebuie să stabilească o valoare care are ca semnificaţie faptul că tripletul este validat sau respins. Pentru limba română, fişierul de intrare conţine 119 întrebări, iar pentru fiecare întrebare sunt între 1 şi 9 răspunsuri posibile, în total fiind 497 de triplete.

Tabel 1: Formatul datelor de intrare

```
<q id="1" lang="RO">
  <q_str>Câte zile avea aprilie înainte de 700 î.Hr.?</q_str>
  <a id="0001_1" value="">
    <a_str>30</a_str>
    <t_str doc="Aprilie.html">
     Înainte de anul 700 î.Hr., luna aprilie era a doua lună a anului în calendarul roman şi
     avea 29 de zile. După ce Iuliu Cezar a introdus calendarul iulian în 45 î.Hr., luna aprilie
     avea 30 de zile şi devenea a patra lună a anului.
    </t_str>
  </a>
  <a id="0001_6" value="">
    <a_str>29 de zile</a_str>
    <t_str doc="1">
     Hr., luna aprilie era a doua lună a anului în calendarul roman şi avea 29 de zile. După ce
     Iuliu Cezar a introdus calendarul iulian în 45 î .
    </t_str>
  </a>
  <a id="0001_7" value="">
    <a_str>de anul 700</a_str>
    <t_str doc="1">
     Numele lunii aprilie (latină: Aprilis ) vine de la cuvântul latinesc aperio, ire = a
     deschide, deoarece în aprilie se deschid mugurii plantelor. Înainte de anul 700 î .
    </t_str>
  </a>
  <a id="0001_8" value="">
    <a_str>cu aceeaşi zi a săptămânii în toţi anii</a_str>
    <t_str doc="1">
     Aprilie începe cu aceeaşi zi a săptămânii ca şi Iulie în toţi anii şi ca Ianuarie în anii
     bisecţi.
    </t_str>
  </a></q>
```

În tabelul 1 putem vedea formatul datelor de intrare (unde tag-ul "**q_str**" conţine întrebarea, tag-urile "**a**" corespund fiecărui răspuns posibil, acesta fiind propriu-zis conţinut în tag-ul "**a_str**", iar fragmentele de text apar în tag-ul "**t_str**").

2.2. Formatul datelor de ieşire

Răspunsurile oferite de participanţi trebuie să fie în următorul format:

q_id a_id [VALIDAT| SELECTAT|RESPINS] scor_de_încredere

unde semnificaţia răspunsurilor este următoarea:

- **VALIDAT**: indică faptul că răspunsul este corect şi este suportat de paragraful de text asociat. Nu există nici o restricţie asupra numărului de răspunsuri validate (pot fi toate validate sau nici unul).
- **SELECTAT**: indică faptul că răspunsul este VALIDAT şi reprezintă cel mai probabil răspuns al unui posibil sistem de tip ÎR. Fiecare întrebare va avea doar un singur răspuns selectat. Cel puţin unul dintre răspunsurile valide trebuie să fie selectat.
- **RESPINS**: indică faptul că răspunsul este incorect (sau că nu există suficiente dovezi care să-i demonstreze corectitudinea). Nu există nici o restricţie asupra numărului de răspunsuri respinse (pot fi toate sau nici unul).
- **scor_de_încredere**: Opţional, pentru fiecare tripletă se poate acorda un scor de încredere (care poate lua valori din intervalul [0, 1]): unde 0 – reprezintă faptul că suntem nesiguri de răspunsul dat, iar 1 – reprezintă faptul că suntem siguri de răspunsul oferit.

2.3. Proveniența datelor de intrare

Ca şi în ediţiile precedente ale competiţiei AVE, datele folosite în antrenarea şi testarea sistemelor de apreciere a răspunsurilor provin din fişierele cu evaluarea sistemelor de tip ÎR folosite în competiţia QA@CLEF, cu unele mici completări şi modificări.

Transformarea evaluării răspunsurilor sistemelor de tip ÎR în date de test pentru competiţia AVE2008 s-a făcut în modul următor (Rodrigo et al., 2008):

- un răspuns care a fost evaluat CORECT în competiţia sistemelor de tip ÎR va fi evaluat ca fiind VALID în datele de test AVE;
- un răspuns evaluat ca fiind GREŞIT sau NESUPORTAT în QA@CLEF va fi considerat RESPINS în AVE;
- un răspuns evaluat ca fiind INEXACT sau NEEVALUAT în QA@CLEF va avea valoarea NECUNOSCUT în AVE (şi nu va fi considerat în evaluarea sistemelor).

Deoarece colecţia datelor de test pentru competiţia AVE s-a construit pe baza tuturor fişierelor trimise de participanţii pe o anumită limbă, a fost nevoie să se stabilească nişte reguli pentru îmbunătăţirea calităţii acestora:

- eliminarea răspunsurilor redundante;

- dacă între răspunsurile posibile pentru o întrebare există răspunsuri care se conţin unele pe altele se va recurge la următoarea abordare: se vor elimina răspunsurile care au lungimea cea mai scurtă;
- eliminarea întrebărilor fără răspuns (care au răspuns nul).

Spre deosebire de competiţia sistemelor de tip întrebare-răspuns, unde întrebările au fost grupate pe domenii, întrebările nu au fost grupate în acelaşi mod în competiţia AVE.

Testarea colecţiilor folosite s-a făcut în 9 runde de test, pentru fiecare limbă fiind generat un test individual. Pentru limba română sistemul creat a avut următoarele rezultate pentru 119 întrebări şi 497 de răspunsuri:

- 48,58 % de răspunsuri au fost evaluate ca fiind VALIDE din mulţimea de răspunsuri posibile;
- 52 de răspunsuri SELECTATE, 406 de răspunsuri RESPINSE şi 39 de răspunsuri NECUNOSCUTE.

3. Sistemul pentru limba română construit pentru competiţia AVE

Structura sistemului construit pentru limba română este asemănătoare cu structura sistemului construit pe limba engleză (Iftene, Balahur-Dobrescu, 2008). Diferenţele faţă de acesta sunt legate de modulele şi resursele folosite. Sistemul primeşte la intrare triplete de forma (*întrebare*, *răspuns*, *fragment de text*) şi oferă la ieşire evaluarea fiecărui răspuns în parte (Vezi figura 1 de mai jos).

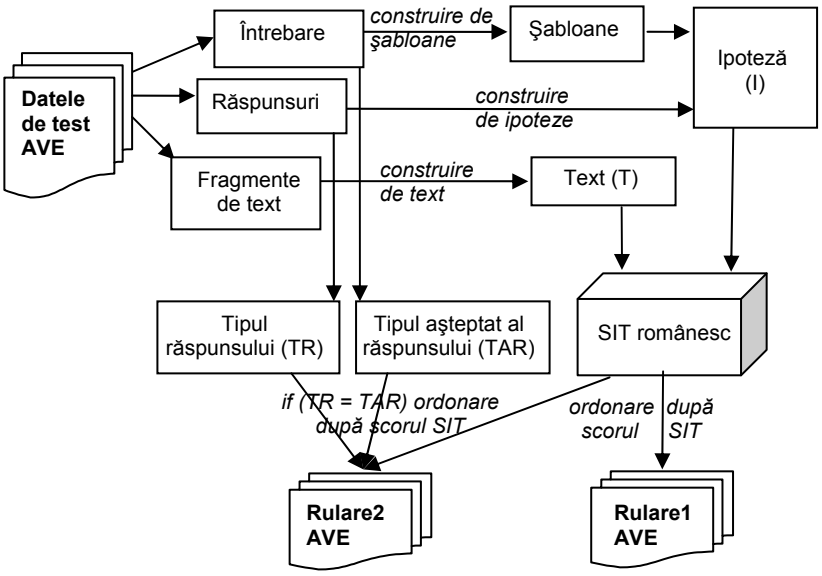

Figura 1: Sistemul AVE românesc

Principalele componente ale acestui sistem realizează următorii paşi:

- Construiesc *ipoteze* necesare *sistemului de inferenţe textuale* (SIT) folosind şabloanele construite din întrebări şi răspunsurile din datele de test AVE.
- Consideră fragmentele de text ca fiind *textul* necesar unui SIT.

- Calculează folosind SIT-ul românesc un scor de potrivire pentru fiecare pereche (*Text*, *Ipoteză*).
- Prelucrează întrebarea cu tehnici specifice sistemelor de tip întrebare-răspuns şi identifică *tipul aşteptat al răspunsului* (TAR).
- Aplică şabloane specifice şi resurse de entităţi de tip nume pentru a identifica *tipul răspunsului* (TR).
- În final, pe baza scorului de potrivire, a tipului răspunsului şi al tipului aşteptat al răspunsului am trimis două rulări.

Vom vedea în continuare modul de funcţionare a fiecărei componente din sistemul prezentat mai sus.

3.1. Construirea şabloanelor

Pentru a putea folosi sistemul de inferenţe textuale am construit din întrebările iniţiale o mulţime de şabloane folosind tehnici asemănătoare celor prezentate în (Bar-Haim et al., 2006). Astfel, pentru întrebarea 1 din datele de test:

Întrebarea 1: *Câte zile avea aprilie înainte de 700 î.Hr.?*

şablonul construit are forma:

Şablon 1: *Aprilie înainte de 700 î.Hr. avea NUMĂR zile.*

unde NUMĂR reprezintă o variabilă ce va fi înlocuită cu toate răspunsurile posibile pentru această întrebare. Pentru limba română am considerat şabloane specifice pentru următoarele tipuri de răspunsuri: DATĂ CALENDARISTICĂ (DATĂ, AN), DEFINIŢIE, MĂSURĂ, LOCAŢIE (ŢARĂ, ORAŞ), NUMĂR, PERSOANĂ, ORGANIZAŢIE, ALTCEVA. Se observă că atunci când a fost posibil am identificat tipuri cât mai specifice pentru tipul răspunsului. Tabelul de mai jos prezintă exemple de astfel de şabloane pentru fiecare tip în parte:

Tabel 2: Şabloane asociate întrebărilor

Tipul aşteptat al răspunsului	**Exemplu de întrebare**	**Şablon**
NUMĂR	Câţi jucători participă la jocul de bridge?	NUMĂR jucători participă la jocul de bridge.
MĂSURĂ	Ce lungime are Biserica Neagră din Braşov?	Biserica Neagră din Braşov are MĂSURĂ.
LOCAŢIE	Unde s-a născut Emil Constantinescu?	Emil Constantinescu s-a născut în LOCAŢIE.
ORAŞ	În ce oraş s-a născut Charlie Chaplin?	Charlie Chaplin s-a născut în ORAŞ.
PERSOANĂ	Ce zeiţă, soră a lui Ares, este fiica lui Metis?	PERSOANĂ, soră a lui Ares, este fiica lui Metis.
ORGANIZAŢIE	Din ce organizaţie teroristă face parte Osama bin Laden?	Din ORGANIZAŢIE teroristă face parte Osama bin Laden.
DATĂ	Când s-a vândut primul produs	La DATĂ s-a vândut primul

	Apple?	produs Apple.
AN	În ce an a fost produs filmul românesc Furia?	În AN a fost produs filmul românesc Furia.
ALTCEVA	Din ce se produce caşcavalul?	Caşcavalul se produce din ALTCEVA.

O situaţie deosebită a fost pentru cazul întrebărilor de tip definiţie. În acest caz am considerat doar răspunsul ca fiind ipoteza ce va fi trimisă sistemului de inferenţe textuale, fără a mai lua ceva din întrebarea iniţială.

3.2. Construirea ipotezelor şi a textelor

În şabloanelor construite ca mai sus am înlocuit variabilele folosind răspunsurile din datele de intrare şi am construit ipotezele. Astfel, pentru întrebarea 1 în şablonul "*Aprilie înainte de 700 î.Hr. avea NUMĂR zile.*" am înlocuit variabila NUMĂR cu toate cele 4 valori posibile ale răspunsurilor corespunzătoare din tabelul 1. În urma înlocuirii am obţinut cele 4 ipoteze de mai jos:

*I_{1_1}: Aprilie înainte de 700 î.Hr. avea **30** zile.*

*I_{1_6}: Aprilie înainte de 700 î.Hr. avea **29 de zile** zile.*

*I_{1_7}: Aprilie înainte de 700 î.Hr. avea **de anul 700** zile.*

*I_{1_8}: Aprilie înainte de 700 î.Hr. avea **cu aceeaşi zi a săptămânii în toţi anii** zile.*

Pentru aceste ipoteze, cele 4 texte le obţinem din fragmentele de text corespunzătoare din tabelul 1:

T_{1_1}: Înainte de anul 700 î.Hr., luna aprilie era a doua lună a anului în calendarul roman şi avea 29 de zile. După ce Iuliu Cezar a introdus calendarul iulian în 45 î.Hr., luna aprilie avea 30 de zile şi devenea a patra lună a anului.

T_{1_6}: Hr., luna aprilie era a doua lună a anului în calendarul roman şi avea 29 de zile. După ce Iuliu Cezar a introdus calendarul iulian în 45 î.

T_{1_7}: Numele lunii aprilie (latină: Aprilis) vine de la cuvântul latinesc aperio, ire = a deschide, deoarece în aprilie se deschid mugurii plantelor. Înainte de anul 700 î.

T_{1_8}: Aprilie începe cu aceeaşi zi a săptămânii ca şi Iulie în toţi anii şi ca Ianuarie în anii bisecţi.

3.3. Folosirea sistemului de inferenţe textuale pentru limba română

Sistemul de inferenţe textuale folosit pentru limba română (Iftene, Balahur-Dobrescu, 2007) primeşte la intrare perechi de tip (*ipoteză*, *text*) şi oferă la ieşire un scor de potrivire, iar în plus precizează dacă există probleme cu entităţile de tip nume. Problemele de acest tip apar în cazurile în care în ipoteză avem o entitate de tip nume căreia nu-i găsim corespondent în text. Acest lucru se întâmplă dacă entitatea cu probleme apare în întrebare, dar nu apare în fragmentul de text care ar trebui să justifice alegerea răspunsului curent, sau în cazul în care entitatea nu apare în fragmentul de text justificator. În ambele cazuri considerăm ca nejustificată alegerea răspunsului curent şi stabilim răspunsul final ca fiind RESPINS.

În tabelul de mai jos avem scorurile asociate celor 4 perechi (text, ipoteză) de mai sus:

Tabel 3: Scorurile asociate perechilor (T, I) corespunzătoare întrebării 1

Perechea	**Scor de potrivire**	**Entitatea de tip nume cu probleme**
($T_{1\ 1}$, $I_{1\ 1}$)	0.727	-
($T_{1\ 6}$, $I_{1\ 6}$)	0.889	-
($T_{1\ 7}$, $I_{1\ 7}$)	0.636	î.Hr.
($T_{1\ 8}$, $I_{1\ 8}$)	0.563	î.Hr.

3.4. Identificarea tipurilor răspunsurilor şi a tipului aşteptat al răspunsurilor

Scopul acestui pas este de a elimina din start cazurile în care aceste valori sunt diferite pentru întrebarea curentă şi un răspuns curent al acesteia.

Pentru identificarea *tipurilor răspunsurilor* (TR) pentru limba română am folosit din GATE[3] următoarele tipuri de entităţi de tip nume: Oraş, Companie, Ţară, Organizaţie, Persoană, Regiune. În plus am folosit şabloane specifice pentru identificarea NUMERELOR, DATELOR CALENDARISTICE, ANILOR şi a MĂSURILOR.

La identificarea *tipului aşteptat al răspunsului* (TAR) am utilizat aceleaşi valori ca cele folosite la construirea şabloanelor din întrebări, prezentate în tabelul 2: DATĂ CALENDARISTICĂ (DATĂ, AN), DEFINIŢIE, MĂSURĂ, LOCAŢIE (ŢARĂ, ORAŞ), NUMĂR, PERSOANĂ, ORGANIZAŢIE, ALTCEVA.

Pentru întrebarea 1 avem următoarele valori:

Tabel 4: Întrebarea 1 - Valoarea TAR şi valorile TR asociate răspunsurilor

TAR	**Răspuns**	**TR**	**Scor potrivire între TAR şi TR**
NUMĂR	30	NUMĂR	1
	29 de zile	MĂSURĂ	0.5
	de anul 700	ALTCEVA	0.25
	cu aceeaşi zi a săptămânii în toţi anii	ALTCEVA	0.25

unde scorul de potrivire dintre TAR şi TR s-a calculat similar modului în care am calculat această valoare pentru limba engleză. Tabelul 5 ne prezintă principalele situaţii întâlnite:

Tabel 5: Calcularea scorului de potrivire dintre TAR şi TR

Situaţie	**Scor de potrivire**
TAR = TR	1
(TAR = “DEFINIŢIE”) şi (TR = “ALTCEVA”)	1
TAR şi TR sunt în aceeaşi clasă de entităţi: {ORAŞ, ŢARĂ, REGIUNE, LOCAŢIE} sau {AN, DATĂ} sau {NUMĂR, MĂSURĂ, AN}	0.5
(TR = “ALTCEVA”) sau (TAR = “ALTCEVA”)	0.25
În celelalte cazuri	0

[3] GATE: http://www.gate.ac.uk/

3.5. Caracteristicile rulărilor trimise

Pe limba română am trimis două rulări, diferenţa dintre ele constând în faptul că am folosit sau nu tabelul 5 pentru a compara valorile TAR cu valorile TR.

Rularea 1: nu foloseşte comparaţia dintre TAR şi TR, ci doar ieşirea oferită de SIT-ul românesc. Astfel, răspunsurile pentru care avem probleme cu entităţile de tip nume sunt considerate ca fiind RESPINSE (cazurile răspunsurilor cu id-urile 7 şi 8 de la întrebarea 1). Toate celelalte sunt considerate ca fiind VALIDATE (cazurile răspunsurilor cu id-urile 1 şi 6 de la întrebarea 1). Ca SELECTAT este considerat răspunsul VALIDAT care are scorul de potrivire cel mai mare oferit de SIT-ul românesc (răspunsul cu id-ul 6 de la întrebarea 1). Toate valorile pentru întrebarea 1 sunt prinse în acest caz în tabelul 6. Putem observa cum în acest caz din cele 4 răspunsuri am obţinut valoarea corectă în 3 cazuri.

Tabel 6: Rularea 1: Valorile obţinute pentru răspunsurile de la întrebarea 1

Răspuns	**Valoare obţinută**	**Valoare corectă**
30	VALIDAT	RESPINS
29 de zile	SELECTAT	VALIDAT
de anul 700	RESPINS	RESPINS
cu aceeaşi zi a săptămânii în toţi anii	RESPINS	RESPINS

Rularea 2: ca mai sus, dar foloseşte în plus comparaţia dintre TAR şi TR. Comparaţia dintre TAR şi TR este folosită astfel: se consideră răspunsuri RESPINSE cele care au probleme cu entităţile de tip nume sau cele pentru care scorul de potrivire dintre TAR şi TR este 0. Pentru întrebarea 1, deoarece în tabelul 4 nu avem scoruri de potrivire 0, vom considera ca fiind RESPINSE cazurile răspunsurilor cu id-urile 7 şi 8, care au probleme cu entităţile de tip nume. Restul răspunsurilor se consideră ca fiind VALIDE (cazurile răspunsurilor cu id-urile 1 şi 6 de la întrebarea 1). Dintre răspunsurile VALIDE pentru a decide care dintre răspunsuri este SELECTAT folosim şi comparaţia dintre TAR şi TR. Vom considera ca fiind SELECTAT, răspunsul cu cel mai mare scor de potrivire întors de SIT-ul românesc, dintre cele cu cel mai mare scor de potrivire dintre TAR şi TR. Deoarece avem un singur răspuns cu cel mai mare scor de potrivire dintre TAR şi TR, răspunsul 1, care are scorul de potrivire 1, este şi cel ales ca fiind selectat. Valorile pentru întrebarea 1 sunt prinse în tabelul 7 de mai jos. De observat faptul că numărul răspunsurilor în care am dat răspunsul corect este 3.

Tabel 7: Rularea 2: Valorile obţinute pentru răspunsurile de la întrebarea 1

Răspuns	**Valoare obţinută**	**Valoare corectă**
30	SELECTAT	RESPINS
29 de zile	VALIDAT	VALIDAT
de anul 700	RESPINS	RESPINS
cu aceeaşi zi a săptămânii în toţi anii	RESPINS	RESPINS

4. Rezultate

Rezultatele oficiale obţinute pentru limba română sunt prezentate mai jos:

Tabel 8: Rezultatele obţinute în competiţia AVE2008 pe limba română

Rezultate	**Rularea 1**	**Rularea 2**
F-measure	0.22	0.23
Precizia	0.12	0.13
Recall	0.92	0.92
qa_accuracy	0.17	0.24
estimated_qa_performance	0.17	0.25

unde *precizia*, *recall* şi *F-measure* au ca scop evaluarea unui sistem care ordonează şi filtrează răspunsurile. Formulele care au fost aplicate pentru acestea sunt următoarele:

$$precizie = \frac{prezise\_corect\_de\_sistem\_ca\_SELECTATE\_sau\_VALIDATE}{prezise\_de\_sistem\_ca\_SELECTATE\_sau\_VALIDATE}$$

$$recall = \frac{prezise\_corect\_de\_sistem\_ca\_SELECTATE\_sau\_VALIDATE}{multimea\_raspunsurilor\_SELECTATE\_sau\_VALIDE}$$

$$F - measure = \frac{2 \times recall \times precizie}{recall + precizie}$$

Iar *qa_accuracy* şi *estimated_qa_performance* au ca scop compararea performanţelor unui sistem de tip ÎR cu un sistem ipotetic de tip ÎR care ar folosi şi un sistem AVE, formulele aplicate fiind:

$$qa\_accuracy = \frac{Raspunsuri\_SELECTATE\_Corect}{Numarul\_Intrebarilor}$$

$$qa\_rej\_accuracy = \frac{Raspunsuri\_RESPINSE\_Corect}{Numarul\_Intrebarilor}$$

$$estimated\_qa\_performance = qa\_accuracy + qa\_accuracy * qa\_rej\_accuracy$$

După cum se observă din tabelul 8 rularea a doua, care foloseşte şi comparaţia dintre TAR şi TR, este mai bună. Analizând rezultatele obţinute am putut observa cum din cele 52 de răspunsuri VALIDE aflate în fişierul de test sistemul nostru oferă 48 dintre ele, din care 28 sunt SELECTATE (de aici valoarea foarte mare a *recall*-ului). Precizia mică obţinută se datorează faptului că sistemul nostru prin modul în care e construit are condiţii foarte stricte pentru a stabili dacă un răspuns este RESPINS şi prin urmare oferă foarte multe răspunsuri VALIDE şi SELECTATE şi foarte puţine răspunsuri RESPINSE. Pe de altă parte, trebuie să precizăm faptul că din cele 73 de răspunsuri RESPINSE date de sistemul nostru, 69 au fost corecte.

5. Concluzii

Lucrarea prezintă principalele componente ale sistemului folosit în competiţia AVE de anul acesta pe limba română. Sistemul construit a avut o comportare foarte bună pentru răspunsurile VALIDE, dar nu a tratat aproape deloc răspunsurile RESPINSE.

De remarcat este faptul că din cele două rulări, rularea a doua are valoarea *estimated_qa_performance* de 0.25 care este superioară preciziei celui mai bun sistem din competiţia sistemelor de tip întrebare-răspuns pe limba română de anul acesta. Acest

lucru ne indică faptul că folosirea acestui sistem de ordonare a răspunsurilor în cadrul unui sistem de tip Întreabare-Răspuns ar duce la creşteri semnificative ale preciziei pentru un astfel de sistem.

Pe viitor avem ca principal obiectiv eliminarea a două neajunsuri majore ale sistemului: Primul este datorat faptului că am considerat o mulţime de şabloane pentru identificarea tipului răspunsului, iar situaţiile noi care pot apare ar fi tratate incorect. Cel de-al doilea este datorat faptului că am considerat condiţii prea stricte pentru identificarea răspunsurilor RESPINSE, iar numărul acestora este prea mic.

Mulţumiri. Mulţumim colegilor de la Facultatea de Informatică care ne-au ajutat la construirea anumitor componente ale sistemului. Lucrul din cadrul acestui proiect este parţial finanţat de proiectul PNCDI II, SIR-RESDEC şi de firma Siemens VDO Iaşi.

Referinţe bibliografice

Bar-Haim, R., Dagan, I., Dolan, B., Ferro, L., Giampiccolo, D., Magnini B., Szpektor, I. (2006). The Second PASCAL Recognising Textual Entailment Challenge. *In Proceedings of the Second PASCAL Challenges Workshop on Recognizing Textual Entailment*. Venice. Italy.

Iftene, A., Balahur-Dobrescu, A. (2007). Improving a QA System for Romanian Using Textual Entailment. *In Proceedings of RANLP workshop "A Common Natural Language Processing Paradigm For Balkan Languages"*. Pages 7-14, September 26, Borovets, Bulgaria.

Iftene, A., Balahur-Dobrescu, A. (2008). Answer Validation on English and Romanian Languages. *In Proceedings of the CLEF 2008 Workshop.* 17-19 September. Aarhus, Denmark.

Peñas, A., Rodrigo, Á., Sama, V., Verdejo. F. (2007) Overview of the Answer Validation Exercise 2006. CLEF 2006, *Lecture Notes in Computer Science* LNCS 4730. Springer, Berlin.

Rodrigo, Á., Peñas, A., Verdejo, F. (2008). Overview of the Answer Validation Exercise 2008. *In working notes of CLEF2008.* September, Aarhus, Denmark.

ECHIVALAREA ÎN LIMBA ROMÂNĂ A UNITĂŢILOR FRAZEOLOGICE INFINITIVALE DIN LIMBA FRANCEZĂ

MARIA HUSARCIUC[1,2,3]

[1]*Facultatea de Litere, Universitatea "Al. I. Cuza", Iaşi*
[2]*Facultatea de Informatică, Universitatea "Al. I. Cuza", Iaşi*
[3]*Institutul de Informatică Teoretică Iaşi, Academia Română*

mhusarciuc@gmail.com

Rezumat

Lucrarea de faţă prezintă o metodă de identificare a unităţilor frazeologice (expresii idiomatice, locuţiuni, structuri proverbiale etc.) într-un corpus paralel francez-român. Etapele urmărite sunt: adnotarea unităţilor frazeologice în textul în limba franceză şi importul acestei adnotări pentru cele două variante de traducere în limba română. Se au în vedere criterii specifice de identificare a acestor unităţi şi a tipurilor lor specifice, precum şi dificultăţile care pot să apară în etapa de import, aplicate pe structuri cu verbul *faire* la infinitiv, în limba franceză, cu diferite posibilităţi de traducere în limba română.

1. Introducere

În realizarea resurselor lexical-semantice superioare dicţionarelor informatizate, problemele cele mai mari apar în cazul mutaţiilor semantice în contexte specifice şi în cazul unităţilor frazeologice. Un imperativ, în acest caz, este descoperirea unui mod, flexibil şi riguros în acelaşi timp, de a „manipula" îmbinările stabile de cuvinte. Astfel a luat naştere un subdomeniu al lingvisticii computaţionale, cunoscut sub denumirea de *frazeologie computaţională* (engl. *computational phraseology*, v. Heid, 2005), disciplină complementară aşa numitei *frazeologii tradiţionale*. Lucrarea de faţă prezintă o metodă de identificare a unităţilor frazeologice într-un corpus bilingv, francez-român, cu particularitatea că, pentru acelaşi text în limba franceză, sunt avute în vedere două versiuni de traducere în limba română. Identificarea acestor unităţi şi stabilirea tipurilor în care ele se încadrează se realizează pornind de la criterii precise.

Lucrarea conţine, în a doua secţiune, o prezentare a premiselor teoretice, cu accent pe criterii de identificare a unităţilor frazeologice, stabilind totodată o tipologie a acestora. A treia secţiune are în vedere analiza bazată pe corpus, cu accent pe structurile care conţin, în limba franceză, verbul *faire* la infinitiv şi pe probleme care apar la importul adnotărilor în textele în limba română. Lucrarea se încheie cu prezentarea concluziilor şi a perspectivelor viitoare de lucru.

2. Premise teoretice

2.1. Criterii de identificare a unităţilor frazeologice

Lăsând la o parte criteriul implicit, conform căruia orice unitate frazeologică trebuie să conţină minim două lexeme, reprezentative sunt frecvenţa, instituţionalizarea, stabilitatea, caracterul idiomatic, variaţia şi caracterul gradual (Corpas Pastor, 1996).

2.1.1. Frecvenţa

Principalul criteriu de identificare a unităţilor frazeologice este frecvenţa cu care ele apar, văzută atât ca frecvenţă de co-ocurenţă a elementelor componente într-un text sau corpus de texte, cât şi ca frecvenţă de utilizare a structurii respective în vorbire. Se pleacă de la ideea conform căreia, cu cât este mai des folosită o structură sintactico-semantică într-o limbă dată, cu atât sunt mai mari şansele ca acea structură să fie sau să devină îmbinare stabilă de cuvinte sau chiar expresie idiomatică.

(Jean David, 1988) demonstrează, cu exemple, faptul că trecerea de la ne-idiomatic la idiomatic în limbă depinde de frecvenţa de utilizare a anumitor structuri. Orice îmbinare stabilă de cuvinte, care are o semnificaţie globală cunoscută, poate avea şi o lectură „compoziţională" şi, reciproc, orice îmbinare liberă de cuvinte, a cărei semnificaţie de bază este semnificaţia elementelor componente, poate câştiga o semnificaţie globală (conotaţie) datorită unor elemente situate în realitatea extra-lingvistică (premise sociale, de exemplu).

2.1.2. Instituţionalizarea

Prin utilizare frecventă, structurile neologice dobândesc un caracter oficial, ajungând să fie reproduse în vorbire fără modificarea formei. Fiind o etapă tranzitorie între intrarea în uz a unei structuri şi stabilizarea ei morfo-sintactică şi semantică, instituţionalizarea constituie un criteriu destul de ambiguu în identificarea unităţilor frazeologice.

2.1.3. Stabilitatea

Proces complex, stabilitatea se realizează în două etape: la nivelul formei (fr. *figement*, sp. *fijacion*) şi la nivelul conţinutului (specializare semantică/ lexicalizare).

Stabilitatea este considerată aproape unanim principala trăsătură definitorie a unităţilor frazeologice. Totuşi, contrar semnificaţiei implicite a termenului, stabilitatea este o trăsătură greu de hotărât cu precizie, fiind mereu relativă la procesul continuu de construire şi re-construire a faptelor de limbă (v. Eric Beaumatin, 1988). De aceea, între stabilitate şi idiomaticitate limitele sunt destul de fragile.

2.1.4. Idiomaticitatea

În literatura de specialitate, stabilitatea şi idiomaticitatea sunt de regulă prezentate împreună, ca două faţete ale aceluiaşi proces, definindu-se una prin cealaltă. Prin stabilizarea unei îmbinări libere de cuvinte, aceasta are şanse mari să se "idiomatizeze" şi, reciproc, nici o expresie idiomatică nu poate exista dacă nu a fost în prealabil stabilizată morfo-sintactic.

(Harald Burger, 1988), analizând cele două concepte de bază ale cercetării frazeologice, pe care el le numeşte stabilitate (*Festigkeit*) şi metaforicitate/ caracter metaforic (*Metaphorizität*), accentuează faptul că numai primul este un concept din sfera frazeologiei, al doilea fiind mai general. Există însă o inter-dependenţă între ceea ce exprimă aceste două concepte, fapt foarte bine ilustrat în lucrările lexicografice în care, adesea, unităţile frazeologice sunt marcate cu trăsătura „metaforic" (sau „figurat"), uneori specificându-se că structura respectivă este o expresie sau o locuţiune.

(Bernd Spillner, 1988), făcând distincţia între *figement syntaxique* (termen asimilabil stabilităţii sintactice şi care stă la baza formării colocaţiilor) şi *figement phraséologique* (termen sinonim cu idiomatizarea, dând naştere frazeologismelor), consideră caracterul idiomatic drept definitoriu pentru unităţile frazeologice, colocaţiile fiind doar fenomene tranzitorii.

2.1.5. Variaţia

Variaţia frazeologică este o regulă lingvistică pe baza căreia se poate stabili gradul de regularitate al unui sistem frazeologic dat. Deşi majoritatea autorilor vorbesc de existenţa a două tipuri de variaţie frazeologică (variantele şi modificările creative), (Duneton & Claval, 1990) demonstrează, cu exemple pentru limba franceză, că aşa-numitele variante frazeologice sunt în realitate expresii vecine, similare morfo-sintactic, şi nu variante ale aceleiaşi expresii. În lucrarea citată sunt analizate structurile *être au bout de son rouleau* şi *être au bout du rouleau*, care, deşi foarte asemănătoare structural (singura diferenţă este prezenţa sau absenţa pronumelui reflexiv) au origini şi sensuri diferite.

Spre deosebire de variante, modificările creative sunt contaminări conştiente între unităţi frazeologice din aceeaşi sferă semantică, de genul *Cine sapă groapa altuia departe ajunge*, dar nu şi modificările inconştiente, datorate lipsei de cultură, care apar în structuri ca *a unsprezecea minune a lumii*.

2.1.6. Gradaţia

Gradaţia este o proprietate a însuşi sistemului frazeologic, bazat pe îmbinări de cuvinte cu diferite grade de idiomaticizare. Trecerea de la ne-idiomatic la idiomatic este posibilă numai datorită caracterului gradual al acestui sistem.

Acestor criterii le mai putem adăuga şi capacitatea unităţilor frazeologice de a conserva termeni ieşiţi din uz: *pe de rost, avoir la berlue*.

2.2. Tipuri de unităţi frazeologice

Unităţile frazeologice sunt enunţuri aparţinând discursului repetat, „prefabricate de vorbire", care reprezintă „tot ceea ce în vorbirea unei comunităţi se repetă într-o formă mai mult sau mai puţin identică de discurs deja făcut" (Coşeriu, 2000).

Dintre termenii folosiţi pentru a desemna unitatea minimală a frazeologiei (izolare, frazeologism etc.), cel mai adecvat din punctul de vedere al consecvenţei modului de definire este termenul *unitate frazeologică*. Poate fi unitate frazeologică orice îmbinare stabilă de cuvinte, fie expresie sau locuţiune expresivă, acestea la rândul lor fiind

subcategorizabile. (Dimitrescu, 1958) stabileşte o serie de trăsături de diferenţiere a locuţiunilor de expresii, recunoscându-le locuţiunilor proprietatea de a se comporta ca o singură parte de vorbire, funcţia gramaticală unică şi posibila coloratură expresivă, în timp ce expresiile sunt întotdeauna marcate stilistic, sunt variabile şi fără o funcţie gramaticală precisă. Criterii diferite de identificare a locuţiunilor găsim în (Branca-Rosoff, 1997): imposibilitatea realizării mutaţiilor interne sau a substituţiilor, caracterul inseparabil al elementelor, „pierderea sentimentului de analicitate" datorită închegării. În plus, locuţiunile verbale oferă constrângeri asupra determinanţilor şi asupra posibilităţilor de pronominalizare.

Alţi autori stabilesc subcategorii de expresii. (Slave, 1966) clasifică expresiile în două tipuri: „consacrate şi în uz propriu, având accepţie tehnică sau fiind folosite ca îmbinări curente, banale" şi „îmbinări folosite numai cu accepţie figurată". Exemplele pe care le alege autoarea (*a aduce la acelaşi numitor, a se spăla pe mâini* etc. pentru prima clasă şi *a avea păr pe limbă, a scoate vorba cu cleştele* pentru a doua) corespund celor două clase de expresii denumite de (Dumistrăcel, 2001) „cópii ale realităţii", respectiv „imaginare". În prezent, se foloseşte foarte mult termenul *colocaţie* pentru a desemna îmbinări stabile de cuvinte care nu au în mod obligatoriu sens conotativ. (Todiraşcu *et al*., 2007)

În linii mari, termenii vehiculaţi reprezintă fie

a) tipuri complementare de structuri lexical-sintactice: expresii (având de regulă sens conotativ), locuţiuni (structuri morfo-sintactice complexe cu rol funcţional, uneori putând avea şi conotaţii), colocaţii (îmbinări stabile de cuvinte care nu sunt nici expresii, nici locuţiuni, de genul formulelor de salut şi a clişeelor lingvistice), realizându-se astfel o clasificare cvasi-exhaustivă a tuturor îmbinărilor stabile de cuvinte,

fie

b) tipuri (cvasi-)concentrice de structuri lexical-semantice: unul dintre termeni reprezintă, în acest caz, o categorie supra-ordona(n)tă căreia i se subordonează celelalte.

3. Analiza bazată pe corpus

3.1. Metoda de lucru

3.1.1. Adnotarea unităţilor frazeologice în textul în limba franceză

Structuri sintactice similare pot reprezenta sau nu unităţi frazeologice. Având în vedere dificultatea diferenţierii îmbinărilor stabile de cuvinte de îmbinările libere, metoda pe care am ales-o este adnotarea manuală a acestor structuri în textul original (în limba franceză) şi importarea adnotărilor în traducerea românească aliniată cu originalul. Textul folosit este *Madame Bovary* de Gustave Flaubert, cu două versiuni de traducere în limba română: a lui Ludovic Dauş (Flaubert, 1915) şi a lui Demostene Botez (Flaubert, 1968). Adnotarea unităţilor frazeologice se realizează pe textul francez, tokenizat în prealabil, folosind adnotatorul PALinkA.

Următorul exemplu prezintă o secvenţă din fişierul XML de ieşire, ce conţine adnotarea unei unităţi frazeologice din textul în limba franceză.

```
<MWE
      DEF="v.TLFI: FAIRE-VALOIR, subst. masc. invar. Mode de
gestion, d'exploitation d'un capital immobilier."
      HEADID="faire"
      ID="0"
      OBS="Structură folosită aici cu sensul de bază"
      OTHER_TYPE=""
      TYPE="IDIOM">
<W id="1298">faire</W>
<W id="1299">valoir</W>
</MWE>
```

Dintre elementele avute în vedere în adnotarea manuală, important este tipul (TYPE), care trebuie ales dintr-o listă reprezentând IDIOM (expresie idiomatică), EXPRESSION (expresie, structură cu sens conotativ) COLLOCATION (colocaţie sau locuţiune) , PROVERB (proverb) sau OTHER (în cazul în care o structură nu corespunde niciuneia din clasele de mai sus, caz în care completarea unei valori pentru OTHER_TYPE este foarte importantă. Definiţia (DEF) va fi preluată din *Le Trésor de la Langue Française Informatisé*. ID-ul unităţii frazeologice adnotate este incrementat automat, iar HEADID-ul este reprezentat de nucleul sintactico-semantic al respectivei structuri. În OBS se pot nota, în timpul adnotării, observaţii suplimentare ce pot fi utile la o prelucrare ulterioară.

3.1.2. Importarea adnotărilor în traducerile în limba română, aliniate cu textul original

Etapa următoare adnotării manuale a unităţilor frazeologice în textul francez este importarea lor în traducerile în limba română. Acest lucru se poate realiza numai după alinierea, la nivel de cuvânt, a textelor. Premisa de la care se pleacă este că importarea unei structuri se poate realiza chiar şi în cazul lipsei unei corespondenţe literale a expresiilor în cele două limbi, deoarece elementele care preced şi succed expresia sunt, de regulă, aliniate corespunzător. Este important să se ţină cont de contextul de apariţie, o aliniere a elementelor din interiorul unor unităţi frazeologice echivalente fiind greu de realizat, datorită unor neconcordanţe de structură (a se vedea, pentru detalii, subsecţiunea 3.3).

Realizarea unor şabloane cu ajutorul cărora să se precizeze structurile morfo-sintactice vizate (de exemplu, precizarea faptului că unei structuri infinitivale din franceză îi poate corespunde, în limba română, o structură cu verbul la conjunctiv) va creşte acurateţea identificării lor în traducerile româneşti.

3.2. *Tipuri de structuri ce-l conţin pe* **faire** *în limba franceză*

Verbul *faire*, relevant pentru polisemantismul său, apare frecvent în *Madame Bovary* la modul infinitiv. Structurile care îl conţin sunt dintre cele mai variate. Exemplele extrase din corpusul adnotat au fost foarte utile în marcarea distincţiei între diferite tipuri de unităţi frazeologice, ţinând cont de tipologia lor şi de criteriile de identificare prezentate în secţiunea a doua.

3.2.1. Expresii idiomatice

O expresie idiomatică este caracterizată prin frecvenţă redusă de apariţie în acelaşi text, stabilitate morfo-sintactică, fiind intrată în uz, deci recunoscută de vorbitorii nativi; este o structură cu sens figurat, metaforic, dezvoltat pe baza unui sens primar. O expresie idiomatică poate fi specializată pentru un anumit domeniu sau poate să apară numai în contexte specifice. Adesea, structurile din această categorie sunt, în corpusul ales, fie traduse prin parafraze (a), fie absente (b).

> *L'officier de santé, chemin faisant, comprit aux discours de son guide que M. Rouault devait être un cultivateur des plus aisés. Il s'était cassé la jambe, la veille au soir, en revenant de _* ***faire les Rois*** *_, chez un voisin.*
>
> (a) *Pe drum, din tot ce-i spunea călăuza, ofiţerul sanitar îşi dădu seama că domnul Rouault trebuie să fie un gospodar din cei mai înstăriţi. Îşi rupsese piciorul, în ajun, seara, întorcându-se de la un vecin unde* ***serbase Boboteaza****.* (Botez)
>
> (b) *Ofiţerul de sănătate, pe drum, înţelese din vorbele călăuzei sale că d. Rouault trebuia să fie un cultivador dintre cei mai bogaţi. Îşi frânsese piciorul, seara din ajun, întorcându-se de la un vecin.* (Dauş)

3.2.2. Expresii

Structuri intermediare între colocaţii şi expresii idiomatice, expresiile sunt frecvente ca mod de utilizare, stabile din punct de vedere morfo-sintactic şi au sens conotativ. În general, echivalenţii de traducere pentru aceste structuri sunt fie calcuri lexical-semantice după limba franceză, fie expresii deja existente şi frecvente.

> *A l'encontre des tendances maternelles, il avait en tête un certain idéal viril de l'enfance, d'après lequel il tâchait de former son fils, voulant qu'on l'élevât durement, à la spartiate, pour* ***lui faire une bonne constitution****.*
>
> *Potrivnic înclinaţiilor materne, avea în cap un anume ideal viril despre copilărie, după care încerca să-şi educe feciorul, voind să fie crescut cu asprime, după moda spartană,* ***ca să ajungă voinic****.* (Botez)
>
> *În contra tendinţelor materne, avea în cap un oare-care ideal viril al copilăriei, după care căuta să-şi formeze feciorul, voind să-l crească cu asprime după moda spartacă* ***ca să-l facă tare de constituţie****.* (Dauş)

3.2.3. Colocaţii

Structuri morfo-sintactice complexe cu rol funcţional, colocaţiile pot prezenta uneori şi sensuri conotative. Diferenţele între colocaţii şi locuţiuni nu sunt foarte clar formalizabile, de aceea am păstrat o singură categorie. Cele mai frecvente structuri din această categorie care-l conţin în limba franceză pe *faire* la infinitiv sunt sintagmele cu determinări nominale: *faire sa toilette, faire du punch, faire de la tapisserie*, care sunt traduse în general literal în limba română.

3.2.4. Predicate verbale compuse

Predicatele verbale compuse din limba franceză, în care verbul *faire* la infinitiv are valoare de auxiliar şi este urmat de un alt infinitiv, sunt de regulă traduse în limba română prin predicate verbale simple, fiind diferite de *predicatele complexe*, care reprezintă structuri compuse din verb la infinitiv şi grup substantival sau prepoziţional, de genul *a aduce atingere, a intra în vigoare* etc. (Todiraşcu *et al.*, 2007).

Un exemplu de predicat verbal compus din limba franceză este *faire étudier la medicine*, structură ale cărei traduceri în limba română vor fi analizate în secţiunea 3.3.1.

3.3. Probleme la alinierea traducerii în limba română a structurilor cu faire

3.3.1. Adesea pot să apară probleme la alinierea celor două versiuni de traducere în limba română, mai ales datorită faptului că Ludovic Dauş recurge mult mai des la traduceri literale ale expresiilor decât Demostene Botez. Un exemplu ar fi structura infinitivală „*pour lui faire une bonne constitution*", care în traducerea lui Dauş apare sub forma „*ca să-l facă tare de constituţie*", în timp ce D. Botez o traduce prin: „*ca să ajungă voinic*".

Ludovic Dauş păstrează de regulă structura sintactică a frazei din originalul franţuzesc şi traduce adesea literal îmbinările de cuvinte (de exemplu, traduce *lui faire étudier la médecine* prin *să-l facă să înveţe medicina*). Demostene Botez simplifică, în unele cazuri, structura frazei, recurge la inversiuni sau parafraze (traduce, de exemplu, aceeaşi structură prin *să-l înscrie la Medicină*).

3.3.2. Problematice sunt şi locuţiunile cu sens general, particularizat în text datorită existenţei unei referinţe anaforice concrete. Prin traducere, se poate păstra referinţa anaforică la o structură exprimată anterior (cum procedează L. Dauş în fraza (2), traducere fidelă a frazei din franceză) sau se poate folosi un echivalent particular, cu explicitarea sensului (ca în (3), unde D. Botez traduce locuţiunea *faire la demande* prin *s-o ceară în căsătorie*, deşi în paragraful anterior este prezentată intenţia lui Charles de a se căsători cu Emma).

(1) „*Charles se promit de* ***faire la demande*** *quand l'occasion s'en offrirait*".

(2) „*Carol îşi făgădui* ***să facă cererea*** *când ocazia se va prezenta*".

(3) „*Charles se hotărî* ***s-o ceară în căsătorie*** *de îndată ce va avea ocazie*".

Aceste diferenţe nu impun dificultăţi în lectura şi înţelegerea textelor, dar în automatizarea procesului de extragere a diferitelor tipuri de unităţi frazeologice, în cazuri similare celui din (3) nu se va putea găsi echivalentul locuţiunii din (1) decât în situaţia în care corpusul va fi procesat suplimentar cu un sistem de rezoluţie a anaforelor.

3.3.3. Contextul este absolut necesar în unele cazuri pentru a discerne între sensul propriu al unei anumite unităţi frazeologice şi un sens particular. Locuţiunea verbală *faire valoir* are în limba franceză atât un sens propriu, general (ca în structura: *faire valoir une excuse*), cât şi sensul particular: „*Loc.* ***Faire valoir*** *ses droits*. Demander (pour soi) à une administration, à une hiérarchie, l'application d'une norme, d'un règlement" (TLFI). Acest sens particular se regăseşte în contextul „*se retira dans la campagne, où il voulut* «faire valoir»", tradus de L. Dauş: „*se retrase la ţară unde căută* să exploateze" şi în mod similar, dar mai explicit, de D. Botez: „*se retrase la ţară, unde voi* să exploateze singur o moşie".

3.3.4. Există cazuri în care anumite expresii fără echivalent idiomatic în limba română sunt pur şi simplu omise la traducere. În exemplul următor, expresia *faire les Rois*, are, în TLFI, următoarea definiţie: „***Faire les Rois*** (vieilli), *tirer les Rois*. Se réunir pour une fête qui consiste à partager la galette des Rois contenant la fève qui rend roi ou reine celui ou celle qui la trouve. [*M. Rouault*] *s'était cassé la jambe, la veille au soir, en revenant de* faire les Rois, *chez un voisin* (FLAUB., *M^{me} Bovary*, t. 1, 1857, p. 13). *Au jour de l'Épiphanie, M^{me} Bavretel conviait les amis d'Armand à venir « tirer les rois »* (GIDE, *Si le grain*, 1924, p. 474)."

Expresia idiomatică, întâlnită în *Madame Bovary* în fraza citată în definiţia din TLFI, nu are echivalent în traducerea lui Dauş („*Îşi frânsese piciorul, seara din ajun, întorcându-se de la un vecin.")*, în timp ce Demostene Botez îi găseşte un echivalent neidiomatic, un fel de definiţie prescurtată: „*Îşi rupsese piciorul, în ajun, seara, întorcându-se de la un vecin unde* serbase Boboteaza."

3.3.5. Expresiilor onomatopeice le sunt uneori găsite echivalenţe de traducere ce nu au în componenţa lor interjecţii, ca în propoziţia „*Ce n'est pas la peine de* faire tant de fla-fla", tradusă prin „*Pentru atâta lucru nu face* să-ţi iei aere" (L. Dauş) sau prin „*Nu-i cazul* să-ţi dai atâtea ifose" (D. Botez). Importarea automată a adnotării în situaţii de acest gen este posibilă numai dacă în faza de verificare a alinierii textelor se recurge la un artificiu, aliniindu-se fie interjecţia din original cu substantivul din traducere, fie direct întreaga expresie cu echivalentul ei în cele două versiuni româneşti.

4. Concluzii

Încercând să prezinte o metodă de identificare a unităţilor frazeologice într-un corpus beletristic bilingv, cu marcarea structurilor echivalente în franceză şi română, această lucrare a avut în vedere şi o încercare de tipologizare a acestor tipuri de unităţi, precum şi a dificultăţilor întâmpinate.

În perspectivă, se urmăreşte definitivarea adnotării şi găsirea de soluţii optime pentru realizarea unei resurse bilingve a unităţilor frazeologice, cu informaţii cât mai complexe.

Mulţumiri Cercetarea întreprinsă beneficiază de finanţare CNCSIS (Grant TD, cod 492) şi nu s-ar putea realiza fără sprijinul continuu al îndrumătorilor de doctorat, Prof. Dr. Eugen Munteanu şi Prof. Dr. Dan Cristea, cărora ţin să le mulţumesc pentru

viziunea critică oferită. Mulţumesc de asemenea membrilor grupului de lingvistică computaţională din Iaşi, care m-au ajutat în diferite etape ale cercetării mele de până acum.

Referinţe bibliografice

Beaumatin, Eric (2000). Langue/ discours/ texte à l'épreuve des faits de figement. Greciano, Gertrud (Ed.), *Micro- et macrolexemes et leur figement discursif* , Actes du colloque international CNRS URA 1035 Langue-Discours-Cognition, 6-7-8 décembre 1998, Saverne, Editions Peeters, Louvain / Paris, p. 3-12.

Branca-Rosoff, Sonia (1997). Modèles de locutionarité et effets de figement dans le discours politique de l'an II. „*La locution: entre lexique, syntaxe et pragmatique. Identification en corpus, traitement, apprentissage*", Textes réunis par Pierre Fiala, Pierre Lafon, Marie-France Piguet, Editions Klincksieck, Paris, 1997, p. 285-286.

Burger, Harald (1988). „BILDHAFT, ÜBERTRAGEN, METAPHORISCH...". Zur Konfusion um die semantischen Merkmale von Phraseologismen. *„EUROPHRAS 88. Phraseologie Contrastive"*, p. 17-29.

Corpas Pastor, Gloria (1996). *Manual de fraseología española*. Editorial Gredos, Madrid.

Coşeriu, Eugeniu (2000). *Lecţii de lingvistică generală*. Editura Arc, Chişinău.

David, Jean (1988). Tous les predicats ne meurent pas idiomes. Mais nul n'est a l'abri. „*EUROPHRAS 88. Phraseologie Contrastive*", Actes du Colloque International Klingenthal – Strasbourg, 12-16 mai 1988, Collection Recherches Germanique N.2, Strasbourg, p. 75-82.

Dimitrescu, Florica (1958). *Locuţiunile verbale în limba română*. Editura Academiei, Bucureşti.

Dumistrăcel, Stelian (2001). *Până-n pânzele albe. Expresii româneşti*. Biografii-motivaţii. Institutul European, Iaşi.

Duneton, Claude; Claval, Sylvie (1990). *Le Bouquet des expressions imagees. Encyclopedie thematique des locutions figurees de la langue francaise*. Editions du Seuil, Paris.

Flaubert, Gustave (1968). *Doamna Bovary*. În româneşte de Demostene Botez. Prefaţă de Aurelian Tănase. Editura pentru Literatură Universală, Bucureşti.

Flaubert, Gustave (1915). *Doamna Bovary*. Traducere de Ludovic Dauş. Ediţia a II-a, revăzută. Editura Minerva, Bucureşti.

Heid, Ulrich (2005). Computational Phraseology. Approaches to the computational analysis and representation of phraseological units and to their extraction from text corpora. *"PHRASEOLOGIE 2005 – La phraséologie dans tous ses états (Colloque interdisciplinaire)"*, Louvain-la-Neuve, p.13-15.

Slave, Elena (1966). *Structura sintagmatică a expresiilor figurate*, LL, XI, p. 397-413.

Spillner, Bernd (2000). Phraséologie et textologie comparées français – allemand. Greciano, Gertrud (Ed.), *Micro- et macrolexemes et leur figement discursif*, p. 23-32.

Todiraşcu, Amalia, Dan Ştefănescu, Christopher Gledhill (2007). Un sistem de extragere a colocaţiilor. *Lucrările atelierului Resurse lingvistice şi instrumente pentru prelucrarea limbii române*. Ed. Universităţii Al. I. Cuza, Iaşi, p. 119-129.

PALinkA, http://clg.wlv.ac.uk/projects/PALinkA/

Le Trésor de la Langue Française Informatisé http://atilf.atilf.fr/tlf.htm

COLECTAREA ŞI PROCESAREA DOCUMENTELOR ROMÂNEŞTI ALE CORPUSULUI JRC-ACQUIS

ALEXANDRU CEAUŞU

[1]*Institutul de Cercetări pentru Inteligenţa Artificială, Academia Română*

aceausu@racai.ro

Rezumat

Partea românească a corpusului JRC-Acquis continuă să crească odată cu creşterea numărului de documente ale "Acquis Communautaire" traduse în română. De asemenea, calitatea corpusului se îmbunătăţeşte deoarece multe din documentele deja traduse trec prin câteva faze de corectare. Lucrarea prezintă o nouă versiune românească a corpusului JRC-Acquis-Ro conţinând peste 30 de milioane de cuvinte în 19211 documente. Numărul de documente româneşti prezente în noua versiune este de trei ori mai mare decât cel din versiunea precedentă.

1. Introducere

Necesarul de corpusuri paralele pentru aplicaţiile de procesare a limbajului natural a cunoscut un trend ascendent accentuat pe parcursul ultimilor ani. Corpusurile paralele sunt folosite în aplicaţiile de traducerea automată sau categorizare multilingvală; pentru a produce resurse lexicale sau semantice multilinguale, cum sunt dicţionarele sau ontologiile; pentru a testa consistenţa procesului de traducere, etc. Cele mai multe corpusuri paralele existente conţin limbi de largă circulaţie şi au un număr redus de perechi de limbi. Dintre acestea, cel mai cunoscut este corpusul francez-englez Hansards (German, 2001). Corpusurile paralele ce conţin mai multe perechi de limbi sunt de mici dimensiuni sau pentru texte foarte specializate cum ar fi biblia (Resnik et al. 1999) sau romanul 1984 al lui George Orwell (Erjavec 2004). Unul dintre cele mai importante corpusuri multlinguale este EuroParl (Koehn, 2005) disponibil pentru 11 din limbile comunităţii europene.

Pentru aplicaţiile de procesare a limbajului natural din limba română cel mai important corpus este JRC-Acquis (Steinberger et al., 2006). Acesta este în prezent cel mai mare corpus multilingual disponibil, conţinând 22 de limbi. Corpusul este disponibil în format XML conform specificaţiilor TEI (Text Encoding Initiative). De asemenea, conţine şi alinierea celor mai mult de 230 de perechi de limbi conţinute de JRC-Acquis. Corpusul creşte pe măsura traducerii legislaţiei europene şi în limbile noilor candidaţi.

„UE Acquis Communautaire" este termenul prin care se face referire la corpul comun de legi şi obligativităţi care leagă toate statele membre ale Comunităţii Europene. Acquis-ul conţine principii şi obiective politice ale diverselor tratate semnate în cadrul Uniunii Europene (UE), legislaţie UE, declaraţii şi rezoluţii, acorduri internaţionale şi obiective comune. Toate ţările acceptate în Uniunea Europeană trebuie să ratifice „Acquis Communautaire". Pe lângă cele 22 de limbi ale Comunităţii Europene, Acquis este tradus şi în limbile croată şi turcă. Datorită efortului depus la ICIA (Institutul de Cercetări pentru Inteligenţa Artificială, Academia Română) pentru colectarea şi

adnotarea documentelor Acquis-ului românesc, limba română a fost prezentă în pachetul de distribuţie JRC-Acquis încă de la prima versiune a acestuia.

Numărul de documente comune în perechea de limbi Engleză-Română este de 11 469. Documentele comune constituie un important corpus paralel conţinând 59 986 838 de cuvinte.

Tabelul 1: JRC-Acquis versiunea 3.0 şi noua versiune de corpus românesc

Limbă	Documente	Caractere	Cuvinte
bulgară	11384	104522671	30146967
cehă	21438	148972981	46832312
daneză	23624	213468135	50944626
germană	23541	232748675	50929652
greacă	23184	239583543	55887003
engleză	23545	210692059	55537910
spaniolă	23573	238016756	62132608
estoniană	23541	192700704	40953424
finlandeză	23284	212178964	40107981
franceză	23627	234758290	62100432
ungară	22801	213804614	46188364
italiană	23472	230677013	57217002
lituaniană	23379	199438258	44392842
letonă	22906	196452051	44703607
malteză	10545	128906748	37883562
olandeză	23564	231963539	56771856
poloneză	23478	214464026	49253537
portugheză	23505	227499418	59606203
română	**19211**	**182631277**	**30832212**
slovacă	21943	179920434	46211035
slovenă	20642	178651767	47643215
suedeză	20243	199004401	46974192
Total	**476430**	**4288962348**	**1053305415**

Pentru indexarea lor, documentele din JRC-Acquis au fost clasificate manual cu ajutorul unui sistem de clasificare (EUROVOC) conţinând peste 6 000 de descriptori organizaţi ierarhic. Versiunea 4.2 a EUROVOC este disponibilă în 21 de limbi ale ţărilor din UE printre care şi limba română. În cadrul Eurovoc-ului, termenii se împart în două categorii: descriptori şi non-descriptori. Descriptorii sunt cuvinte sau expresii care denotă concepte din domeniile tezaurului într-un mod ne-ambiguu, pe când non-descriptorii sunt cuvinte sau expresii reprezentate deja în tezaur de un descriptor echivalent. Versiunea în limba engleză conţine 6 645 de descriptori iar, în comparaţie, tezaurul în limba română conţine doar 4 625 (aproximativ 70% din cel englezesc). Aceşti descriptori sunt organizaţi în 21 de domenii (de la politică şi relaţii internaţionale până la mediu, industrie sau geografie) ce conţin la rândul lor micro-tezaure. Există un total de 519 micro-tezaure (în română doar 508), fiecare din aceştia constituind un arbore în nodurile căruia se găsesc descriptori. Domeniile şi micro-tezaurele au

identificatori unici, independenţi de limbă, asigurându-se astfel o inter-relaţionare multilingvă.

2. Colectarea şi convertirea documentelor JRC-Acquis-Ro

Fişierele româneşti şi bulgăreşti nu sunt disponibile în acelaşi format (HTML) ca şi documentele celorlalte limbi din Acquis, neputând fi procesate de aceleaşi instrumente de convertire HTML-TEI. Fişierele în format HTML disponibile pentru celelalte limbi ale JRC-Acquis conţin şi informaţii cu privire la structura documentului, cum ar fi secţiunile de anexe şi semnături, secţiunile cu textul şi titlul documentului etc. Această structură nu se regăseşte în formatul Microsoft Word, format în care sunt disponibile documentele româneşti ale JRC-Acquis.

Pentru a constitui colecţia de documente în limba română, fişierele au fost descărcate de pe situl „CCVista Translation Database" folosind drept adresă „http://ccvista.taiex.be/Fulcrum/CCVista/RO/<celex>" unde <celex> este numărul unic de identificare al documentului. Numărul total de documente în limba română disponibile pe situl CCVista este de 19 286.

Fişierele au fost convertite din formatul Microsoft Word în formatul XML conform specificaţiilor TEI. Conversia celor 19 286 fişiere a fost făcută automat fiind folosit pachetul de funcţii „Visual Studio Tools for Office". Aceste funcţii permit interacţiunea directă cu aplicaţia Microsoft Office direct din mediul de programare. Datorită particularităţilor formatului, conversia documentelor a implicat şi o serie de etape intermediare:

- au fost înlăturate comentariile traducătorilor;
- au fost şterse notele de subsol şi secţiunile de cap de pagină;
- a fost normalizată folosirea caracterelor diacritice (unele documente foloseau „ş" şi „ţ" cu cedil iar altele foloseau „ş" şi „ţ" cu virgulă).

Dintre cele 19 286 de fişiere în format Microsoft Word au fost convertite 19 211 (restul de documente având erori de format).

În formatul TEI-XML al documentelor româneşti au fost adăugate şi datele de indexare EUROVOC acolo unde acestea erau disponibile.

3. Comparaţie între corpusul JRC-Acquis-Ro şi alte corpusuri româneşti

Pentru a compara diferenţa de vocabular între domeniul legislativ şi alte domenii am folosit corpusul românesc Agenda (colecţie de articole din săptămânalul timişorean *Agenda*). Acest corpus conţine 8 408 185 de cuvinte. După cum se observă din tabelul de mai jos, datorită domeniilor diferite abordate în cele două corpusuri, listele primelor cuvinte-conţinut, ordonate după rangul de frecvenţă, diferă într-o proporţie considerabilă.

Tabelul 2: Primele 25 de cuvinte, sortate în funcţie de ocurenţă, din corpusurile Jrc-Acquis-Ro şi Agenda

Jrc-Acquis Ro		**Agenda**	
art.	Yn	Timişoara	Np
articolul	Ncmsry	este	Vmip3s
nr.	Yn	ora	Ncfsry
membre	Ncfp-n	Timiş	Np
regulamentul	Ncmsry	ani	Ncmp-n
alin.	Yn	pot	Vmip3p
privind	Vmg	România	Np
trebuie	Vmip3s	sunt	Vmip3p
statele	Ncfpry	zona	Ncfsry
este	Vmip3s	mare	Afpfsrn
regulament	Ncms-n	vând	Vmip3p
vedere	Ncfsrn	privind	Vmg
având	Vmg	data	Ncfsry
CEE	Np	fost	Vmp--sm
comisiei	Ncfsoy	perioada	Ncfsry
consiliului	Ncmsoy	an	Ncms-n
prezentul	Afpmsry	apartament	Ncms-n
directiva	Ncfsry	piaţa	Ncfsry
ce	Np	persoane	Ncfp-n
comisia	Ncfsry	anul	Ncmsry
prezenta	Afpfsry	poate	Vmip3s
în_special	Rgp	muncă	Ncfsrn
comunităţii	Ncfsoy	astfel	Rgp
prevăzute	Vmp--pf	are	Vmip3s
membru	Ncms-n	Bucureşti	Np

Aplicaţiile de procesarea limbajului natural care vor avea la bază corpusul JRC-Acquis-Ro trebuie să ia în considerarea şi zgomotul pe care un astfel de corpus îl conţine. Experimentele noastre ne-au arătat că pentru un lexicon de peste 1 200 000 de cuvinte (incluzând aici şi entităţi denumite), în corpusul JRC-Acquis-Ro încă mai găsim foarte multe cuvinte necunoscute - din 30 832 212, 2 796 473 sunt cuvinte necunoscute.

Tabelul 3: Primele 20 de cuvinte necunoscute din JRC-Acquis-Ro

CEE	126951	NC	9872
ex	30926	and	9522
Amtsgericht	25646	BCE	8510
JO	25632	THE	8223
see	22721	the	8201
year	17553	en	4965
please	13805	comarca	4910
pct.	12978	del	4464
EUR	12870	Euratom	4455
from	10951	CECO	3875

Cuvintele necunoscute sunt, în marea lor majoritate, cuvinte aparţinând altor limbi – multe din pasajele din JRC-Acquis-Ro sunt copii din originalul limbii din care au fost

traduse. În tabelul 3 sunt prezentate primele 20 de cuvinte necunoscute ordonate după rangul de frecvenţă.

Pentru a observa diferenţele dintre JRC-Acquis-Ro şi alte corpusuri care au la bază „Acquis Communautaire" am testat modelele de limbă construite folosind SRILM (Stolcke, 2002) din JRC-Acquis-Ro, DGT-TM (Directorate-General for Translation – Translation Memory – http://langtech.jrc.it/DGT-TM.html) şi SEEERANET (Tufiş et. al., 2008). Modelele de limbă sunt de ordinul 3 şi au fost antrenate folosind forma de ocurenţă a cuvintelor. Textul pe care a fost evaluată perplexitatea este unul din fişierele JRC-Acquis-Ro (4 271 de cuvinte). În tabelul 4 se poate observa că deşi corpusurile diferă mult considerând modul în care au fost colectate, perplexitatea raportată pe fişierul de evaluare se îmbunătăţeşte în funcţie de mărimea acestora. Numărul de cuvinte necunoscute este un indicator al uni-gramelor neîntâlnite în corpusul de antrenament.

Tabelul 4: Evaluarea modelelor de limbă ale corpusurilor JRC-Acquis-Ro, DGT-TM Ro şi SEEERANET Ro

	JRC-Acquis Ro	**DGT-TM Ro**	**SEEERANET Ro**
Număr de cuvinte în corpus	30 832 212	2 528 584	1 442 915
Cuvinte necunoscute	79	218	362
Perplexitate	141	174	262

4. Concluzii

Corpusul JRC-Acquis-Ro, cu peste 30 de milioane de cuvinte, este cel mai mare corpus monolingual şi multilingual disponibil pentru limba română. Încă de la lansarea sa, în 2005, odată cu prima versiune a JRC-Acquis, corpusul JRC-Acquis-Ro este folosit în experimente de traducere automată şi pentru dezvoltarea sistemelor de întrebare-răspuns multilinguale, devenind un corpus de referinţă pentru limba română.

Deşi mărimea sa îl recomandă pentru orice aplicaţie de procesare a limbajului natural pentru limba română, corpusul necesită o etapă de filtrare pentru a fi înlăturate paragrafele din altă limbă decât româna. De asemenea, trebuie avută în vedere limitarea strictă la discursul juridic, limitare care nu permite o generalizare a fenomenelor de limbă observate în corpus.

Referinţe bibliografice

Steinberger Ralf, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, Dániel Varga (2006). *The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages*. Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006). Genoa, Italy, 24-26 May 2006

Germann Ulrich (ed.) (2001). *Aligned Hansards of the 36th Parliament of Canada* - Release 2001-1a. http://www.isi.edu/natural-language/download/hansard/

Koehn Philipp (2005). *EuroParl: A Parallel Corpus for Statistical Machine Translation*. Machine Translation Summit 2005. Phuket, Thailand. http://people.csail.mit.edu/koehn/publications/europarl/

Resnik Philip, Mari Broman Olsen & Mona Diab (1999). *The Bible as a Parallel Corpus: Annotating the ‚book of 2000 Tongues'*. Computers and the Humanities, 33(1-2), pp. 129-153. http://www.umiacs.umd.edu/users/resnik/

Andreas Stolcke. (2002). "SRILM - An Extensible Language Modeling Toolkit", in *Proc. Intl. Conf. Spoken Language Processing*, Denver, Colorado, September 2002

Erjavec Tomaž (2004). MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'2004)*, pp. 1535-1538, Paris. http://nl.ijs.si/ME/CD/docs/1984.html.

Dan Tufiş, Svetla Koeva, Tomaž Erjavec, Maria Gavrilidou, Cvetana Krstev. (2008). Building language resources and translation models for machine translation focused on South Slavic and Balkan languages, *FASSBL 2008: The Sixth International Conference Formal Approaches to South Slavic and Balkan Languages*, Dubrovnik, Septembrie 25-28, 2008

EXPERIMENTE DE TRADUCERE AUTOMATĂ BAZATĂ PE EXEMPLE PENTRU LIMBILE ENGLEZĂ/ROMÂNĂ

IRIMIA ELENA

Institutul de Cercetări pentru Inteligenţă Artificială, Academia Română, Bucureşti

elena@racai.ro

Rezumat

Acest articol descrie arhitectura unui sistem de traducere automată bazată pe exemple care a fost implementat în procesul de cercetare doctorală a autorului. Aplicaţia nu este neapărat inovativă în cadrul mai larg al paradigmei EBMT (Example Based Machine Translation), ci reprezintă un experiment care a dorit să combine în mod eficient o parte dintre resursele si instrumentele pentru prelucrarea automată a limbajului natural dezvoltate la ICIA cu tehnici şi algoritmi consacraţi în domeniul traducerii automate.

1. Introducere

Orice aplicaţie de traducere automată este o întreprindere care necesită investirea a importante cantităţi de timp şi energie. De aceea este util ca o astfel de aplicaţie să se construiască pe fundaţia muncii unei echipe de lucru (cum este grupul de PLN de la ICIA), care să poată pune la dispoziţie instrumentele şi resursele indispensabile. Articolul descrie in detaliu aplicaţia până la stagiul actual de implementare şi prezintă pe scurt obiectivele încă neatinse. Din lipsă de spaţiu, numărul exemplelor este foarte redus iar rezultatele sunt prezentate doar cantitativ şi evaluate sumar.

2. Resurse lingvistice utilizate şi aplicaţii de preprocesare ale acestora

Ca resursă fundamentală pentru aplicaţia de traducere bazată pe exemple pe care am implementat-o am ales corpusul paralel multilingv **JRC-Acquis** (Steinberger et al., 2006). Am considerat foarte potrivit faptul că acest corpus este: **omogen** - dedicat unui domeniu specific (cu un conţinut de natură juridică); **consistent** - cel puţin în teorie, orice expresie juridică din corpus trebuie să fie tradusă întotdeauna în acelaşi fel, într-o manieră validată de comunitatea profesioniştilor în domeniu; **actual**: JRC-Acquis este o colecţie dinamică de documente juridice extrase din *Acquis Communautaire* (AC), care reprezintă corpul total de legi ale Uniunii Europene aplicabile în toate ţările membre UE. AC, şi implicit JRQ-Acquis, se îmbogăţeşte constant cu noi documente, pe măsură ce Uniunea Europeană se extinde şi ţările membre îşi aliniază legislaţia la cea comunitară. JRC-Acquis este disponibil în 22 dintre cele 23 de limbi oficiale ale Uniunii Europene (traducerile irlandeze nu sunt încă disponibile) şi reprezintă cel mai mare corpus paralel existent în acest moment, atât ca dimensiune cât şi ca număr de limbi implicate. În forma utilizată de aplicaţia pe care am construit-o, corpusul conţine doar perechea de limbi română-engleză şi este rezultatul unor acţiuni consecutive de procesare: segmentare şi aliniere la nivel de propoziţie, segmentare la nivel de cuvânt, analiză morfo-sintactică, lematizare, adnotare sintactică de suprafaţă (chunking), aliniere lexicală şi analiză a dependenţelor sintactice între cuvinte. Documentele sunt

codificate XML conform DTD-ului (atributele acestuia captează toate adnotările produse de aplicaţiile de pre-procesare, cu excepţia alinierilor lexicale care sunt furnizate într-un fişier separat):

```
<!DOCTYPE text [
       <!ELEMENT text (body)>
       <!ATTLIST text id CDATA #REQUIRED>
   <!ELEMENT body (tu+)>
   <!ELEMENT tu (seg+)>
   <!ATTLIST tu id CDATA #REQUIRED>
   <!ELEMENT seg (s)>
   <!ATTLIST seg lang (en | ro) #REQUIRED>
   <!ELEMENT s (w | c)+>
   <!ATTLIST s id ID #REQUIRED>
   <!ELEMENT c (#PCDATA)>
   <!ELEMENT w (#PCDATA)>
   <!ATTLIST w
              ana CDATA #REQUIRED
              lemma CDATA #REQUIRED
      chunk CDATA #IMPLIED
      wns CDATA #IMPLIED
      head CDATA #IMPLIED
   >
]>
```

Strategia de extragere de exemple de traducere din corpusul JRC-Acquis pe care am implementat-o se bazează pe existenţa unor alinieri la nivel de cuvânt între 2 propoziţii pereche, precum şi pe adnotarea acestora cu dependenţe sintactice; aceste proceduri sunt asigurate de aplicaţiile următoare (implementate în cadrul ICIA):

- **YAWA**. Presupunând că avem de-a face cu o propoziţie p_1 într-o limbă l_1 şi traducerea ei p_2 în limba l_2, o aliniere lexicală presupune stabilirea de corespondenţe între cuvintele din p_1 şi cele din p_2 astfel încât acestea să reprezinte traduceri reciproce. Într-o formă uşor prelucrabilă de către alte aplicaţii, structura de alinieri din Figura 1 este reprezentată de către YAWA prin lista: {(1,1), (2,2), (3,3), (4,4), (5,5), (6,6), (7,7), (8,8). (9,8). (10,9), (11,10), (13,10), (12,11), (14,12), (15,13), (17,13), (16,14), (18,15)}. Elementele acestei liste sunt perechi de poziţii din cele două propoziţii care sunt asociate prin corespondenţa de traducere. Dacă un cuvânt dintr-o propoziţie nu are corespondent în propoziţia echivalentă, atunci absenţa acestuia se marchează prin cifra 0.

- LexPar (Ion, 2007) este o aplicaţie ce se bazează pe Modelul de Atracţie Lexicală a lui Yuret (MAL, (Yuret, 1998)) pentru a analiza legăturile sintactice între cuvinte. În viziunea lui Yuret, atracţia lexicală este o măsură a afinităţii de combinare a două cuvinte într-o propoziţie. Dacă două cuvinte sunt "atrase lexical" într-o propoziţie, atunci probabilitatea ca ele să se combine şi în alte contexte este semnificativă. De aceea, două sau mai multe cuvinte care se atrag lexical, împreună cu traducerile lor într-o altă limbă, se constituie într-un exemplu bun de traducere. Prin intermediul aplicaţiei LexPar, corpusul de lucru este adnotat cu atributul "head", a cărui valoare (un număr întreg) reprezintă poziţia unui cuvânt din propoziţie de care se „leagă" printr-o dependenţă sintactică forma adnotată. De exemplu, pentru Figura 1, prezenţa în corpus a atributului *head=„3"* pentru cuvântul "payments" indică faptul că „payments" se leagă de "all", aflat în poziţia 3 în propoziţie. Atributul „head" nu indică centrul

constituentului sintactic şi sensul relaţiei de dependenţă dintre cele două cuvinte, ci doar faptul că aceste cuvinte sunt legate între ele din punct de vedere sintactic.

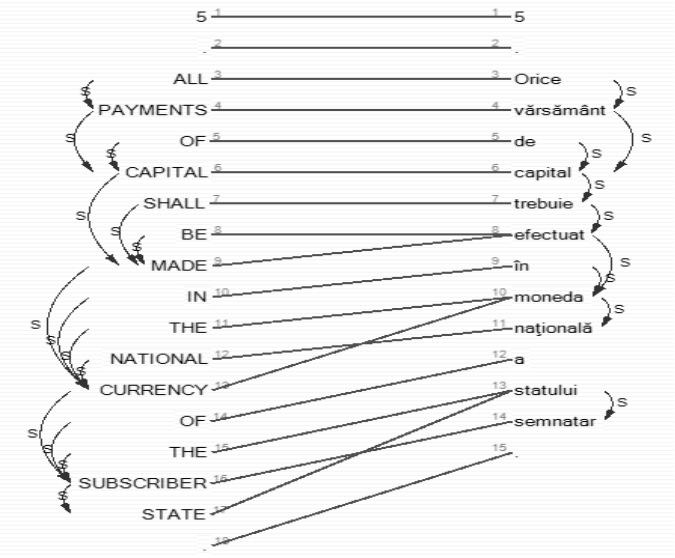

Figura 1. Vizualizarea alinierilor şi legăturilor pentru o unitate de traducere din corpus. Numerele reprezintă poziţiile cuvintelor în propoziţie. Corespondenţele de traducere sunt marcate prin linii. O săgeată marchează existenţa unei legături de dependenţă sintactică între cele două cuvinte pe care le uneşte. Sensul săgeţii este irelevant pentru sensul legăturii de dependenţă.

3. Baza de date cu exemple

În construcţia bazei de date cu exemple am dorit să pornim de la ceea ce Yuret numea atracţie lexicală sau afinitate de combinare între cuvinte. Am menţionat deja că această atracţie lexicală creşte probabilitatea de asociere a două sau mai multe cuvinte şi în alte contexte decât cel în care s-au identificat legăturile, ceea ce face ca secvenţa respectivă de cuvinte să se constituie într-un exemplu de traducere mai bun decât o simplă n-gramă. De asemenea, atunci când descompunem propoziţia de tradus în subsecvenţe care se suprapun peste baza de date folosim acelaşi concept de atracţie lexicală pentru a decide graniţele subsecvenţelor. Se ştie că pentru extragerea unui exemplu de traducere nu este îndeajuns să stabilim o strategie de divizare a propoziţiilor în subsecvenţe de cuvinte, ci este necesară stabilirea de corespondenţe între subsecvenţele dintr-o propoziţie şi traducerile lor în propoziţia echivalentă. În acest scop, vom utiliza alinierile lexicale produse de YAWA .

Revenind la exemplul din Figura 1, se observă că legăturile trasate cu LexPar tind să se grupeze prin imbricare şi să descompună propoziţia prin înlănţuire. Aceste proprietăţi sugerează mai multe descompuneri posibile ale propoziţiei şi implicit extragerea unor subsecvenţe de lungimi diferite, dar care să fie compuse din cuvinte între care există fenomenul de atracţie lexicală de care am vorbit.

Am denumit *superlegătură* un vector de forma $S = (poziţie_1, ..., poziţie_s)$ unde:
- $poziţie_i$, cu $i \in [1,s]$ reprezintă poziţia unui cuvânt într-o propoziţie dată P;
- *s* este lungimea lui S;
- există un vector de legături de forma [($poziţie_1$, $poziţie_2$), ($poziţie_1$, $poziţie_3$),..., ($poziţie_1$, $poziţie_s$)] *sau* un vector de legături de forma [($poziţie_1$, $poziţie_s$), ($poziţie_2$, $poziţie_s$),..., ($poziţie_{s-1}$, $poziţie_s$)] ce caracterizează secvenţa de cuvinte descrisă de

poziţiile din S; deoarece legăturile care se intersectează sunt filtrate de LexPar, o superlegătură nu poate avea decât o formă imbricată precum cea din Figura 1.
Am denumit *lanţ* un vector de forma L = *(poziţie$_1$, ..., poziţie$_l$)* unde:
- *poziţie$_i$,* cu $i \in [1,l]$ reprezintă poziţia unui cuvânt într-o propoziţie dată P;
- *l* este lungimea lui L;
- există un vector de legături de forma [(poziţie$_1$, poziţie$_2$), (poziţie$_2$, poziţie$_3$),..., (poziţie$_{s-1}$, poziţie$_s$)] ce caracterizează secvenţa de cuvinte descrisă de poziţiile din L.
Trebuie remarcat că o pereche de poziţii (poziţie$_i$, poziţie$_{i+1}$) nu reprezintă în mod obligatoriu poziţii consecutive în propoziţie.

Extractorul de exemple de traducere pe care l-am construit, *ExTract*, primeşte ca date de intrare corpusul de lucru precum şi un fişier care conţine pentru fiecare propoziţie din corpusul de lucru, alinierea lexicală asociată de YAWA Vom descrie modul de procesare a unei singure unităţi de traducere din corpus *U*, care se repetă până la prelucrarea întregului document. O unitate de traducere din corpusul de lucru este alcătuită dintr-o propoziţie în limba engleză *Pen* şi traducerea ei *Pro* în limba română, împreună cu toate adnotările ce rezultă în urma prelucrărilor descrise în capitolul 2. Aplicaţia lucrează în două etape:

Etapa 1. Se construiesc superlegăturile şi lanţurile posibile atât pentru *Pen* cât şi pentru *Pro*. Procedeul este independent de limbă, deci putem simplifica descrierea acestuia prin generalizare. Fie *P* o propoziţie oarecare dintre cele două propoziţii ale unei unităţi de traducere. Informaţia despre atracţia lexicală între cuvintele lui *P* este conţinută de către atributul „head". Pentru a face această informaţie accesibilă unor prelucrări ulterioare, construim vectorul de legături *L,* care, similar formalizării alinierilor din secţiunea 3.3. conţine *perechi de poziţii* ale unor cuvinte. Elementele unei perechi de poziţii din *L* sunt însă poziţii ale unor cuvinte din *aceeaşi propoziţie (P)* iar relaţia care le aduce împreună în aceeaşi pereche este cea de atracţie lexicală între cuvintele pe care le indexează. Vectorul de legături se construieşte parcurgând propoziţia *P* cuvânt cu cuvânt iar pentru fiecare cuvânt care deţine atributul „head", se introduce în *L* perechea *(poziţie_cuvânt, valoare_atribut_head)*. De exemplu, pentru propoziţia în limba engleză din Figura 1 vectorul *L* conţine perechile: (3,4), (4,6), (5,6), (6,9), (7,9), (8,9), (9,13), (10.13), etc. Vectorul rezultat al etapei 1) este similar vectorului *L,* dar poate conţine nu doar perechi de poziţii, ci liste de dimensiune variabilă care păstrează proprietatea de atracţie lexicală între cuvintele pe care le indexează în P. Voi denumi acest vector *Lfinal* iar popularea lui cu elemente are loc în doi paşi:

Pasul 1. Identificarea superlegăturilor şi introducerea lor în *Lfinal*.

Lfinal ← *null;*

Pentru fiecare p, unde *p* este o poziţie în *P*

{ *Pentru fiecare* pereche (x,x') din *L*

{ *Dacă* ((x=p) *sau* (x' = p))

{ Lista ← *null;*

Dacă ((x' – x)>1 şi (x' – x)<=4) Lista ← x, x+1,…,x-1,x *//completează poziţiile care lipsesc;*

Altfel Lista ← x,x'*;*

Dacă (Lista $\neq$ *null* şi Lista $\notin$ *Lfinal*) *Lfinal* $\leftarrow$ Lista} } }

Observaţie: În cazul în care avem de a face cu o legătură la distanţă mare, probabilitatea ca aceasta să reprezinte o legătură simplă şi nu un indiciu pentru o superlegătură creşte. De exemplu, în propoziţia *„(1)Dacă (2)la (3)încheierea (4)exerciţiului (5)financiar (6)se (7)constată..."*, o legătură de tipul (1,7) nu atrage după sine existenţa unor legături care să producă superlegătura (1,2,3,4,5,6,7). Algoritmul de extragere a superlegăturilor a fost gândit pe baza observaţiilor legăturilor din corpusul de lucru şi reflectă proprietăţile sintactice ale limbilor implicate. Pragul peste care o legătură nu poate fi tratată de către algoritm ca posibilă superlegătură este 4.

Pasul 2. Identificarea lanţurilor şi introducerea lor în *Lfinal*

Pentru identificarea unor subsecvenţe de tip lanţ în care să descompunem propoziţia P, vom parcurge *Lfinal* de la cap la coadă efectuând următoarele operaţii:

- *fie* l_i *un element din Lfinal; din lista de elemente* l_{i+1}, ..., l_n *ce i se succed, alegem prima listă l' care verifică proprietăţile* $l_i \cap l' \neq \phi$ *şi* $l' - l_i \equiv \phi$ *;*

- *lanţ* $\leftarrow l_i \cup l'$ *;*

- *Lfinal* $\leftarrow$ *lanţ;*

- *Repetăm operaţiile 1), 2) şi 3) substituind* l_i *cu lanţ;*

Astfel, vom construi *lanţuri* formate din legături simple şi superlegături, limitând numărul de elemente care se înlănţuiesc la 3. Această limitare, care este motivată de dorinţa de a nu supraîncărca baza de date în mod inutil, a fost stabilită tot pe baza observaţiilor datelor din corpusul de lucru şi reflectă faptul că probabilitatea ca un cuvânt să fie atras lexical de un cuvânt aflat la o distanţă mai mare de 2 verigi ale unui lanţ este foarte mică. De asemenea, proprietăţile morfologice precum acordul între substantiv şi verb sau între substantiv şi adjectiv se transmit arareori la distanţe care să nu fie acoperite de un lanţ cu 3 elemente. De exemplu, în Figura 1 acordul între substantivul "vărsământ" şi adjectivul "efectuat" este surprins de lanţul cu 3 verigi *(vărsământ de capital, trebuie, efectuat)*.

Etapa 2. Presupunând că am calculat (conform Etapei 1) cei doi vectori $Lfinal_{en}$ şi $Lfinal_{ro}$, în acest moment trebuie să stabilim corespondenţe între elementele acestora pentru a construi exemplele de traducere. Fie *A* vectorul de alinieri asociat unităţii de traducere *U*. Am specificat în secţiunea 3.3 că un astfel de vector conţine perechi de poziţii ale unor cuvinte, *(poziţie*$_i$*, poziţie*$_j$*)*, unde *poziţie*$_i$ reprezintă poziţia unui cuvânt w_i din *Pen*, *poziţie*$_j$ reprezintă poziţia unui cuvânt w_j din *Pro*, iar w_j este traducerea lui w_i. Structura care va formaliza corespondenţele dintre listele lui $Lfinal_{en}$ şi cele ale lui $Lfinal_{ro}$ este o listă *ET* de perechi de forma : $((p_1, p_2, ..., p_k),(p_1', p_2', ..., p_s'))$, unde $(p_1, p_2, ... p_k) \in Lfinal_{en}$,, $(p_1', p_2', ...p_s') \in Lfinal_{ro}$, iar pentru fiecare p_i există un p_j' astfel încât $(p_i,p_j') \in A$. Algoritmul de construire a listei *ET* este :

Pasul 1. Grupează toate alinierile *1:n* şi *n:1* din *A*. Introduce în *ET* toate alinierile 1:n şi n:1, precum şi alinierile din *A* care nu sunt implicate în alinierile multiple. Acest pas

este necesar deoarece se doreşte ca baza de date cu exemple să conţină şi echivalenţe de traducere la nivel lexical, care vor compensa absenţa unui dicţionar;

Pasul 2.

Pentru fiecare element listă l din $Lfinal_{en}$

{ Lista l' ← null;

Pentru fiecare element p al listei l

{ Extrage din A lista C de perechi (x,x') pentru care p = x;

Pentru fiecare pereche (p, x') din C, l' ← x'; }

Dacă l' $\in Lfinal_{ro}$, *ET ← (l,l') }*

Etapa 3 presupune o simplă recuperare a secvenţelor de cuvinte indexate de listele de poziţii din *ET* şi afişarea lor într-un fişier ale cărui intrări au forma descrisă la pagina 4. Informaţiile legate de lema şi eticheta MSD a cuvântului sunt extrase din atributele „lemma" şi „ana" cu care este adnotat corpusul de lucru. În plus, în această etapă, fiecărui exemplu de traducere i se asociază un scor de încredere după cum urmează: dacă cele două liste de poziţii (corespunzătoare membrului în limba engleză, respectiv membrului în limba română al exemplului de traducere) respectă condiţia de consecutivitate (nu există elemente p_i, p_{i+1} astfel încât $p_{i+1} - p_i > 1$) atunci scorul de încredere este 10; dacă cel puţina dintre cele două liste de poziţii nu respectă condiţia de consecutivitate (implică legături la distanţă), scorul de încredere este 5.

EXEMPLE DE TRADUCERE: LIMBA ENGLEZĂ	EXEMPLE DE TRADUCERE: LIMBA ROMÂNĂ
ALL(all,Di3)	Orice(orice,Di3--r---e)
PAYMENTS(payment,Ncnp)	vărsământ(vărsământ,Ncms-n)
OF(of,Sp)	de(de,Spsa)
CAPITAL(capital,Ncns)	capital(capital,Ncms-n)
SHALL(shall,Vaip)	trebuie(trebui,Vmip3s)
BE(be,Van) MADE(make,Vmps)	efectuat(efectua,Vmp--sm)
THE(the,Dd) NATIONAL(national,Afp) CURRENCY(currency,Ncns)	moneda(monedă,Ncfsry) naţională (naţional,Afpfsrn)
ALL(all,Di3) PAYMENTS(payment,Ncnp)	Orice(orice,Di3--r---e) vărsământ (vărsământ,Ncms-n)
OF(of,Sp) CAPITAL(capital,Ncns)	de(de,Spsa) capital(capital,Ncms-n)
PAYMENTS(payment,Ncnp) OF(of,Sp) CAPITAL(capital,Ncns)	vărsământ(vărsământ,Ncms-n) de(de,Spsa) capital(capital,Ncms-n)
SHALL(shall,Vaip) BE(be,Van) MADE(make,Vmps)	trebuie(trebui,Vmip3s) efectuat(efectua,Vmp--sm)
IN(in,Sp) THE(the,Dd) NATIONAL(national,Afp) CURRENCY(currency,Ncns)	în(în,Spsa) moneda(monedă,Ncfsry) naţională(naţional,Afpfsrn)
THE(the,Dd) SUBSCRIBER(subscriber,Ncns) STATE(state,Ncns)	statului(stat,Ncmsoy) semnatar(semnatar,Ncms-n)
PAYMENTS(payment,Ncnp) OF(of,Sp) CAPITAL(capital,Ncns) SHALL(shall,Vaip) BE(be,Van) MADE(make,Vmps)	vărsământ(vărsământ,Ncms-n) de(de,Spsa) capital(capital,Ncms-n) trebuie(trebui,Vmip3s) efectuat(efectua,Vmp--sm)
NATIONAL(national,Afp) CURRENCY(currency,Ncns) OF(of,Sp) THE(the,Dd) SUBSCRIBER(subscriber,Ncns)	moneda(monedă,Ncfsry) naţională(naţional,Afpfsrn) a(al,Tsfs) statului(stat,Ncmsoy) semnatar(semnatar,Ncms-n)
PAYMENTS(payment,Ncnp) OF(of,Sp) CAPITAL(capital,Ncns) SHALL(shall,Vaip) BE(be,Van) MADE(make,Vmps) IN(in,Sp) THE(the,Dd) NATIONAL(national,Afp)	vărsământ(vărsământ,Ncms-n) de(de,Spsa) capital(capital,Ncms-n) trebuie(trebui,Vmip3s) efectuat(efectua,Vmp--sm) în(în,Spsa) moneda(monedă,Ncfsry) naţională(naţional,Afpfsrn)

CURRENCY(currency,Ncns)	

Tabelul 1. O parte dintre rezultatele ExTract pentru unitatea de traducere din Figura 3.2.

Corpusul de lucru a fost împărţit în corpusul de date pentru extracţie de exemple (99% din corpusul total) şi corpusul de date de test (1% din corpusul total). După rularea *ExTract pe* unităţile de traducere rezervate construirii bazei de date cu exemple, rezultatele au fost numărate şi a rezultat un fişier cu 900.000 de exemple de traducere diferite, asociate frecvenţelor lor în corpus. *ExTract* este o aplicaţie de sine stătătoare şi nu face parte din fluxul de traducere. Construcţia bazei de date cu exemple se face o singură dată şi este urmată de o procedură de reorganizare a informaţiei în 5 fişiere diferite, conectate printr-un index comun, pentru ca procedura de matching să nu supraîncarce memoria şi să nu dureze foarte mult. Astfel, unei intrări din fişierul de ieşire al lui ExTract i se asociază un index de exemplu de traducere (*iet*) şi toate informaţiile conţinute de intrarea respectivă sunt distribuite în cele 5 fişiere după cum urmează: fişierul **en_forme:** *iet "enf_1 enf_2 ...enf_n" MD5("enf_1 enf_2 ...enf_n");* fişierul **en_leme**: *iet "enl_1 enl_2 ...enl_n" MD5("enl_1 enl_2 ...enl_n")*; fişierul **ro_forme**: *iet "rof_1 rof_2 ...rof_n" MD5("rof_1 rof_2 ...rof_n")*; fişierul **ro_leme**: *iet "rol_1 rol_2 ...rol_n" MD5("rol_1 rol_2 ...rol_n")*; fişierul **info**: *iet enm_1 enm_2 ... enm_n rom_1 rom_2... rom_3 frecvenţă, scor încredere;*

unde: ***MD5*** este o funcţie hash utilizată des în criptografie, care asociază unui şir de caractere un număr natural pe 16 octeţi, reprezentat în mod uzual ca o secvenţă de 32 de cifre hexazecimale; ***enl*** = lemă cuvânt în limba engleză, ***enf*** = formă cuvânt în limba engleză, ***enm*** = MSD cuvânt în limba engleză, **_1, _2, _n** = poziţiile cuvintelor în şirul extras în limba engleză; analog, ***rol*** = lemă cuvânt în limba română**,** ***rof*** = formă cuvânt în limba română, ***rom*** = MSD cuvânt în limba română, **_1, _2, _m** = poziţiile cuvintelor în şirul extras în limba română .

4. Suprapunerea propoziţiei de tradus peste baza de date cu exemple (matching-ul)

Datorită formei în care a fost organizată baza de date, procedura de matching devine una de căutare într-o listă de numere naturale şi este mult mai eficientă. De asemenea, separarea informaţiei legată de forma cuvintelor şi cea legată de lemă în fişiere diferite este utilă în etapa de matching, când aplicaţia poate urma una dintre următoare două direcţii:

1) suprapunere la nivel de formă ocurenţă: *pasul 1)* încarcă fişierul *en_forme*; *pasul 2)* descompune propoziţia de tradus *P* în fragmente utilizând algoritmii de identificare a superlegăturilor şi lanţurilor din secţiunea 4.1 (de fapt, etapa de descompunere este corespunzătoare Etapei 1 din secţiunea 4.1 aplicată membrului în limba engleză al unităţii de traducere şi produce un vector similar cu $Lfinal_{en}$); *pasul 3)* pentru fiecare dintre elementele vectorului de legături: recuperează din propoziţie secvenţa de forme pe care o indexează – *şir_forme* – şi calculează MD5(*şir_forme*); caută MD5(*şir_forme*) în lista MD5 din *en_forme* şi extrage indexul *iet* pentru toate intrările identificate astfel; un singur identificator MD5 poate aveam mai mulţi indecşi *iet* asociaţi, deoarece unui şir de forme în limba engleză îi pot corespunde mai multe

traduceri diferite în baza de date cu exemple; pentru fiecare *iet,* recuperează informaţia din *ro_forme* şi *info*.

2) suprapunere la nivel de lemă: se execută aceiaşi paşi ca în 1), înlocuind *en_forme* cu *en_leme, ro_forme* cu *ro_leme*, *şir_forme* cu *şir_leme*.

Pentru eficientizare, informaţia asociată unui exemplu de traducere candidat pentru traducerea finală se organizează într-un obiect din clasa *trans_ex*, cu următoarele proprietăţi: *iet, en_formă, en_lemă, ro_formă, ro_lemă, en_msd, ro_msd, frecvenţă, scor_încredere, md5, poziţie*: aceste atribute primesc valori în etapa de *matching*; atributul `poziţie` are ca valoare lista de poziţii din $Lfinal_{en}$ corespunzătoare listei de cuvinte `en_formă` şi distinge între secvenţe de cuvinte identice aflate în aceeaşi propoziţie, în poziţii diferite; *scor_aliniere, scor_traducere, lungime_suprapunere_en, lungime_suprapunere_ro, cel_mai_bun_scor_traducere*: aceste atribute primesc valori în procesul de *recompunere a propoziţiei* (vezi secţiunea 5). La finalul etapei de matching se construieşte o listă *Frg* de obiecte de tip *trans_ex* care reprezintă mulţimea tuturor fragmentelor în care P se poate descompune ce au fost găsite în baza de date cu exemple. Din această listă *Frg* se va recompune cea mai bună traducere posibilă a propoziţiei P în limba ţintă.

5. Etapa de recombinare şi adaptare

Pentru această etapă am ales Metoda Suprapunerii Maximale (Hutchinson et all, 2003), care combină fragmente care se suprapun şi ale căror traduceri sunt consistente". Autorii acestei metode exploatează intuiţia că, atunci când două exemple de traducere se suprapun atât la nivelul fragmentelor sursă cât şi la cel al fragmentelor ţintă, probabilitatea ca o combinaţie a acestor exemple să producă o traducere corectă este crescută. În accepţiunea algoritmului (Hutchinson et all, 2006), suprapunere a două şiruri înseamnă de fapt suprapunere la stânga, adică: două şiruri $s = \{w_1, w_2, \ldots w_n\}$ şi $s' = \{w_1', w_2', \ldots w_m'\}$ se suprapun dacă există un întreg $p<m,n$, astfel încât $w_1' = w_{n-p}$, $w_2' = w_{n-p=1}, \ldots, w_p' = w_n$.

Exemplu: (in this Regulation, în acest regulament), (this Regulation, prezentul Regulament): combinarea acestor exemple nu este indicată deoarece: 1. pentru fragmentul în limba engleză, condiţia se suprapunere este îndeplinită doar parţial – există un întreg p=2 care reprezintă lungimea suprapunerii „this Regulation", dar p=m; 2. pentru fragmentul în limba română condiţia de suprapunere nu este îndeplinită.

Combinarea exemplelor de traducere este ghidată de o funcţie de evaluare *s(E)* – E este un exemplu de traducere reprezentat ca un obiect din clasa *trans_ex*. *s(E)* este calculată ţinând cont doar de un alt exemplu de traducere, considerat a fi predecesorul lui E în soluţia finală şi depinde de următorii parametrii: *overlapp_length_en (respectiv overlapp_length_ro)* – lungimea suprapunerii între membrul în limba engleză (respectiv română) al lui E şi membrul în limba engleză (respectiv română) al predecesorului său; *length_en*: lungimea fragmentului în limba engleză al lui E (un fragment mai lung este preferat); *gap*: distanţa dintre primul cuvânt în fragmentul în limba engleză al lui E şi ultimul cuvânt al fragmentului în limba engleză din predecesorul lui E; *alignment*:

scorul de aliniere, calculat ca medie ponderată a frecvenţei, scorului de încredere şi a unui scor MSD dat de un model de traducere pe MSD-uri (scorul este calculat pentru perechea *(en_msd, ro_msd)* asociată lui E iar modelul este extras din fişierul *info*).

$$s(E) = g*gap + s'(E),$$

$$s'(E) = 1/(a*alingment + or* overlapp\_length\_ro + oe* overlapp\_length\_en + l* length\_en + 1).$$

Coeficienţii *g, a, oe, or* şi *l* sunt optimizaţi în mod experimental. Funcţia de evaluare totală *s(P)*, este aditivă pe mulţimea exemplelor de traducere care descompun *P*. Pentru a minimiza numărul de calcule, (Hutchinson et all, 2003) propun o *tehnică best-first cu rază limitată de acţiune*, expandând primul cel mai bun candidat neevaluat la un moment dat şi păstrând în memorie numai primii cei mai buni n candidaţi neevaluaţi. Algoritmul produce un vector de exemple de traducere (l-am numit *Soluţie*) a căror combinaţie ar trebui să formeze cea mai bună traducere posibilă pentru propoziţia P în condiţiile unei anumite baze de date cu exemple. Indicaţiile (Hutchinson et all, 2003) au fost implementate destul de exact, cu doar câteva modificări: s-a introdus ca parametru şi lungimea suprapunerii pentru fragmentele în limba română ; s-a fixat un coeficient $l = \sqrt{length\_en}$ care poate favoriza (mai bine decât un scalar) un exemplu mai lung, asigurând valori mai bune pentru funcţia de evaluare decât suma funcţiilor de evaluare a mai multor exemple scurte; s-a fixat o rază de acţiune *n= 40*.

O etapă finală de adaptare este necesară pentru transformarea informaţiei conţinute de vectorul *Soluţie* într-o propoziţie în limba ţintă. Atât pentru opţiunea de matching pe leme cât şi pentru cea de matching pe forme, adaptarea va implica eliminarea dublurilor produse de suprapunere, concatenarea secvenţelor de cuvinte şi câteva reguli de reordonare bazate pe secvenţele de MSD-uri asociate secvenţelor de cuvinte. Aceste proceduri au fost implementate, dar lista de reguli de rescriere poate fi extinsă. Pentru opţiunea de matching pe leme, este necesară şi integrarea unui mecanism de generare a formelor ocurenţă înainte de etapa de adaptare. Această etapă nu a fost încă implementată, dar se va baza pe resursele şi instrumentele dedicate generării morfologice dezvoltate la ICIA (vezi (Irimia, 2007) şi (Tufiş et al., 2008)).

6. Concluzii

Am rulat componentele de matching la nivel de formă, recombinare şi adaptare pe datele de test (600 de exemple de traducere) şi am calculat scorul BLEU în raport cu o singură traducere referinţă. Această evaluare este prematură (scorul BLEU este sensibil la numărul de traduceri referinţă), dar un rezultat de 0,232 reprezintă un bun motiv pentru a continua implementarea componentei de matching pe leme şi a concluziona printr-o evaluare şi o comparare a celor două direcţii. Analiza bazei de date cu exemple a condus la observarea unor erori sistematice produse de aliniatorul lexical (de exemplu, articolul hotărât în limba engleză este, adeseori, nealiniat cu substantivul în limba română sau aliniat incorect cu un alt substantiv decât cel corespunzător), a unor erori generate de LexPar (în special legături care nu pot fi captate) precum şi a unui mic număr de erori generate de adnotatorul morfologic. Testarea algoritmului de extragere a exemplelor pe un corpus de mici dimensiuni (200 de unităţi de traducere) corectat la

nivel de aliniere lexicală a produs exemple de traducere corecte în proporţie de 99%. Considerăm că o creştere a performanţelor instrumentelor de aliniere şi analiză a legăturilor poate îmbunătăţi semnificativ calitatea exemplelor din baza de date şi, implicit, a rezultatelor aplicaţiei de traducere. O altă soluţie evidentă este creşterea dimensiunilor corpusului din care se extrage baza de date, dar această abordare presupune scăderea vitezei de răspuns a aplicaţiei (spaţiu mai mare de căutare) şi necesită găsirea unor metode de optimizare a acesteia.

Referinţe bibliografice

Brants, Thorsten (2000). TnT – a statistical part-of-speech tagger. In *Proceedings of the 6th Applied NLP Conference*, ANLP-200, pages 224-231, Seattle, WA.

Erjavec T., R.Pavlov, L.Dimitrova, L.Sinapova, K.Simov, M.Tadi, V.Petkevi, HJ.Kaalep, N.Ide, G.Priest-Dorman, L.Tihanyi, T.Vradi, C.Oravecz D.Tufis, AM.Barbu, P Holozan, V Gorjanc, M. Stabej (2001). Specifications and Notation for MULTEXT-East Lexicon Encoding. *MULTEXT-East Report*, Concede Edition D.

Hutchinson Rebbeca, Paul N. Bennett, Jaime Carbonell, Peter Jansen, Ralf Brown (2003). *Maximal Lattice Overlap in Example-Based Machine Translation.* School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213

Ion, Radu (2007). *Word Sense Disambiguation Methods Applied to English and Romanian.* PhD thesis (in Romanian). Romanian Academy, Bucharest.

Irimia Elena (2007). ROG- a Parafigmatic Morphological Generator for Romanian. *Proceedings of the 3rd Language Conference: Human Languages Technologies as a Challenge for Computer Science and Linguistics*, Poznan, Poland, pages 408-412 ISBN 978-83-7177-407-2.

Steinberger Ralf, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, Dániel Varga (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. *Proceedings of the 5th International Conference on Language Resources and Evaluation* (Genoa, Italy).

Tufiş Dan, Elena Irimia, Radu Ion, Alexandru Ceauşu (2008). Unsupervised Lexical Acquisition for Part of Speech Tagging. In *Proceedings of LREC 2008*, May 26 - June 1, Marakkech, Morocco. ELRA - European Language Resources Association. ISBN: 2-9517408-4-0.

Yuret Deniz (1998). *Discovery of linguistic relations using lexical atrraction.* PhD thesis, Department of Computer Science and Electrical Engineering, MIT.

CONAN – DETECŢIA POSIBILELOR CONOTAŢII ALE UNUI TEXT

DAN ŞTEFĂNESCU, DAN TUFIŞ

Institutul de Cercetări pentru Inteligenţă Artificială
Academia Română

{danstef, tufis}@racai.ro

Rezumat

Ambiguitatea limbajului natural trece adesea neobservată, fiind de cele mai multe ori generată în mod involuntar. O propoziţie în cazul căreia intenţia autorului este evidentă în contextul original, poate fi înţeleasă total diferit atunci când este pusă într-un alt context, mai ales în situaţia în care ea conţine ambiguităţi neobservate. Uneori acest lucru poate fi amuzant, alteori stânjenitor. Lucrarea prezintă o aplicaţie pe care am numit-o CONAN, realizată la ICIA, cu ajutorul căreia astfel de ambiguităţi pot fi detectate şi, în funcţie de intenţia autorului, înlăturate, diminuate sau amplificate.

1. Introducere

Există diverse metode de modelare a proceselor de clasificare a opiniilor şi diferite grade de granularitate în definirea acestor modele. De exemplu, în cazul clasificării recenziilor se poate evalua opinia/aprecierea generală (pozitivă, negativă sau neutră) a autorului cu privire la un anumit subiect al discuţiei. Clasificarea opiniei la nivelul documentelor este însă considerată a fi prea puţin granulară în cazul majorităţii aplicaţiilor (de pildă un document poate exprima diferite opinii în raport cu diverse aspecte ale subiectului tratat - un produs poate fi excelent dar preţul său mult prea mare, o piesă de teatru poate avea o distribuţie foarte bună dar un scenariu prost, etc.). De aceea mai toate sistemele avansate de analiză/clasificare opiniilor iau în considerare nivelul propoziţional. La acest nivel, problemele tipice care apar includ identificarea propoziţiilor ce exprimă o opinie, a relevanţei acestei propoziţii faţă de subiectul de interes, a agentului care exprimă opinia (care poate fi autorul textului sau o sursă citată de autor), polaritatea (pozitivă, negativă sau neutră) opiniei precum şi intensitatea ei (puternică, slabă).

Toate metodele sau algoritmii (abia la început) folosiţi pentru analiza subiectivităţii exploatează cuvinte sau expresii deja procesate, acestea fiind unităţi lexicale purtătoare de opinie sau sentiment. Aceste unităţi lexicale (pe care le vom numi senti-cuvinte) sunt în general manual codificate, extrase din corpusuri sau marcate în lexicoane precum General Inquirer sau SentiWordNet (Esuli & Sebastiani, 2006). În SentiWordNet fiecare sens al fiecărui cuvânt are asociat un marcaj <O, P, N> în care O, P şi N reprezintă scoruri apriori (independente de context) de obiectivitate (O), subiectivitate pozitivă (P) şi respectiv subiectivitate negativă (N) şi în plus, pentru oricare triplet <O, P, N> ce marchează înţelesurile din dicţionar există relaţia O+P+N=1.

În timp ce stabilirea faptului că o propoziţie are caracter subiectiv (exprimă o opinie) şi a detectării faptului că această propoziţie se referă la subiectul de interes al investigaţiei sunt mai puţin controversate, polaritatea ei poate fi în schimb problematică. Dificultatea

este generată de polisemia majorităţii cuvintelor şi de faptul că în multe cazuri polaritatea subiectivităţii apriori (engl. "*prior subjectivity*") a senti-cuvintelor depinde de sensul lor contextual (uneori local, alteori global). Aparent, noţiunea de sens, aşa cum este ea definită în cadrul SentiWordNet, ar rezolva problema. În realitate însă nu este aşa. După cum arată (Tufiş, 2008a), este necesară stabilirea unei distincţii clare între cuvintele intrinsec purtătoare de subiectivitate cu polaritate specifică, şi cuvintele a căror polaritate ar trebui luată în considerare în funcţie de context. Cazul al doilea face referire la situaţii de felul: „timpul de răspuns al monitorului este *lung*" comparativ cu „viaţa *lungă* a unui motor". Se poate observa aici că polaritatea cuvântul *lung* depinde de substantivul pe care îl modifică.

Cercetările în această direcţie au fost până nu demult de tip monolingv, orientate în marea lor majoritate către limba engleză datorită bogatelor resurse existente, necesare acestor tipuri de analiză[1] (Mihalcea et al., 2007). În ultimii ani însă, pentru tot mai multe limbi (printre care şi limba română) au fost dezvoltate astfel de resurse, în general prin exploatarea textelor paralele şi a lexicoanelor multilingve. O ipoteză fundamentală în transferul cros-lingual al caracterizării apriorice de subiectivitate a unui sens al unui element lexical dintr-o limbă este că aceasta (caracterizarea) este validă şi pentru sensul corespunzător al echivalentului de traducere într-o altă limbă. După cunoştinţele noastre nu există nici un experiment care să infirme această ipoteză, cel puţin la nivelul calitativ. Desigur, există posibilitatea ca raportul dintre subiectivitatea apriori pozitivă şi cea negativă să varieze între echivalenţii de traducere pentru o pereche de limbi, dar acest subiect necesită cercetări şi experimente care depăşesc cadrul investigaţiei noastre prezente. Deocamdată, adoptând ipoteza corespondenţei subiectivităţii apriori a echivalenţilor lexicali de traducere, metoda transferului cros-lingual al informaţiei de subiectivitate reprezintă o modalitate comodă de creare a unor resurse lingvistice necesare detectării şi evaluării automate a opiniilor exprimate textual. Deşi SentiWordNet a fost dezvoltat pentru limba engleză, având în vedere existenţa de dicţionare semantice monolingve[2] cu structură identică sau foarte asemănătoare cu Princeton WordNet (Fellbaum, 1998) pentru mai mult de 40 de limbi şi că majoritatea acestor dicţionare folosesc ca index interlingual chiar Princeton WordNet, marcajul de subiectivitate poate fi transferat (în virtutea echivalenţei interlinguale a înţelesurilor) şi exploatat în oricare dintre limbile respective (inclusiv limba română). Vom denumi în continuare orice wordnet astfel îmbogăţit cu informaţia de subiectivitate, cu termenul de sentiwordnet. Apariţia resurselor de tip sentiwordnet va conduce la amplificarea cercetărilor de subiectivitate pentru tot mai multe limbi, iar în viitorul imediat, la realizarea de noi aplicaţii în acest domeniu.

O astfel de aplicaţie va fi prezentată în cele ce urmează.

2. Detecţia posibilelor conotaţii textuale

Majoritatea reclamelor publicitare pe care le vedem zilnic exploatează în mod inteligent ambiguitatea limbajului utilizând jocuri de cuvinte, asocieri surprinzătoare, imagini care

[1] citat: "mainly explained by the availability of resources for subjectivity analysis, such as lexicons and manually labeled corpora", (Mihalcea et al., 2007).

[2] http://www.globalwordnet.org/

împing către un anumit context de interpretare, pentru a promova diverse produse şi/sau servicii. Multe din scurtele propoziţii folosite în acest sens, atunci când sunt folosite în texte obişnuite, pot avea o parte din posibilele conotaţii mascate de context şi astfel să fie nesesizate de către cititorul obişnuit. Observaţia este validă însă şi în sens invers: anumite propoziţii luate din contextul original şi plasate în contexte noi, eventual cu grijă alese, pot fi purtătoare de mesaje complet noi, adesea nedorite. Este relativ uşoară identificarea, mai ales în interiorul textelor argumentative, a unor astfel de propoziţii ce pot fi folosite într-un mod maliţios în contexte care să poată induce o interpretare complet diferită şi chiar opusă faţă de cea originală.

Metodele de analiză a subiectivităţii decid în general dacă o propoziţie este subiectivă sau nu, iar în caz afirmativ stabilesc polaritatea şi scorul de subiectivitate. Acest lucru se poate face uşor atât timp cât fiecare cuvânt din propoziţie este dezambiguizat la nivel de sens, iar pentru fiecare sens identificat există înregistrat într-un dicţionar sentiwordnet un scor de subiectivitate. Pe lângă acest tip de analiză, aplicaţia descrisă in continuare, numită CONAN (CONnotation ANalyzer), poate fi folosită pentru o prelucrare textuală mai complexă: estimarea *variabilităţii conotative* a unei propoziţii, reprezentând potenţialul unui enunţ textual de a-şi modifica, în funcţie de context, intensitatea sau chiar polaritatea opiniei subiective. Aplicaţia estimează pe o scară de la 0 la 1, măsura în care o propoziţie, independent de contextul său curent de ocurenţă, poate fi interpretată obiectiv (O), subiectiv-pozitiv (P) sau subiectiv-negativ (N), elementele de interes pentru această estimare fiind senti-cuvintele din propoziţia prelucrată. În mai toate cazurile, aceste scoruri sunt diferite. Experimentele noastre arată că propoziţiile pentru care scorurile de subiectivitate (pozitivă şi respectiv negativă) sunt ridicate şi comparabile pot fi uşor folosite în „jocuri conotaţionale" de care autorul poate fi conştient sau nu.

3. CONAN (CONotation ANalyzer)

Sistemul CONAN a fost implementat astfel încât să fie independent de limbă şi ca atare el poate fi folosit pentru diferite limbi, atât timp cât textele de analizat sunt preprocesate corespunzător şi atât timp cât există sentiwordnet-uri pentru aceste limbi.

Preprocesarea textelor, aşa cum este ea necesară sistemului CONAN, include: segmentarea la nivel de unitate lexicală (tokenizare), adnotarea cu descriptori morfo-sintactici (tagging) şi grupuri gramaticale (chunking). Acestea constituie operaţii fundamentale pentru aproape orice aplicaţie NLP, existând implementări pentru majoritatea limbilor. Instrumente ce realizează astfel de prelucrări pentru limba română au fost implementate în mai multe colective din ţară sau din străinătate şi există o bibliografie semnificativă. Pentru o prezentare detaliată a unor astfel de instrumente, implementate la Institutul de Cercetări pentru Inteligenţă Artificială (ICIA), a se vedea (Tufiş, 2008b). Recent, majoritatea programelor de preprocesare realizate la ICIA au fost făcute publice prin intermediul unor servicii web[3] nu numai pentru limba română, dar şi pentru limba engleză (Tufiş et al., 2008).

După preprocesarea textelor, în faza a doua sunt identificate toate senti-cuvintele – cuvinte cu cel puţin o interpretare subiectivă (scorul de obiectivitate este mai mic decât

[3] http://tutankhamon.racai.ro /ttlws.wsdl

1) – folosind sentiwordnet-ul specific limbii prelucrate (în prezent, limba română şi limba engleză). Literatura de specialitate indică faptul că, în problema analizei subiectivităţii, abordările ce ignoră ordinea si relaţiile dintre cuvinte (eng.: *bag-of-words* – BoW) nu sunt potrivite deoarece subiectivitatea atribuită iniţial poate fi schimbată de către contextul local al propoziţiei prin aşa-numiţii modificatori de valenţă (eng.: *valence-shifters*): intensificatori, moderatori şi negaţii. Primii doi modificatori cresc şi respectiv descresc scorurile de subiectivitate în timp ce negaţia complementează valoarea subiectivităţii apriori a senti-cuvântului de sub incidenţa modificatorilor. Cum modificatorii nu acţionează în mod necesar doar asupra senti-cuvântului din imediata vecinătate, adnotarea la grupuri gramaticale este importantă pentru a delimita raza de influenţă a acestora. De exemplu, în propoziţia „NU este FOARTE *simpatic*$^{+}$”, cuvântul *simpatic*$^{+}$ este un senti-cuvânt pozitiv, în timp ce cuvintele scrise cu majuscule (NU, FOARTE) sunt modificatori: negaţie, respectiv intensificator. Intensificatorul acţionează asupra senti-cuvântului, în timp ce negaţia acţionează asupra rezultatului NU(este FOARTE(*simpatic*)). În consecinţă, propoziţia de mai sus are un scor negativ de subiectivitate. În (Tufiş, 2008a) am arătat că majoritatea scorurilor de subiectivitate care sunt greşit atribuite în SentiWordNet, se datorează abordării de tip BoW în cazul analizei definiţiilor sensurilor. Cele mai multe mulţimi de sinonime (eng. *synset*) cu astfel de scoruri greşit calculate au în interiorul definiţiilor modificatori care aparent au fost ignoraţi. Acest lucru ar putea explica de ce cuvintele *honest* (sensul 1) şi sinonimul său *honorable* (primul sens) sunt considerate ca având o conotaţie mult mai negativă (0.5) decât pozitivă (0,25). Glosa ataşată acestei serii sinonimice este: NOT DISPOSED to *cheat*$^{-}$ or *defraud*$^{-}$; NOT *deceptive*$^{-}$ or *fraudulent*$^{-}$.

CONAN acceptă texte de prelucrat atât de la tastatură cât şi din fişiere. În cazul în care la intrare avem un fişier, aplicaţia presupune că fişierul este deja preprocesat şi codificat în acelaşi mod în care platforma de servicii web a ICIA codifică documentele prelucrate – formatul XCES. În figura 1 se poate observa codificarea unei propoziţii aparţinând corpusului SEMCOR[4], preprocesat de platforma TTL (Ion, 2007).

Figura 2 prezintă capturi de ecran ale aplicaţiei având la intrare un fişier (ce conţine şi propoziţia din figura 1). Utilizatorul specifică un mod de interpretare a propoziţiilor din textul de intrare (obiectivă/pozitivă/negativă).

Fereastra din stânga-jos afişează analiza propoziţiilor din fişierul de intrare. Propoziţiile sunt ordonate în funcţie de interpretarea obiectivităţii sau a polarităţii selectate (detaliate în continuare). Fereastra din mijloc-jos afişează scorurile de interpretare ale propoziţiilor din fereastra din partea stângă-jos. Fereastra din dreapta-jos afişează informaţii precum indecşii interlinguali şi definiţiile seriilor de sinonime din care face parte cuvântul selectat de către utilizator în fereastra de analiză (fereastra stânga-jos).

[4] http://www.cs.unt.edu/~rada/downloads.html

```
<s id="br-a01.4.4.ro"><c>"</c>
<w lemma="doar" ana="14+,Rgp" chunk="Ap#1" wns="ili:ENG20-
00004331-b">Doar</w>
<w lemma="un" ana="21+,Timsr" chunk="Np#1">un</w><w
lemma="num&abreve;r" ana="1+,Ncms-n"
chunk="Np#1">num&abreve;r</w>
<w lemma="relativ" ana="14+,Rp" chunk="Np#1,Ap#2">relativ</w>
<w lemma="mic" ana="1+,Afpms-n" chunk="Np#1,Ap#2">mic</w>
<w lemma="de" ana="5+,Spsa" chunk="Pp#1,Ap#3">de</w>
<w lemma="asemenea" ana="1+,Afp"
chunk="Pp#1,Ap#3,Np#2">asemenea</w>
<w lemma="raport" ana="1+,Ncfp-n"
chunk="Pp#1,Np#2">rapoarte</w>
<w lemma="avea" ana="3+,Va--3s" chunk="Vp#1">a</w>
<w lemma="fi" ana="3+,Vap--sm" chunk="Vp#1">fost</w>
<w lemma="primi" ana="1+,Vmp--sm" chunk="Vp#1,Ap#4"
wns="ili:ENG20-00508949-v">primit</w><c>"</c><c>,</c>
<w lemma="avea" ana="3+,Va--3s" chunk="Vp#2">a</w>
<w lemma="spune" ana="1+,Vmp--sm" chunk="Vp#2,Np#3,Ap#5"
wns="ili:ENG20-00976600-v">spus</w>
<w lemma="juriu" ana="1+,Ncmsry" chunk="Np#3" wns="ili:ENG20-
07903245-n">juriul</w><c>,</c><c>"</c>
<w lemma="considera" ana="1+,Vmg"
chunk="Vp#3">consider&acirc;nd</w>
<w lemma="interes" ana="1+,Ncmsry" chunk="Np#4"
wns="ili:ENG20-05354775-n">interesul</w>
<w lemma="r&abreve;sp&acirc;ndi" ana="1+,Vmp--sm"
chunk="Np#4,Ap#6,Vp#4">r&abreve;sp&acirc;ndit</w>
<w lemma="&icirc;n" ana="5+,Spsa" chunk="Pp#2">&icirc;n</w>
<w lemma="alegere" ana="1+,Ncfp-n" chunk="Pp#2,Np#5"
wns="ili:ENG20-00171672-n">alegeri</w><c>,</c>
<w lemma="num&abreve;r" ana="1+,Ncmsry" chunk="Np#6"
wns="ili:ENG20-12816962-n">num&abreve;rul</w>
<w lemma="aleg&abreve;tor" ana="1+,Ncmpoy" chunk="Np#6"
wns="ili:ENG20-10058086-n">aleg&abreve;torilor</w>
<w lemma="&scedil;i" ana="31+,Crssp">&scedil;i</w>
<w lemma="m&abreve;rime" ana="1+,Ncfsry" chunk="Np#7"
wns="ili:ENG20-04819645-n">m&abreve;rimea</w>
<w lemma="acest" ana="2+,Dd3mso---e" chunk="Np#8">acestui</w>
<w lemma="ora&scedil;" ana="1+,Ncms-n" chunk="Np#8"
wns="ili:ENG20-08005407-n">ora&scedil;</w><c>.</c><c>"</c></s>
```

Figura 3: Propoziţie codificată conform standardului XCES

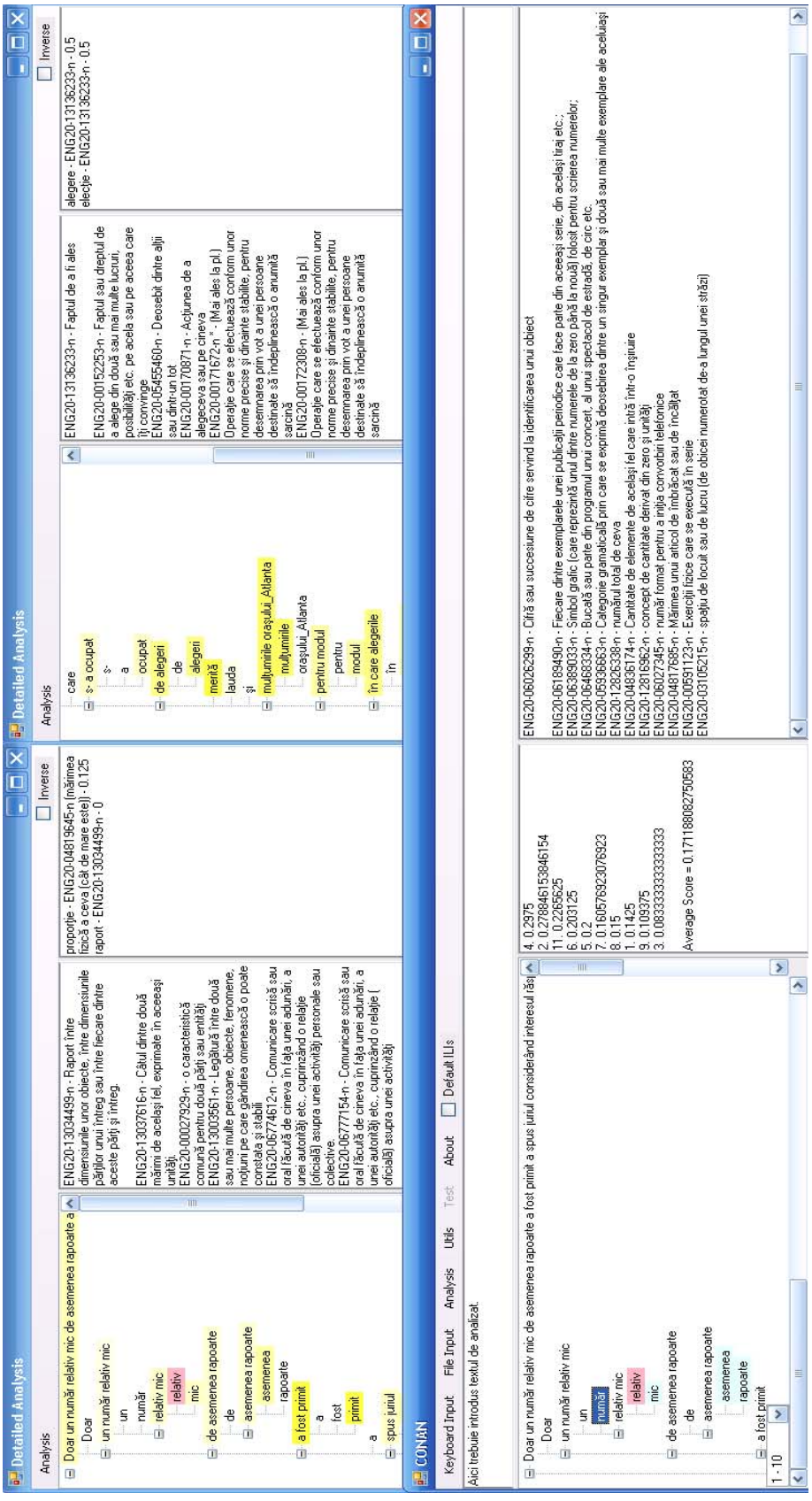

Figura 2: CONAN

CONAN exploatează informaţia furnizată de TTL despre grupurile gramaticale (eventual imbricate) identificate în propoziţia curentă şi construieşte structuri arborescente similare structurilor recursive de constituenţi. Reprezentările astfel obţinute sunt folosite pentru a calcula scoruri pentru interpretări de subiectivitate şi obiectivitate. Primul pas presupune selecţionarea înţelesurilor senti-cuvintelor cu scorurile cele mai ridicate relativ la interpretarea selectată de către utilizator. Algoritmul calculează apoi recursiv scorurile de interpretare pentru fiecare nod al arborelui făcând media aritmetică a scorurilor nodurilor sale copil. Pornind de la frunze (ce conţin senti-cuvintele), scorurile se propagă până când scorul întregii propoziţii este calculat. Prin selecţia din meniu a opţiunii *Analysis*, utilizatorul are posibilitatea de a-şi concentra analiza pe o propoziţie aleasă (vezi figura 2, partea superioară). Aplicaţia permite multiple alegeri de acest fel cu posibilitatea deschiderii mai multor asemenea ferestre concomitent. Astfel, utilizatorul poate compara diversele grade de subiectivitate ale diferitelor propoziţii. După cum afirmam, programul oferă opţiuni multiple de interpretare în direcţii de polaritate sau obiectivitate diverse: interpretare pozitivă, negativă sau obiectivă (pe acestea le vom numi interpretări principale), iar pe lângă acestea se oferă posibilitatea forţării interpretărilor în direcţii dorite: forţează cea mai pozitivă interpretare, forţează cea mai negativă interpretare, forţează cea mai obiectivă interpretare, forţează cea mai non-negativă interpretare, forţează cea mai non-negativă pozitivă, forţează cea mai non-subiectivă interpretare. Cea mai simplă operaţiune constă în afişarea polarităţii tuturor propoziţiilor textului, în eventualitatea în care cuvintele au fost în prealabil dezambiguizate în ceea ce priveşte sensul.

În cazul în care textul este introdus de la tastatură, aplicaţia detectează dacă acesta (format din una sau mai multe propoziţii) este sau nu preprocesat. În cazul în care nu este, textul brut este trimis serviciilor lingvistice web ala ICIA, servicii de care aminteam mai sus. Restul operaţiilor se petrec ca în descrierea de la paragraful anterior.

Aşadar, utilizatorul poate cere o analiză a tipurilor de interpretare principale. În acest caz sensurile considerate pentru senti-cuvinte sunt cele cu scorurile de polaritate/obiectivitate cele mai ridicate. pentru interpretarea dorită. Cuvintele care nu sunt senti-cuvinte sunt considerate de obiectivitate maximă:1.

În timp ce scorurile de mai sus pot fi calculate doar pentru întreg textul de la intrare, opţiunea *Analysis* oferă posibilitatea schimbării rapide între diferite interpretări ale unei propoziţii. Mai mult, utilizatorul poate cere două tipuri de interpretări forţate: în direcţia unei polarităţi, sau opusă direcţiei unei polarităţi. Aceste două tipuri de interpretare sunt mai elaborate şi le vom numi *tipuri de analiză complexă*, deoarece, spre deosebire de cazul interpretărilor principale, aplicaţia nu numai că face aceeaşi analiză a propoziţiilor, dar şi sugerează înlocuirea anumitor cuvinte în funcţie de direcţia interpretării cerute. Aşadar, pentru ambele tipuri de analiză complexă, pasul iniţial este acelaşi cu cel al tipurilor principale: cuvintelor le sunt atribuite înţelesurile cu scorurile cele mai mari pentru interpretarea dorită şi apoi scorurile pentru toate propoziţiile sunt calculate. Pentru partea a doua însă, în cazul primului tip de analiză complexă, cuvintele sunt înlocuite de sinonime selectate din seriile sinonimice corespunzătoare înţelesurilor deja atribuite, sinonime reprezentate de literali ce pot avea sensuri cu scoruri mai mari (evident în alte serii sinonimice) pentru interpretarea selectată. Pentru a formaliza, să presupunem că avem cuvântul w cu sensurile m_1, m_2, …, ,m_n (corespunzând evident la n

serii sinonimice). Pentru o anumită interpretare *I*, să presupunem că sensul cu scorul cel mai ridicat este m_l. Pentru acest sens, *w* are următoarele sinonime s_1, s_2, …, s_k. În mod evident, sinonimele corespund unor sensuri ale altor literali care mai pot avea şi alte sensuri. Cu alte cuvinte, fiecare s_i este un sens pentru un literal L_i care poate avea multiple alte sensuri: m_{i1}, m_{i2}, …, m_{it}. Algoritmul selectează acel literal care are un sens având cel mai mare scor pentru interpretarea curentă (dintre toate celelalte sensuri ale tuturor literalilor consideraţi). Literalul câştigător este acela care corespunde expresiei:

$$\max_{i \text{ pentru } I}(score(m_{i1}), score(m_{i2}),..., score(m_{i|L_i|}))$$

Este clar că, în anumite cazuri, selecţia literalilor pe acest criteriu poate conduce schimbarea sensului propoziţiilor. Este tocmai ce ar dori să obţină un utilizator ce foloseşte această opţiune. În acest caz, utilizatorul vrea să-l forţeze pe eventualul cititor să interpreteze textul într-o anumită direcţie. Înlocuirea cuvintelor se face la nivelul formei ocurenţă prin utilizarea lemei literalului (furnizată de WordNet), descriptorul morfo-sintactic (msd) asociat cuvântului original în faza de preprocesare şi un tabel conţinând forme de ocurenţă a cuvintelor limbii, împreună cu lemele şi msd-urile corespunzătoare (tblwordform).

Aşa cum am mai menţionat, am numit al doilea tip de analiză complexă *interpretare care forţează în direcţia opusă direcţiei unei polarităţi*. Diferenţa dintre această interpretare şi cea anterioară constă în modul în care se face selecţia literalilor care să înlocuiască cuvintele originale. În acest caz, pentru a înlocui un cuvânt, algoritmul selectează literalul având sensul cu cel mai mic (şi nu mai mare) scor, în interpretarea inversă (şi nu curentă). Literalul câştigător corespunde expresiei:

$$\min_{i \text{ pentru } I}(score(m_{i1}), score(m_{i2}),..., score(m_{i|L_i|}))$$

Motivaţia acestei opţiuni constă în dorinţa de a selecta sinonime pentru cuvintele originale astfel încât să se poată evita interpretările cu anumite polarităţi.

În cadrul aplicaţiei se folosesc nuanţe de culoare cu ajutorul cărora utilizatorul are posibilitatea de a identifica imediat gradul în care diferite noduri din structura arborescentă corespunzătoare unei propoziţii contribuie la scorul final al acesteia. Un simplu clic pe un nod afişează scorul corespunzător nodului.

O altă opţiune pe care CONAN o oferă utilizatorilor săi este aceea de a calcula scorul de *interpretativitate* a propoziţiilor unui text. Definim scorul de interpretativitate al unei propoziţii (ISP) ca fiind o mărime cantitativă a potenţialului unei propoziţii de a-şi schimba conotaţia. Interpretativitatea unei propoziţii se calculează ca sumă normalizată a scorurilor de interpretabilitate subiectivă ale propoziţiei (Eq2). Scorul de interpretativitate al unei propoziţii este strâns legat de scorurile de interpretativitate ale cuvintelor ce o compun. Cu cât scorurile pentru cuvinte sunt mai mari, cu atât cel calculat pentru propoziţie va fi mai mare.

Interpretativitatea unui cuvânt (ISW) se defineşte ca în ecuaţia (Eq1) pe baza scorurilor de maximă pozitivitate şi respectiv maximă negativitate asociate sensurilor sale.

$$ISW(sw_k) = \frac{0.5 * (\max P(sw_k) + \max N(sw_k))}{1 + |\max P(sw_k) - \max N(sw_k)|} \quad \text{(Eq1)}$$

$$ISP\ (\text{propoziţie}_k) = \frac{0.5 * (\max P(\text{propozitie}_k) + \max N(\text{propozitie}_k))}{1 + |\max P(\text{propozitie}_k) - \max N(\text{propozitie}_k)|} \quad \text{(Eq2)}$$

Justificarea intuitivă a acestor formule empirice constă în faptul că, în condiţiile în care un senti-cuvânt are cel puţin două sensuri, unul marcat cu un scor mare de interpretare pozitivă iar celălalt cu un scor mare de interpretare negativă, acel cuvânt are un impact major în schimbarea conotaţiei unei propoziţii. O propoziţie are o variabilitate conotativă cu atât mai mare cu cât scorul ei de interpretativitate este mai mare.

Scorul de interpretativitate al unui senti-cuvânt poate fi maxim 1 când are cel puţin două sensuri, unul de pozitivitate maximă (1), iar altul de negativitate maximă (1). În SentiWordNet-ul românesc, cuvintele cu interpretativitatea cea mai ridicată sunt adjectivele *prost* şi *imoral* şi substantivul *generozitate* (toate având 0.875).

Trebuie să menţionăm că modificatorii de valenţă se găsesc în trei fişiere externe, uşor editabile, care sunt citite la fiecare lansare a aplicaţiei. În momentul de faţă toţi modificatorii au o influenţă uniformă asupra senti-cuvintelor: intensificatorii şi modificatorii măresc sau scad scorurile argumentelor lor cu 20%, în timp ce negaţia comută între scorurile P şi N. O abordare mai elaborată, aflată în construcţie, va specifica mai multe trăsături pe care le poate avea un modificator: categoria gramaticală preferată, argumentul preferat şi chiar numărul de sens preferat al argumentului, dacă este cazul. În plus se vor putea defini diferite grade de influenţă a modificatorului în funcţie de argument.

4. Rezultate şi Concluzii

Majoritatea experimentelor au fost realizate pe corpusul SEMCOR, un corpus paralel Română-Engleză preprocesat, cu conţinut divers. Versiunea de lucru utilizată conţine 8146 de propoziţii a căror analiză, indiferent de interpretarea dorită, durează doar câteva minute. Au fost efectuate analize de variabilitate a conotaţiei atât pentru limba engleză cât şi pentru limba română. Valorile obţinute pentru limba română sunt uşor diferite de cele obţinute pentru limba engleză, deşi ordinea propoziţiilor sortate după scorul de interpretabilitate se păstrează (lucru uşor previzibil datorită ipotezei de lucru enunţate în secţiunea de introducere). Diferenţa între scorurile de interpretativitate se datorează în primul rând numărului mult mai mic de adjective (principalele senti-cuvinte) existente în wordnetul pentru limba română faţă de wordnetul pentru limba engleză. De asemenea, cuvintele din wordnetul pentru limba română au, în general, mai puţine sensuri implementate decât cuvintele din Princeton WordNet.

Deşi analiza rezultatelor acestui experiment este abia la început, au fost identificate o serie de propoziţii (în ambele limbi) cu scoruri de interpretabilitate contrazicând intuiţia comună. Prin inspectarea mai amănunţită a acestor propoziţii şi a senti-cuvintelor componente au fost detectate multe marcaje <O,P,N> din SentiWordNet cu valori cel puţin discutabile (a se vedea discuţia din secţiunea 3).

Cercetările noastre viitoare vor avea în vedere, pe lângă o evaluare cros-lingvală a ipotezei de validitate a importului cros-lingual de marcaje de subiectivitate pe baza echivalenţei de traducere, o mărire substanţială a inventarului de adjective în wordnetul

românesc precum şi o îmbunătăţire a metodei de calcul al scorurilor de subiectivitate lexicală (apriori) pentru sensurile adjectivelor.

Referinţe bibliografice

Esuli A.& F. Sebastiani. (2006). SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. *Proceedings of the 5th Conference on Language Resources and Evaluation LREC-06*, Genoa, Italy, pp. 417-422.

Fellbaum C., editor. (1998). *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Ion, Radu (2007). *Word Sense Disambiguation Methods Applied to English and Romanian*. PhD thesis (in Romanian). Romanian Academy, Bucharest.

Mihalcea R.; Banea C.; Wiebe J. (2007). Learning Multilingual Subjective Language via Cross-Lingual Projections. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*,.Prague, Czech Republic, June, pp. 976-983.

Polanyi L. & Zaenen A. (2006). Contextual Valence Shifters. In James G. Shanahan, Yan Qu, and Janyce Wiebe, editors, *Computing Attitude and Affect in Text: Theory and Application*. Springer Verlag.

Tufiş D.; Ion R.; Ceauşu A.; Ştefănescu D. (2008). RACAI's Linguistic Web Services. In *Proceedings of 6th Conference on Language Resources and Evaluation LREC-08*, Marrakech, Marocco.

Tufiş D. (2008a). *Subjectivity mark-up in WordNet: does it work cross-lingually? A case study on Romanian Wordnet*. Invited talk on the Panel "Wordnet Relations" at the Global WordNet Conference, January 22-25, 2008.

Tufiş D. (2008b). Algorithms and Data Design Issues for Basic NLP Tools. In Sergei Nirenburg and Oleg Kapanatze (eds). *Advances in Language Engineering for Low- and Middle-Density Languages*, NATO-ASI, 48 pages, IOS Press

COMPLETAREA AUTOMATĂ A RESURSELOR LINGVISTICE ROMÂNEŞTI

PETIC MIRCEA

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

mirsha@math.md

Rezumat

În articol sunt examinate unele aspecte de completare a resurselor lingvistice utilizând proceduri de derivare automată. De asemenea este prezentată o descriere a particularităţilor afixelor româneşti, precum şi a metodelor de generare automată a derivatelor cu prefixe şi sufixe pentru a completa ulterior Resursele Reutilizabile pentru Tehnologia Limbajului Natural.

Cuvinte cheie: generarea automată a derivatelor, flexionare automată, validarea cuvintelor

1. Introducere

Aplicaţiile ce ţin de prelucrarea limbajului natural presupun crearea, completarea şi folosirea resurselor lingvistice electronice. Completarea automată sau/şi semiautomată a acestor resurse cu cuvinte generate în baza celor deja existente reprezintă o sursă importantă de îmbogăţire a vocabularului prin mijloace exclusiv interne.

Scopul acestui articol va fi studierea particularităţilor afixelor româneşti şi a metodelor de generare automată a derivatelor cu sufixe şi prefixe pentru a completa ulterior Resursele Reutilizabile pentru Tehnologia Limbajului Natural[1] (RRTLN).

Iniţial vom efectua o trecere în revistă a definiţiilor şi noţiunilor ce ţin de derivare împreună cu anumite clasificări, precum şi a metodelor existente de derivare automată. Un compartiment aparte este destinat resurselor lingvistice electronice folosite în studiul mecanismelor de recunoaştere şi generare a derivatelor noi cu prefixe şi sufixe. Ne vom opri asupra trei momente:

a) Identificarea derivatelor,

b) Derivarea cu prefixe şi sufixe, urmată de flexionarea derivatelor obţinute,

c) Validarea derivatelor generate automat.

În această ordine de idei procesul de identificare a derivatelor este caracterizat de unele momente relevante în proiectarea ulterioară a algoritmilor de generare automată a derivatelor cu prefixe şi sufixe.

În completarea resurselor lingvistice prin derivare automată apare tendinţa firească de a folosi cele mai frecvente afixe. Însă practic cele mai productive afixe se dovedesc a fi problematice datorită comportamentului neregulat. De aceea pentru cercetările noastre am ales acele afixe, care ne-au permis să stabilim nişte legităţi de comportament mai simple, fără a invoca prea multe excepţii.

[1] Lexiconul se conţine pe site-ul http://imi201.math.md/elrr/

Din aceste considerente am operat cu prefixele *ne-* şi *re-*, precum şi cu sufixele *–tor* şi *–bil*, ultimele fiind, la rândul său, frecvente în procesul de derivare cu prefixul *ne-*. Am inclus în examinarea noastră sufixul lexical verbal *–iza*, care este de origine neologică şi foarte productiv la momentul de faţă cu o relaţie foarte strânsă cu sufixele lexicale *–ism* şi *-ist*.

2. Particularităţile procesului de derivare

Derivarea reprezintă unul din mijloacele de îmbogăţire a vocabularului care foloseşte resursele proprii ale limbii, pornind de la cuvinte existente în limbă, în particular, formarea de cuvinte noi ori cu sens nou prin adăugarea unor afixe la bazele lexicale existente. Prin *afix* se înţelege orice morfem care rămâne în afara rădăcinii, atunci când segmentăm un cuvânt. În denumirea globală de afixe se includ prefixele şi sufixele. Cuvântul care este format prin adăugarea unui prefix sau sufix se numeşte *cuvânt derivat*. Unităţile de la care se formează cuvintele derivate se numesc *baze* sau *primitive*. De multe ori, prefixele şi sufixele nu se adaugă direct la rădăcină, ci la aşa-numita temă lexicală. Aceasta este comună tuturor formelor flexionare sau gramaticale ale unui cuvânt şi este formată, în mod obligatoriu, dintr-o rădăcină şi, cel puţin, un sufix sau prefix (Hristea, 1984).

În conformitate cu structura lor morfologică afixele sunt divizate în: *simple* (când nu pot fi divizate în unităţi mai mici) şi *complexe* (când structura lor permite identificarea unor unităţi mai mici, dar întregul complex funcţionează ca un element unic de derivare).

După poziţia pe care o ocupă faţă de rădăcină, elementele adăugate la bază se împart în două mari categorii, şi anume: unele care sunt plasate înaintea rădăcinii şi se numesc *prefixe*, iar altele care sunt ataşate la sfârşitul ei poartă denumirea de *sufixe*.

Sufixele sunt *lexicale* sau *derivative* în cazul în care ele servesc la formarea de noi cuvinte şi *flexionare*, *morfologice* sau *gramaticale* când servesc la realizarea unor forme ale aceluiaşi cuvânt; sufixele lexicale se menţin în toate formele flexionare ale derivatului respectiv, pe când cele flexionare caracterizează anumite forme. În structura cuvintelor derivate sufixele lexicale sunt urmate de cele flexionare, iar acestea de desinenţe. De cele mai multe ori, sufixele conferă cuvintelor noi create o anumită valoare semantică şi morfologică.

Numărul prefixelor simple, incluse în (Graur&Avram, 1978) în urma identificării în cel puţin un derivat este de 86, iar a sufixelor lexicale de peste 600 (Hristea, 1984).

3. Metode existente de derivare automată

Dicţionarele moderne se confruntă cu anumite deficienţe, care sunt obiecte de cercetare pentru lexicografi. Cu toate că dicţionarele sunt permanent completate cu intrări noi, graţie dezvoltării continue a limbii, sarcina elaborării unui vocabular complet rămâne una practic imposibilă. Mai mult ca atât, cunoaştem că în fiecare zi sunt create ad-hoc o mulţime de cuvinte noi, totuşi în majoritatea cazurilor, dar nu în totalitate, ele rămân pentru o perioadă îndelungată nevalidate, cei care creează neologismele fiind mai puţin preocupaţi de „legalitatea” prezenţei acestor cuvinte într-o limbă sau alta.

În cadrul modelului de derivare cu sufixe pentru limba italiană descris în (Carota, 2006) sunt investigate principalele modalităţi de formare a cuvintelor complexe prin sufixare în baza cuvintelor italiene morfologic simple. Pentru principalele tipuri de derivate italiene cu sufixe au fost identificate unele nuclee semantice. Astfel, pe de o parte se tratează interfaţa dintre morfo-sintaxă şi semantică, iar pe de altă parte o interfaţă între sufixare şi flexionare (Carota, 2006).

Caracteristicile relaţiei dintre morfologia derivaţională şi sinonimie raportată la un dicţionar electronic sunt studiate în lucrarea (Duško&Krstev, 2005) fiind ilustrate în baza derivării cuvintelor în limba sârbă. În acest context au fost generate noi leme cu sensuri previzibile. Acest procedeu a fost numit derivare regulară. Acest tip de derivare este utilizat în prelucrarea textului folosind dicţionarul electronic morfologic al limbii sârbe şi o colecţie de traducere cu constrângeri lexicale.

În (Vilares et al., 2001) se descrie modul de lucru al unei aplicaţii de prelucrare a limbajului natural pentru extragerea informaţiei. Autorii aplicaţiei propun generarea familiilor morfologice ale unui cuvânt, fapt care va reduce varietatea lingvistică de documente indexate în limba spaniolă. Principalele caracteristici ale acestui sistem sunt: utilizarea minimă a resurselor lingvistice, costul mic „computaţional" şi independenţă faţă de motorul de indexare.

Pentru limba arabă a fost proiectat un sistem MORPHE (Leavitt, 1994) care reprezintă un analizor/generator elaborat ca o componentă a tehnologiei de traducere automată KANT (Nyberg&Mitamura, 1992). Deşi MORPHE a fost proiectat ca să fie folosit atât în analiză, cât şi în generare, în practică el este folosit doar pentru generare şi doar în limbile care conţin prefixe şi sufixe.

În (Santana et al., 2004) este descris un instrument capabil să recunoască, să genereze şi să manipuleze relaţiile morfo-lexicale ale cuvintelor cât şi stabilirea cuvintelor primitive de la care au fost formate derivatele. Totodată permite lucrul cu prefixele şi sufixele în parte cât şi stabilirea relaţiilor între ele şi cuvinte derivate cu afixele respective.

Una din primele aplicaţii de derivare automată pentru limba română a fost sistemul FAVR, realizat în mediul Mac-ELU (Tufiş et al., 1996) care a avut drept scop acoperirea completă a morfologiei flexionare. Odată însă cu migrarea descrierii FAVR în acest mediu s-a abordat şi descrierea proceselor lexicale. Sub aspect derivativ, sufixele şi prefixele lexicale au un potenţial productiv mare. În urma analizei atributelor specifice fiecărei părţi de vorbire în parte, în descrierea morfologică implementată în Mac-ELU s-au utilizat 20 de categorii gramaticale. Clasificarea s-a efectuat nu numai în baza cerinţelor prelucrărilor morfo-lexicale, ci şi a granularităţii necesare analizei şi, respectiv, a generării sintactice.

O altă aplicaţie ce merită atenţie este AnMorph (Cristea&Forăscu, 2006). Ea reprezintă un mediu de dezvoltare şi actualizare a modelului morfologic paradigmatic al unei limbi neaglutinative (modelul de bază al cuvântului flexionat consideră cuvântul ca fiind compus din rădăcină şi o terminaţie). Programul compară formele introduse de utilizator cu acele care pot fi generate pornind de la o paradigmă deja existentă în baza de date a programului, şi dacă asemănarea este confirmată, se generează restul formelor. Când se întâmplă acest lucru, utilizatorul doar verifică şi validează partea tabelului generată automat. În afară de interfaţa pentru dezvoltare-actualizare, mediul oferă un editor

pentru dicţionar şi o colecţie de paradigme, o componentă care permite verificări de consistenţă a datelor şi a lematizorului.

4. Resursele lingvistice electronice ale limbii române

Aplicaţiile ce ţin de procesarea unui limbaj natural necesită în mare parte resurse lingvistice, care reprezintă cunoştinţele lingvistice împreună cu datele suport (structurate într-o formă prestabilită) şi programe asociate (Tufiş&Barbu, 2002).

O resursă importantă este dicţionarul morfologic de limba română (DMLR) (Lombard& Gâdei, 1981). Acest dicţionar conţine 28932 de cuvinte care sunt împărţite în clase de flexionare în dependenţă de modul de formare a acestora. Pornind de la DMLR, au fost elaborate programe de flexionare pentru limba română (Cojocaru, 1997). Ele au contribuit substanţial la acumularea resurselor lingvistice.

Pachetul de programe "Produse program pentru aplicaţii lingvistice" a fost utilizat cu succes la implementarea corectorului de texte pentru limba română RomSP (Boian et al., 2000). Dezvoltarea celui din urmă a condus la implementarea RRTLN, care conţin o bază de date cu informaţie lingvistică la nivel de cuvânt şi un set de programe de gestionare (Boian et al., 2005a, 2005b). Astfel, lexiconul conţine nu doar reprezentarea grafică a cuvântului, dar şi informaţia despre partea de vorbire al lui. RRTLN are aproximativ 100000 de cuvinte de bază şi circa un milion de flexiuni. De menţionat, că un cuvânt poate avea mai multe intrări pentru diferite părţi de vorbire.

5. Identificarea derivatelor

Drept sursă pentru recunoaşterea derivatelor cu prefixe a servit lexiconul RRTLN şi o listă de prefixe simple cu formele lor fonologice care sunt înregistrate în (Graur&Avram, 1978). Ţinând cont de particularităţile prefixelor precum şi a derivatelor lor, a fost elaborat un algoritm de extragere automată a cuvintelor derivate cu prefixe simple.

Cu mici schimbări algoritmul menţionat mai sus s-a folosit pentru studierea problemelor de extragere a derivatelor cu sufixe din lexiconul RRTLN, în baza derivatelor cu sufixele *–tor* şi *–bil*. Selectarea acestor sufixe e motivată prin existenţa unui număr mare de cuvinte cu aceste particule în lexicon. În urma verificării s-a stabilit că nu au fost găsite toate cuvintele derivate. Motivul este prezenţa alternanţelor vocalice în desinenţele verbelor, fapt ce nu este luat în calcul în algoritmul (Petic, 2007a). În afară de aceasta lexiconul nu conţine toate verbele de la care au fost formate derivatele cu sufixele *–tor* şi *–bil*. În plus, s-a constatat că verbele de la care s-au format derivate cu sufixul *-tor* fără alternanţe vocalice/consonantice se termină în *a*, *i, ă* şi *î*. Acest algoritm a fost implementat într-un program în limbajul de programare C++, în mediul Windows (Petic, 2007a). Totodată, programul este util în extragerea atât a prefixelor compuse cât şi a sufixelor lexicale compuse în cuvintele derivate din acelaşi lexicon.

Verbele care se termină în *e* formează substantive în *-tor* doar cu ajutorul alternanţelor vocalice/consonantice. Numărul verbelor care se termină în *î* este foarte mic, în primul caz 7, în cel de-al doilea doar 1 – *dogorî*. În ceea ce priveşte litera *a* se observă că dacă verbul se termină în *ja, ua, va, xa*, atunci există o probabilitate destul de mare că se vor

forma derivate cu *-tor* fără alternanţe vocalice şi consonantice. Numărul unor astfel de terminaţii pentru litera *i* este mai mare: *îi, şi, ţi, ai, bi, di, ei, fi, hi, ii, ji, li, mi, ni, si, vi* şi *zi*. Totodată este ambiguă situaţia cu verbele în alte terminaţii, precum *a* şi *i*, din cauza că pot fi atestate cuvinte derivate atât cu alternanţă cât şi fără alternanţe. În plus, numărul unor astfel de derivate este destul de mic.

S-a constatat că verbele de la care s-au format derivate cu sufixul *-bil* fără alternanţe vocalice/consonantice se termină *a* şi *i*. Verbele care formează derivate cu sufixul *-bil* cu alternanţe vocalice/consonantice se termină în *a, e* şi *i*. Este cert că verbele care se termină în *e* formează substantive în *-bil* doar cu ajutorul alternanţelor vocalice/consonantice. Se pune în evidenţă repetarea literelor *a* şi *i*. În ceea ce ţine de litera *a* se poate spune că dacă verbul se termină în *şa, ţa, ba, ga, ja, la, ma, ua, va* şi *xa* atunci există o probabilitate destul de mare că se vor forma derivate cu *-bil* fără alternanţe vocalice şi consonantice. Numărul unor astfel de terminaţii pentru litera *i* este mai mic: *ăi, li, ri, ni, si* şi *ti*. Ca şi în cazul sufixului *-tor* rămâne ambiguă situaţia cu verbele în alte terminaţii, precum *a* şi *i,* din cauza că pot fi atestate cuvinte derivate atât cu alternanţă cât şi fără alternanţe. În plus, numărul unor astfel de derivate este destul de mic.

Este interesantă situaţia cu alternanţa literei *e* la sfârşitul verbului pentru formarea derivatelor în sufixele *–tor* şi *–bil*. În cazul derivării sufixale de la verbe cu afixele *-tor* şi *-bil* se observă unele alternanţe vocalice şi consonantice atât la desinenţe cât şi în rădăcină, unele nu se atestă la flexionare.

6. Derivarea automată

6.1. Derivarea automată cu prefixele ne- şi re-

Studiind particularităţile prefixelor *ne-* şi *re-* (Iordan, 1970), s-au obţinut următoarele legităţi pentru prefixul *ne-* care permit îmbogăţirea resurselor lingvistice:

- de la adjectivele derivate cu sufixele *–tor, -bil, -os* se formează adjectivele derivate cu prefixul *ne-* (de exemplu: *neconductor*, *nenobil*, *neinvidios*)
- de la participiile terminate în alomorfele *–at, -it, -ut* se formează adjectivele derivate cu prefixul *ne-* (de exemplu: *nelaureat*, *neiubit*, *nenăscut*);
- de la gerunzii se formează adjectivele derivate cu prefixul *ne-* (de exemplu: *nesuferind*).

Respectiv, legităţile pentru prefixul *re-* sunt următoarele:

- de la infinitivul verbelor se formează verbe derivate cu prefixul *re-* (de exemplu, a *regenera*);
- de la infinitivul verbelor se formează substantive derivate atât în sufixul *–re*, cât şi în prefixul *re-* (de exemplu, *recitire*).

Legităţile formulate mai sus necesită cunoaşterea doar a reprezentării grafice a cuvântului şi a părţii de vorbire a lui. La stabilirea lor s-a operat cu DMLR. Algoritmul de derivare analizabilă cu prefixele *ne-* şi *re-* constă în examinarea cuvintelor din

lexicon şi concatenarea cu prefixe din clasa celor, care se încadrează în categoriile stabilite de legităţile de mai sus (Petic, 2007b).

Astfel, programul elaborat în baza algoritmului menţionat a îmbogăţit lexiconul cu 397 cuvinte derivate cu prefixul *ne-* şi 8556 cuvinte derivate cu prefixul *re-* (Petic, 2008a).

6.2. Derivarea automată cu sufixe lexicale

6.2.1. Cazul sufixului lexical verbal –iza

Analizând particularităţile de derivare a sufixului lexical verbal *–iza* s-a constat:

- cuvintelor care se termină în *–an* sau *–ian* le poate fi ataşat sufixul *–iza*, fără alternanţe vocalice sau consonantice, aşa cum sufixul *–an* este mai scurt, deci afixul *–ian* poate fi inclus în *–an* (de exemplu: *alcaniza*);
- nu este similar cazul cu *–ean*, deoarece aici apare alternanţa vocalică *ea->e* (de exemplu: *europeniza*);
- în cazul terminaţiilor *–atic* (de exemplu: *dramatiza*), *-etic* (de exemplu: *cosmetiza*), *-otic* (de exemplu: *patriotiza*) şi *–ific* (de exemplu: *ştiinţifiza*) se înlătură ultimele două litere şi se adaugă sufixul *–iza*, iar în alte cazuri la *–ic* se alipeşte, pur şi simplu, *-iza*;
- în cazul terminaţiei *–ură* ultima vocală este înlăturată şi se alipeşte *–iza* (de exemplu: *caricaturiza*);
- există o relaţie strânsă între verbele în *-iza* şi substantivele şi adjectivele în *–ism* şi *–ist*, care se manifestă prin apariţia a numeroase serii de derivate de la aceleaşi teme (Petic, 2008b).

Ţinând cont de cele expuse mai sus pentru generarea noilor cuvinte se vor verifica nu doar terminaţiile, dar şi existenţa substantivelor şi adjectivelor respective cu sufixele lexicale neologice *-ist* şi *-ism*. În plus, se va verifica existenţa cuvintelor obţinute în lexicon. Drept sursă pentru generarea noilor verbe în *-iza* a servit lexiconul RRTLN. Deoarece în lexicon sunt 1178 de cuvinte în *-ism* şi 1285 în *-ist* este posibilă verificarea multiplă a terminaţiilor cuvintelor înainte de a fi formate cuvinte noi.

Examinând cuvintele din lexicon şi concatenându-le sufixul lexical *-iza* respectiv celor, care se încadrează nu doar în categoriile stabilite terminaţiilor în (Petic, 2008b), dar care lipsesc în lexicon şi permit formarea de serii de la aceleaşi teme cu unul din sufixele lexicale neologice *-ism* sau *-ist*, cu unele alternanţe vocalice, s-a construit un algoritm de derivare cu sufixul lexical verbal *–iza*, în baza căruia a fost elaborat un modul în limbajul C în mediul de programare KDevelop (sistemul de operare Linux OpenSuse 10.3), care generează cuvinte noi. În baza algoritmului expus în (Petic, 2008c) în cazul terminaţiilor *-atic, -etic, -otic* şi *-ific* s-au generat automat 420 de verbe.

6.2.2. Cazul sufixelor lexicale –bil şi –tor

Din cele stabilite anterior se formează derivate cu sufixele *-tor* şi *-bil* de la infinitivul prezent al verbelor (de exemplu: *cititor*, *caracterizabil*), în unele cazuri cu careva alternanţe vocalice sau consonantice (de exemplu: *dogorî→dogoritor,*

dispune→disponibil). Drept sursă pentru generarea noilor derivate cu sufixe a servit lexiconul RRTLN, în care sunt înregistrate 7796 de verbe distincte. Ţinând cont că în lexicon sunt derivate cu sufixul *-tor* şi *–bil*, pentru care au fost utilizate verbe, rămâne să fie folosite celelalte verbe pentru a fi generate noi cuvinte pentru lexiconul RRTLN.

Astfel pentru *-tor*, mai întâi, se vor genera derivate de la verbele care se termină în *a* şi *i* fără alternanţe vocalice/consonantice, ca după această să se încerce să se genereze cele în *a, i* şi *e* cu alternanţe vocalice şi consonantice. Numărul de cazuri posibile pentru *-tor* de la verbe în *a* şi *i* fără alternanţe este de 1140.

Pentru *-bil* se va proceda la fel doar pentru verbele în *a* şi *i* fără alternanţe şi după aceea cu alternanţe pentru literele *a, i* şi *e*. Numărul de cazuri posibile pentru *-bil* de la verbe în *a* şi *i* fără alternanţe este de 1962.

7. Problema validării şi flexionării derivatelor generate automat

Derivatele noi, care au fost generate, ar trebui să fie corecte din punct de vedere morfologic şi semantic. Unul din procedeele de validare a derivatelor constă în validarea manuală a fiecărui cuvânt generat în corespundere cu cerinţele regulilor morfologice şi semantice. Garantând calitatea rezultatului (în cazul când procedeul este efectuat de către un specialist în domeniu) ne confruntăm cu dezavantajele specifice unui lucru manual: resurse considerabile de timp, precum şi posibilitatea comiterii unor erori.

Un alt mod de validare constă în verificarea prezenţei cuvintelor derivate în documentele electronice existente pe Internet. Căutarea în Internet trebuie să se realizeze pentru documentele culese doar pentru limba română. Aici însă ne confruntăm cu o serie de dificultăţi. Chiar cu opţiunea cu privire la limba setată este posibil să se găsească cuvinte în alte limbi. Este cazul cuvintelor *maciza* (limba spaniolă)*, bariza* (limba arabă)*, neautomobil* (limba cehă)*, nemonolit* (limba croată) care au fost găsite de către motorul Google la căutare pentru limba română. În plus, apar deficienţe create de o eventuală segmentare a cuvântului căutat. Astfel, de exemplu, la încercarea de a valida în acest mod verbul *fataliza* s-a găsit *„...o fată, Liza...”,* în loc de *cristianiza* s-a găsit *Cristian Iza.* Există şi cuvinte care reprezintă substantive proprii, în particular denumiri de companii, validitatea cărora trezeşte dubii, de exemplu, *SRL „Daniza”* şi *SRL „Cariza”* găsite de către motorul de căutare nu pot confirma validitatea verbelor omonime generate automat.

Dincolo de cele expuse mai sus apare şi dificultatea stabilirii condiţiilor în care un cuvânt este valid. S-ar părea că numărul de apariţii ale cuvântului în listă ar fi un criteriu obiectiv. De exemplu, pentru cuvântul *catiza* s-au găsit mai multe intrări, dar cu greu s-ar găsi argumente pentru a-l valida.

Mai pot apărea cazuri în care se formează un cuvânt derivat cu alt sens, de exemplu *negros* format de la adjectivul *gros*, ar trebui să aibă sensul „subţire”. În DEX există un astfel de cuvânt *negros* dar are alt sens „brun, brunet, negricios”.

O altă problemă este cea a stabilirii părţii de vorbire pentru derivatele obţinute în mod automat şi flexionării lor ulterioare.

Tabel 1: Date statistice despre numărul cuvintelor derivate automat

Afixul	**Num. de cuvinte generate automat**	**Num. de cuv. validate manual**	**Num. de cuv. validate automat**	**Num. de flexiuni**
ne-	397 (100%)	362 (91%)	187 (47%-52%)	3740
-iza	420 (100%)	317 (75%)	76 (18%-24%)	2920
Total	817 (100%)	679 (82%)	263 (32%-39%)	6660

Cuvintele derivate cu prefixul *ne-* validate în mod manual, cât şi folosind mijloacele *google.com* ca parte de vorbire sunt adjective. Aceste cuvinte vor moşteni clasa de flexionare de la baza derivatului. În procesul de flexionare au apărut unele situaţii ambigue soluţionate cu ajutorul mijloacelor motorului de căutare *google.com*.

În cazul flexionării cuvintelor derivate cu sufixul lexical verbal *–iza* a apărut o situaţie problematică la stabilirea verbelor personale şi impersonale. Determinarea acestui lucru s-a realizat manual, prin consultarea cu un specialist filolog.

La flexionarea derivatelor verbale cu prefixul *re-* de asemenea trebuie de stabilit dacă verbul este personal sau impersonal. Cuvintele derivate, conform legităţilor de derivare cu prefixe, vor moşteni clasa de flexionare de la baza derivatului.

Sufixul *–re,* care transformă verbele la infinitiv în substantive de genul feminin, se flexionează ca şi toate celelalte substantive care se termină în *–re*, astfel cuvintele derivate nu vor moşteni clasa de flexionare de la baza derivatului.

Deoarece cuvintele sufixate cu *-tor* şi *–bil* pot fi atât substantive cât şi adjective, în fiecare caz aparte trebuie să decidem, dacă într-adevăr le putem flexiona ca substantive sau ca adjective.

8. Concluzii

Aplicaţiile ce ţin de derivarea automată s-au dovedit a fi destul de utile, în particular, pentru completarea resurselor lingvistice electronice, stabilirea legăturii între derivatele cu sufixe şi semantică, generarea prin derivare a unor leme cu sensuri previzibile, generarea familiilor morfologice ale unui cuvânt pentru a reduce varietatea lingvistică de documente indexate.

Studierea particularităţilor procesului de derivare a permis stabilirea asemănărilor şi deosebirilor care apar la procesul de prefixare şi sufixare în limba română. Aceasta a condus la elaborarea algoritmilor necesari în prelucrarea cuvintelor derivate atât cu sufixe cât şi cu prefixe. Totuşi n-a fost posibil de a recunoaşte toate cuvintele derivate, în special, cu sufixe, cauza majoră fiind prezenţa alternanţelor vocalice în procesul de derivare şi lipsa în lexicoane a cuvintelor lemă ale tuturor cuvintelor derivate.

Rezolvarea problemei generării unor derivate noi inexistente în dicţionare a fost ilustrată în baza unor afixe concrete. Aceasta a permis evaluarea rezultatelor pentru fiecare afix în parte. Totodată nu toate cuvintele generate pot fi considerate acceptabile, ci doar cele care au trecut printr-un proces de validare.

Cuvinte considerate acceptate au fost flexionate în mod automat cu ajutorul programelor existente, astfel valorificând posibilităţile şi completând arsenalul lexicografic al lexiconul RRTLN al limbii române.

Mulţumiri. Sunt recunoscător dnei. dr. hab. Svetlanei Cojocaru şi dnei dr. Elena Boian pentru ajutorul acordat la realizarea acestei lucrări.

Referinţe bibliografice

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

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

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

Carota F. (2006). Derivational Morphology of Italian: Principles of Formalization, *Literary and Linguistic Computing*, Vol. 21, Suppl. Issue, pp. 41-53.

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

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

Duško V., Krstev C. (2005). Derivational Morphology in a E-Dictionary of Serbian, In Zygmunt Vetulani (ed.), *Proceedings of the 2nd Language & Technology Conference*, Poznan, Poland, pp. 139-143.

Graur Al., Avram M. (1978). *Formarea cuvintelor în limba română*, vol. II Editura Academiei, Bucureşti.

Hristea T. (1984). *Sinteze de limba română*, Bucureşti, pp. 66-99.

Iordan I. (1970). *Limba română contemporană*, Editura Academiei, Bucureşti.

Leavitt, JR. (1994). MORPHE: A Morphological Rule Compiler. Technical Report, CMU-CMT-94-MEMO.

Lombard A., Gâdei C. (1981). *Dictionnaire morphologique de la langue roumain*, Bucureşti, Editura Academiei, 232 p.

Nyberg, E. H., Mitamura, T. (1992). The KANT System: Fast, Accurate, High Quality Translation in Practical Domains. In: *Proceedings of COLING92*.

Petic M. (2007). Automatic extraction of the analysable formations with simple prefixes. *Proceedings of the Second International Conference of Young Scientists „Computer Science and Engineering-2007”*, Lvov, pp. 215-217.

Petic M. (2007). Derivarea automată cu prefixele ne- şi re- pentru adjective şi verbe. *Proceedings of the International Conference BIT+2007*, Chişinău.

Petic M. (2008). Specific features in automatic processing of the formations with prefixes, *Computer Science Journal of Moldova*, 4 1(7), pp. 209-222.

Petic M. (2008). Probleme în popularea resurselor lingvistice electronice prin derivarea automată cu sufixul lexical verbal –iza. *Proceedings of the International Conference BIT+2008*, Chişinău.

Petic M. (2008). Generarea automată a verbelor cu sufixul lexical –iza. *The 2nd International Conference „Telecommunications, Electronics and Informatics*. *Procedeengs*. Volume I, Chişinău, pp. 441-446.

Santana O, Perez J., Carreras F., Rodriges G. (2004). Suffixal and Prefixal Morpholexical Relationships of Spanish. *Lecture Notes in Artificial Intelligence*, Ed. Springer-Verlag, pp. 407-418.

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

Tufiş D., Diaconu L., Barbu A. M., Diaconu C. (1996). Morfologia limbii române, o resursă lingvistică reversibilă şi reutilizabilă, *Limbaj şi Tehnologie*, Editura Academiei Române, Bucureşti, pp. 59-65.

Vilares J., Cabrero D. M., Alonso A. (2001) Applying Productive Derivational Morphology to Term Indexing of Spanish Texts Source Lecture Notes In Computer Science; Vol. 2004, *Proceedings of the Second International Conference on Computational Linguistics and Intelligent Text Processing*, pp. 336 — 348.

INDEX DE AUTORI