Revista Română de Interacţiune Om-Calculator 8 (2) 2015, 101-120 

# RACAI-RoTb: nucleu de corpus de limbă română adnotat sintactic cu relaţii de dependenţă

Elena Irimia, Verginica Barbu Mititelu

Institutul de Cercetări pentru Inteligenţă Artificială "Mihai Drăgănescu", Academia Română
Calea 13 Septembrie nr. 13, CASA ACADEMIEI, 050711, Bucureşti
*E-mail: elena@racai.ro, vergi@racai.ro*

**Rezumat.** Articolul prezintă activitatea de creare a unei colecţii de arbori sintactici (eng. *treebank*) pentru limba română, alcătuite din 5000 de propoziţii adnotate sintactic cu formalismul gramaticii de dependenţe. În introducere se argumentează necesitatea construirii unei astfel de resurse în contextul unui important deficit de reprezentare electronică a limbii române, în raport cu alte limbi de circulaţie internaţională. Pe plan internaţional, *treebank*-uri de mari dimensiuni (sute de mii de propoziţii) au început să fie dezvoltate încă din anii '90 (secţiunea 2.1), în timp ce puţinele iniţiative dedicate limbii române nu numără mai mult de câteva mii de propoziţii (secţiunea 2.2). Vom detalia în continuare modul de selectare a propoziţiilor în corpus (Secţiunea 3), gramatica de dependenţe utilizată (Secţiunea 4), metodologia de adnotare (Secţiunea 5), precum şi rezultatele evaluării adnotării automate în raport cu corectura manuală ulterioară (Secţiunea 6). Pentru a grăbi procesul de adnotare, se pleacă de la o adnotare automată statistică (cu un model statistic de analiză sintactică pentru limba spaniolă) care trece prin două etape succesive de corectură manuală, cu doi specialişti lingvişti.

**Cuvinte cheie**: corpus, colecţie de arbori sintactici, adnotare automată, adnotare sintactică, gramatică de dependenţe.

## 1. Introducere

În toate domeniile bazate pe interacţiunea om-maşină, unde limbajul verbal reprezintă o componentă esenţială, capacitatea instrumentelor software de a procesa şi de a genera limbaj natural la performanţe cât mai ridicate este o preocupare centrală. Cu cât limbajul natural este mai amănunţit procesat şi adnotat, cu cât aplicaţiile dispun de mai multe informaţii lingvistice de orice nivel (lexical, morfologic, sintactic, semantic, discurs etc.) şi cu atât mai profesionist şi performant se comportă sisteme complexe precum cele de traducere automată, răspuns automat la întrebări, e-learning, procesare de text (tip Microsoft Word, cu modulele sale de corectare ortografică şi

sintactică), aplicații inteligente din diverse domenii (telecomunicații, transport, medicină, servicii online sau telefonice orientate către clienți, activitate editorială, industrie, securitate etc.). Filip şi Leiviska (2009) califică subsistemul de limbaj şi comunicare drept una dintre cele trei componente esențiale ale oricărui sistem complex de mari dimensiuni (şi ale sistemelor de suport pentru decizii din cadrul acestora), alături de componenta cunoaşterii şi cea a procesării de probleme.

Nivelul de analiză sintactică a limbajului se află în aria de interes a comunității ştiințifice şi industriale internaționale preocupate de modelarea automată a limbajului deja de mult timp, dar perspectiva națională întârzie să se alinieze acestor abordări. De exemplu, în Traducerea Automată Statistică, diverse încercări de modelare statistică la nivel sintactic au început încă din anii 2000 (vezi (Och et al., 1999), (Marcu şi Wong, 2002), (Yamada şi Knight, 2002), (Zens şi Ney, 2004), (Chiang, 2007)), în timp ce în România încă nu se cunosc astfel de abordări, aplicațiile limitându-se la informația lexicală şi morfologică pentru a propune traduceri. Depunerea efortului de cercetare pentru construirea unei resurse electronice de tipul colecție de arbori sintactici (vom folosi în continuare termenul englezesc *treebank*) este pe deplin justificată de utilitatea acesteia drept corpus de antrenare pentru un analizor sintactic, instrument ce deschide calea către integrarea acestui nivel de analiză în diverse aplicații precum cele menționate în paragraful anterior.

Din perspectiva lingvisticii teoretice, existența unui corpus adnotat la nivel morfosintactic va oferi posibilitatea căutărilor avansate: înlănțuiri de cuvinte, înlănțuiri de etichete morfologice şi chiar lanțuri de relații sintactice. Pe baza rezultatelor găsite se pot susține, completa sau ajusta teoriile lingvistice. De exemplu, pentru limba engleză, G. Sampson (în Abeille, 2003) ilustrează cum studiile pe un corpus adnotat la nivel sintactic au scos în evidență faptul că, în limba engleză, propozițiile de tipul subiect-verb intranzitiv sunt mult mai puțin frecvente decât se susținea în anumite manuale lingvistice.

În cadrul proiectului META-NET, într-un studiu amplu ce îşi propunea inventarierea resurselor şi instrumentelor lingvistice pentru 31 de limbi europene (http://www.meta-net.eu/whitepapers/overview), în raportul despre limba română (Trandabăț et al., 2012) se remarcă, printre alte lipsuri de aşteptat pentru o limbă slab reprezentată electronic, absența unui treebank de referință. Răspunzând semnalului de alarmă tras de autorii

raportului şi împărtăşind opinia acestora că o tehnologie adecvată a limbajului este esențială pentru supraviețuirea unei limbi în era informațională, ne-am propus să contribuim la alcătuirea unui treebank de referință pentru limba română.

În mod concret, urmărim să realizăm un set de câteva mii de fraze analizate sintactic, cuprinzând verbe frecvent utilizate în limbă, cu structuri sintactice complexe şi diverse. Am avut în vedere acoperirea mai multor stiluri funcționale şi domenii, pentru a obține o resursă diversă şi adecvată ştiințific, oferind un model la scară redusă al tiparelor sintactice din limba română. De asemenea, pentru că o asemenea întreprindere ar necesita un efort uman important dacă ar fi îndeplinită manual, ne propunem o automatizare cât mai completă a procesului de construire a băncii de arbori: de la extragerea automată a verbelor cu frecvență ridicată în corpusul de interes, la selectarea propozițiilor ce conțin aceste verbe, în funcție de criterii precum lungimea propoziției sau prezența unui verb predicativ în structura propoziției, la folosirea unui adnotator statistic pentru adnotarea automată a corpusului şi folosirea unui editor grafic pentru corectarea acestuia, la evaluarea rezultatelor obținute utilizând instrumente software consacrate în competițiile internaționale de analiză sintactică automată.

În acest moment (iunie 2015), corpusul conține 2000 de fraze (cu lungime între 10 şi 30 de cuvinte) adnotate automat şi corectate manual de doi specialişti lingvişti. Până în toamna anului 2015 sperăm să ajungem la un nucleu de treebank românesc cu 5000 de fraze. În secțiunea 3 vom detalia proveniența şi modul de selecție a acestor fraze.

## 2. Stadiul actual în domeniu

### 2.1. Contextul internațional

Utilizarea corpusurilor electronice, de către lingvişti şi ingineri din domeniul Prelucrării Limbajului Natural (PLN) deopotrivă, are deja o istorie de zeci de ani, în special în context internațional. Aplicațiile de prelucrare şi modelare a limbajului natural au fost inițial bazate pe reguli construite prin efortul susținut al cercetătorilor lingvişti. Cu timpul au luat avânt metodele statistice care funcționează extrăgând automat modele lingvistice din corpusuri electronice de mari dimensiuni. Reducând foarte

mult efortul uman, aplicațiile statistice au, în acelaşi timp dezavantajul de fi dependente de particularitățile datelor de antrenare şi de a nu fi capabile să gestioneze fenomene lingvistice pe care nu le regăsesc în aceste date. De aceea, în ultimii ani au câştigat teren *metodele hibrid*, care combină cunoştințe linvistice explicite cu metode de extragere automată a cunoştințelor implicit codificate în corpusurile electronice.

Astfel, deşi inițial modelele statistice se bazau pe text neprocesat şi neadnotat, cu timpul au apărut abordări care presupun adnotarea textului înainte de învățarea modelelor, la diferite niveluri lingvistice: la început doar la nivel morfo-lexical, ulterior la nivel sintactic şi chiar semantic.

Primele corpusuri analizate sintactic au fost treebank-ul Lancaster (LPC, eng. Lancaster Parsed Corpus, Garside et al., 1992) şi Penn TreeBank (Taylor et al., 2003). Realizate în anii '90, au constituit modele de urmat pentru numeroase alte proiecte asemănătoare precum băncile de arbori germane NEGRA (Skut at al., 1997), TIGER (Brants at al., 2004), corpusurile scrise sau vorbite TüBa, realizate la Tübingen pentru limbile germană, engleză şi japoneză (http://www.sfs.uni-tuebingen.de/en/ascl/resources/corpora.html), treebank-ul cehesc Prague Dependency Treebank (Hajič et al., 2001), pentru a le enumera doar pe cele mai importante. Interesul pentru acest tip de resursă a crescut continuu, conducând la dezvoltarea de treebank-uri pentru limbile arabă, bulgară, catalană, chineză, coreeană, croată, daneză, ebraică, estoniană, finlandeză, franceză, greacă, hindu, islandeză, italiană, latină, norvegiană, olandeză, persană, poloneză, portugheză, română, rusă, slovenă, spaniolă, suedeză, thai, turcă, ungară, urdu, vietnameză.

Majoritatea corpusurilor adnotate la nivel sintactic enumerate sunt resurse de mari dimensiuni, atingând un număr de sute de mii de propoziții, în timp ce unele proiecte, inclusiv corpusurile româneşti, numără doar câteva mii de propoziții. În cazul corpusurilor mari, performanțele se datorează unor echipe de lucru numeroase, cuprinzând atât informaticieni, cât şi lingvişti, care au înțeles importanța ştiințifică, culturală şi strategică a unei astfel de resurse şi au investit uneori aproape un deceniu în atingerea acestui scop.

## 2.2. Contextul național

Interesul pentru realizarea unei bănci de arbori sintactici pentru limba română s-a manifestat încă de la începutul anilor 2000. Dovadă stă

realizarea unei astfel de resurse în cadrul proiectului RORIC-LING (Hristea şi Popescu, 2003). Formalismul gramatical utilizat este gramatica de dependențe, iar propozițiile reflectă stilul jurnalistic. Rezultatul proiectului este un treebank de 4042 de arbori (i.e., de propoziții adnotate), a căror lungime medie este de nouă cuvinte. Este, în mod evident, un corpus cu propoziții scurte. De altfel, autorii au evitat cazurile lingvistice problematice prin includerea exclusiv a propozițiilor, nu şi a frazelor. Frazele au fost segmentate în propoziții, fiecare dintre acestea fiind analizate separat, manual (http://www.phobos.ro/roric/DGA/dga.html). Acest mod de analiză nu este adecvat: el eşuează în a reflecta, de exemplu, cazurile în care un argument verbal se realizează ca subordonată. Autorii au dezvoltat şi o interfață grafică de adnotare (Popescu, 2003), care porneşte de la text complet neadnotat, fără nici un fel de informație morfo-lexicală. Un rezultat important al acestui proiect, la care ne-am raportat şi în munca noastră, este crearea unui inventar de relații de dependență pentru limba română (Hristea şi Popescu, 2003).

În ciuda caracteristicilor sale (simplitate, inadecvare a metodei de lucru), acest treebank a fost, totuşi, folosit la antrenarea unui analizor sintactic (în engleză *parser*) pentru limba română (Călăcean şi Nivre, 2009). Considerăm că, din cauza neajunsurilor corpusului de antrenare, acest parser ar fi util într-o măsură foarte mică, dacă ar fi accesibil.

Un alt treebank pentru limba română, nefinalizat şi nedisponibil în momentul în care am început documentarea şi munca la proiectul nostru, este descris în (Perez, 2014) şi (Mărănduc şi Perez, 2015). Cele 4.500 de propoziții (115.000 de cuvinte, cu o lungime medie de 37 de cuvinte/propoziție) sunt adnotate manual cu relații specifice gramaticii de dependențe, cu ajutorul unei interfețe dezvoltate special, TreeAnnotator. Corpusul acoperă diferite stiluri funcționale din diferite perioade istorice: traducerea în limba română a FrameNet[1]-ului englezesc, traducerea romanului „1984” al lui G. Orwell, texte beletristice româneşti, documente din Wikipedia şi din Acquis-ul Comunitar[2], texte politice etc. Complexitatea sintactică a corpusului este evaluată în termeni de abundență a punctuației şi

[1] https://framenet.icsi.berkeley.edu/fndrupal/about

[2] totalitatea drepturilor si a obligatiilor comune care decurg din statutul de stat membru al Uniunii Europene

de acoperire a tipurilor de relații sintactice: autorul raportează că relația de punctuație reprezintă 12% din totalul relațiilor din textele beletristice şi că toate relațiile de dependență din gramatica limbii române sunt reprezentate în treebank, în procente variind de la 13,64% la 0,04% (unele relații nu sunt niciodată reprezentate în secțiunile Wikipedia sau Acquis Comunitar).

Pe o parte din acest treebank a fost antrenat Malt Parser (Colhon şi Simionescu, 2012), cu rezultate destul de bune, evidențiate în faza de testare. Din păcate, contextul în care am desfăşurat proiectul de față nu a fost favorabil utilizării acestui parser.

Seretan et al. (2010) anunță adaptarea analizorului Fips pentru adnotarea textelor româneşti. Nici regulile gramaticale specifice limbii formulate de autori, nici corpusul rezultat după rularea parserului nu sunt disponibile. Mai mult, nici parserul accesibil pe site-ul proiectului (http://latlapps.unige.ch/Parser) nu este disponibil pentru limba română.

Un alt corpus românesc (jurnalistic) adnotat la nivel sintactic este raportat în Bick şi Greavu (2010). Adnotarea se face cu un parser (VISL) a cărui gramatică a fost scrisă prin adaptarea celei pentru limba italiană. Formalismul gramatical adoptat în VISL este Constrained Grammar. Corpusul (de peste 21 de milioane de cuvinte) poate fi vizualizat prin căutări efectuate la adresa http://corp.hum.sdu.dk/cqp.ro.html.

## 3. Asamblarea corpusului

Pentru ca resursa construită de noi să constituie o bază solidă pentru crearea unui model statistic de analiză sintactică cât mai fiabil, am ales să pornim de la un corpus balansat[3], ROMBAC, considerând că acesta poate oferi un model la scară redusă al şabloanelor sintactice din limba română. Corpusul acoperă 4 stiluri funcționale (beletristic, publicistic, oficial şi ştiințific), distribuite în cinci secțiuni (proză, juridic, medical, academic şi jurnalistic).

ROMBAC este un corpus balansat adnotat la nivel morfosintactic (lema, partea de vorbire şi categoriile gramaticale specifice fiecăreia) şi la nivel de grup sintactic, dezvoltat la ICIA (Ion et al., 2012) şi disponibil gratuit prin platforma META-SHARE (http://www.meta-share.eu/).

[3] Un corpus balansat (general sau specializat) acoperă un spectru larg de categorii de texte, fiecare categorie reprezentată aproximativ prin acelasi număr de cuvinte.

Folosind informația morfo-sintactică din ROMBAC, am putut extrage automat liste cu cele mai frecvente 500 de verbe în fiecare din cele 5 secțiuni. Luând drept criteriu de selecție frecvența în corpus a verbelor, ne concentrăm pe cel mai des întâlnite structuri sintactice verbale în limba română. Cum nu avem pretenția de a crea o resursă de mari dimensiuni, ne dorim să includem în acest nucleu de treebank informație cât mai relevantă pentru un model statistic.

În mod natural, anumite verbe se vor regăsi în mai multe sau chiar în toate subdomeniile, lucru ce va fi util unor cercetări lingvistice: de exemplu, se poate studia comportamentul sintactico-semantic al unui verb în diverse registre lingvistice sau chiar domenii ştiințifice, remarcându-se fie identitatea comportamentală de la un domeniu/registru la altul, fie diferențe în natura semantică a unor argumente (în cazul în care verbul este folosit cu acelaşi cadru în toate domeniile/registrele), sau în numărul de argumente, deci utilizarea verbului cu cadre diferite de la un domeniu/registru la altul.

Pentru fiecare din cele 2500 de verbe selectate astfel, s-au recuperat în mod automat din sub-corpusurile corespunzătoare toate propozițiile care le conțin, au o lungime mai mare de 10 cuvinte şi cel puțin un verb predicativ în structură. Din frazele recuperate am selectat câte 2 pentru fiecare verb, în total 5000 de fraze, care, o dată analizate sintactic, vor constitui un nucleu de treebank.

## 4. Descrierea teoretică a gramaticii folosite în adnotare

### 4.1. Caracteristicile gramaticii de dependențe

Gramatica de dependențe este o clasă de teorii sintactice a căror reprezentare sintactică de bază este relația de dependență. Aceasta se stabileşte între două cuvinte dintr-o propoziție, unul fiind centru, celălalt dependent şi se reprezintă ca un arc orientat între centru şi dependent.

Principiile gramaticii de dependență sunt:

- un cuvânt dintr-o propoziție depinde doar de un alt cuvânt (numit centru) din aceeaşi propoziție, cu excepția verbului (din propoziția principală), care nu depinde de niciun alt cuvânt;
- mai multe cuvinte pot depinde de acelaşi centru;

- arcele nu trebuie să se intersecteze; graful format din aceste arce nu conține cicluri.

Structura rezultată în urma analizei cu gramatica de dependență:

- este minimală – fiecare nod din structură este un cuvânt din propoziția analizată; nu există noduri artificiale în graf;
- este ordonată – ordinea nodurilor reflectă ordinea cuvintelor în propoziție;
- permite dependențe multiple – un centru poate stabili relații sintactice cu mai mulți dependenți în cadrul aceleiași propoziții.

## 4.2 Inventarul de relații de dependență folosit

Metodologia de adnotare (detaliată în Secțiunea 5) presupune o etapă de adnotare automată cu un model sintactic statistic antrenat pe corpusul IULA LSP, adnotare care vine cu un set de etichete specific, denumit în continuare *iulaLSPdep*. Într-o etapă ulterioară, de corectură manuală a acestei adnotări, am avut ocazia să intervenim asupra acestui set de etichete, pe care l-am îmbogățit cu relații din proiectul Universal Dependency (*UD*) (http://universaldependencies.github.io/docs/u/dep/index.html), dar şi cu altele propuse de noi (pentru a descrie relații specifice limbii române). Intenția noastră a fost crearea unui set de relații specifice (denumit în continuare *ROdep*), care să conducă la rezultate echilibrate în următoarele privințe: 1) corectură manuală cât mai redusă a analizei automate; 2) realizarea unui corpus în conformitate cu standardele gramaticii româneşti; 3) alinierea, pe cât posibil, la inventarul UD.

În tabelul de mai jos prezentăm, comparativ, inventarul de relații de dependență folosit pentru analiza limbii române, pe cel folosit pentru limba spaniolă (*iulalLSPdep*) şi pe cel utilizat în proiectul UD:

*Tabelul 1. Tablou comparativ al relațiilor sintactice în română, spaniolă şi Universal Dependecy*

| ROdep | iulaLSPdep | UD |
|---|---|---|
| acl | MOD | acl |
| advcl | MOD | advcl |
| advmod | MOD, PP-LOC, PP-DIR, ADV | advmod |
| agc | BYAG | - |
| amod | SPEC | amod |
| appos | MOD | appos |
| aux | AUX | aux |
| auxpass | | auxpass |
| cc | COORD | cc |

| | | |
|---|---|---|
| compound | | compound |
| conj | CONJ, ENUM | conj |
| correl | | |
| dblclitic | | expl |
| det | SPEC | det |
| dislocated | | dislocated |
| dobj | DO | dobj, ccomp, iobj |
| foreign | | foreign |
| goeswith | | goeswith |
| iobj | IO | iobj, ccomp |
| list | | list |
| mark | SPEC | mark |
| mwe | | mwe |
| name | | name |
| discourse | | discourse |
| neg | NEG | neg |
| nmod | MOD | nmod |
| parataxis | | parataxis |
| passmark | PASSM | - |
| pmod | MOD, PP-LOC, PP-DIR | - |
| pobj | OBLC | - |
| poss | | - |
| possclitic | | - |
| post | | case |
| pred | PRD, ATR | root, xcomp, ccomp |
| prep | COMP | case |
| punct | PUNCT | punct |
| reflclitic | PRNM, IMPM | expl |
| remnant | SUBJ-GAP, COMP-GAP, MOD-GAP | remnant |
| reparandum | | reparandum |
| root | ROOT | root |
| sc | | - |
| secobj | | dobj |
| spe | OPRD | xcomp |
| subj | SUBJ | nsubj, nsubjpass, csubj, cubjpass |
| voc | VOC | vocative |
| xcomp | OPRD | xcomp |

Diferența majoră între adnotarea *iulaLSPdep* şi cea *UD* (caz în care noi am optat pentru strategia de adnotare *iulaLSPdep*) constă în modul de tratare a cuvintelor funcționale:

- în *UD,* prepozițiile şi conjuncțiile nu pot fi centre de grup sintactic, ci doar determinanți: prepozițiile sunt legate prin relația *case* de nominalul pe care-l însoțesc, conjuncțiile coordonatoare sunt

dependente de primul dintre conjuncți, iar cele subordonatoare sunt *mark* față de centrul propoziției subordonate.

- în *iulaLSPdep*, prepozițiile sunt centre de grup sintactic, conjuncțiile coordonatoare sunt dependente de primul dintre conjuncți, iar conjuncțiile subordonatoare sunt centrul propoziției pe care o introduc (centrul verbal din subordonată stabilește relația *sc* cu conjuncția).

Pe de altă parte, am preluat din *UD* tratarea verbelor la moduri nepredicative (infinitiv, participiu, supin și gerunziu) ca centre de propoziții subordonate. Această abordare este distinctă de cea a gramaticii românești, însă ne permite adnotarea consecventă a verbelor și a argumentelor lor.

### 4.2.1. Relații preluate din *iulaLSPdep*

Modul de preluare și adaptare de către noi a relațiilor din *iulaLSPdep* cunoaște următoarele forme:

1. **preluare fără modificări:**

- în *iulaLSPdep* nu există diferențe de adnotare între realizarea lexicală și cea propozițională a unui argument al unui predicat: un subiect exprimat prin substantiv, pronume, numeral sau subordonată este întotdeauna analizat ca *SUBJ* în *iulaLSPdep* și ca *subj* în treebank-ul nostru. În schimb, în *UD*, subiectul este de patru tipuri: *nsubj* (realizat nominal într-o propoziție cu verbul la diateza activă), *nsubjpass* (realizat nominal într-o propoziție cu verbul la diateza pasivă), *csubj* (realizat ca subordonată într-o propoziție cu verbul la diateza activă) și *csubjpass* (realizat ca subordonată într-o propoziție cu verbul la diateza pasivă);
- alte relații din *iulaLSPdep* care au fost preluate ca atare: *ROOT* (centrul propoziției), *NEG* (pentru marcatorul de negație), *VOC* (pentru vocativ) și *PUNCT* (pentru marcarea punctuației).

2. **preluare prin schimbarea numelui relației, dar folosirea pentru aceleași fenomene lingvistice:**

- obiectul direct, indiferent de realizarea sa, este *DO* în *iulaLSPdep* și *dobj* la noi, iar obiectul indirect este *IO* în *iulaLSPdep* și *iobj* la noi. Numele acestor relațiilor sunt preluate în *ROdep* din *UD*, dar modul de analiză este cel din *iulaLSPdep*. În *UD*, obiectul direct și cel indirect marchează tipuri diferite de relații doar în cazul realizării lor

nominale; dacă sunt realizate propoziţional, se foloseşte pentru ambele cazuri aceeaşi etichetă de relaţiei: *ccomp*;

- *BYAG* devine la noi *agc* (complement de agent): În *UD*, el nu se marchează printr-o relaţie specială, ci ca *nmod*, cu un determinant legat prin *case* (prepoziţia care îl introduce). Deşi element cu ocurenţă opţională în propoziţie, complementul de agent apare în cadrul de subcategorizare al verbului centru al propoziţiei, motiv pentru care am convenit să-l marcăm diferit de alţi modificatori substantivali;
- *OBLC* devine la noi *pobj* (obiect prepoziţional): acesta este un determinant obligatoriu al predicatului, care are ca centru o prepoziţie selectată de acesta. Statutul de complinire obligatorie a predicatului ne determină să-l tratăm diferit de grupurile prepoziţionale care funcţionează ca modificatori (şi care sunt analizate ca *pmod*, vezi mai jos). Imposibilitatea prepoziţiei de a fi centru de grup în *UD* face ca acestei relaţii să nu îi corespundă vreo relaţie în *UD*;
- *COMP* devine la noi *prep* (complementul prepoziţiei în grupul prepoziţional);
- *OPRD* devine la noi *spe* (predicativul suplimentar).

**3. rafinarea relaţiilor prea generale:**

- *AUX*: am marcat diferit verbele auxiliare în funcţie de diateza la care se află verbul pe care îl însoţesc: pentru diateza activă am folosit relaţia *aux*, iar pentru diateza pasivă am folosit relaţia *auxpass,* după modelul oferit de *UD*;
- *MOD*: în setul *iulaLSPdep*, desemnează orice tip de modificator (element a cărui apariţie în propoziţie este facultativă), indiferent de partea de vorbire care îl realizează; după model *UD*, am ales să distingem între: *nmod* (modificator realizat ca substantiv), *advmod* (modificator realizat ca adverb) şi *appos* (apoziţia, ce poate avea şi realizare propoziţională). În plus faţă de *UD*, adnotăm şi *pmod*, un modificator realizat ca grup propoziţional;
- *SPEC* (specificator) are ca echivalenţi în adnotarea noastră *amod* (modificator adjectival) şi *det* (determinatorii, adică articolele), după modelul oferit de *UD*.

**4. unificarea unor relaţii:**

- am considerat că diferenţierea între *PP-LOC* (complement circumstanţial de loc) şi *PP-DIR* (complement circumstanţial de loc,

ce indică direcția) nu este justificată dacă nu se fac şi alte diferențieri semantice între adjuncți. În consecință, aceste tipuri de complemente circumstanțiale au fost adnotate ca *advmod* sau *pmod*, în funcție de realizarea lor;

- în *iulaLSPdep*, elementele unei enumerări se marchează cu relația *ENUM*, cu excepția ultimului element din enumerare, care este marcat, ca şi elementele unei coordonări (de doi termeni), cu relația *CONJ*. Noi am decis să folosim doar relația *conj* pentru a marca orice fel de coordonare, inclusiv a elementelor dintr-o enumerare;
- pentru limba spaniolă s-au folosit două relații diferite pentru adnotarea numelui predicativ: *ATR* (atunci când verbul copulativ este *ser* sau *estar*) şi *PRD* (pentru numele predicative ale celorlalte verbe copulative). În spiritul lingvisticii româneşti, am decis să nu folosim etichete diferite pentru aceeaşi funcție sintactică, indiferent de regentul ei, aşadar am folosit relația *pred*;
- în funcție de valoarea sa, reflexivă sau impersonală, în *iulaLSPdep* se folosesc două relații pentru pronumele reflexiv: *PRNM* şi *IMPM*. Ambele cazuri, precum şi utilizarea cu voloare reciprocă a aceluiaşi pronume, sunt acoperite de relația *reflclitic* în limba română.

Există relații în *iulaLSPdep* pe care nu le-am preluat: *COORD, SUBJ-GAP, COMP-GAP, MOD-GAP*. Prima se foloseşte pentru adnotarea conjuncției coordonatoare, iar ultimele marchează cuvintele cu rol de subiect, complement al unei prepoziții, al unui adverb sau adjectiv, respectiv modificator în structuri eliptice. După cum vom descrie mai jos, modul de analiză a acestor structuri în limba română este preluat din *UD*, împreună cu etichetele corespunzătoare ale relațiilor.

### 4.2.2. Relații preluate din UD

Pentru o apropiere cât mai mare de modul de adnotare folosit în *UD*, am decis să preluăm o parte dintre relațiile din acest proiect, atunci când ele oferă o analiză suficient de apropiată de spiritul gramaticii tradiționale româneşti.

Relațiile preluate din *UD* sunt de următoarele tipuri:

1. **unele adnotează fenomene sintactice:** *acl* – propoziții subordonate atributive. Reamintim faptul că, în adnotarea noastră, subordonatele pot avea drept centru şi un verb la un mod nepredicativ; *advcl* – propoziții subordonate corespunzătoare complementelor

circumstanțiale; *advmod* – complemente sau atribute exprimate printr-un adverb; *amod* – atributul adjectival; *nmod* – atributul substantival sau pronominal neprepozițional; *appos* – apoziția; *xcomp* – complementele circumstanțiale exprimate prin adjectiv; *cc* – conjuncția coordonatoare; *remnant* – elementele ocurente într-o structură eliptică; *mark* – pentru adverbele care ajută la formarea gradelor de comparație, pentru apozeme, pentru prepoziția supinului, a infinitivului şi conjuncția care însoțește fenomenul de dublare a subiectului;

2. **altele adnotează fenomene morfologice:** *auxpass* – auxiliarul de pasiv; *compound* – cuvinte compuse; *mwe* – termeni multicuvânt; *name* – nume de persoane şi entități;
3. **iar altele adnotează fenomene de discurs** (ne-adnotate în *iulaLSPdep*): *dislocated* – elemente dislocate din poziția normală în propoziție; *goeswith* – părți de cuvânt în mod greşit separate în text; *list* – pentru liste de elemente de acelaşi fel (de ex., adrese, numere de telefon etc.); *discourse* – în special pentru interjecții şi cuvinte de umplutură (conform terminologiei din comunicare) (*Ăăă*..., *păi*); *parataxis* – pentru împletirea vorbirii directe cu intervențiile naratorului, pentru propoziții incidente; *reparandum* – pentru disfluențe în vorbirea directă; *foreign* – pentru secvențe de cuvinte străine.

### 4.2.3. Relații specifice limbii române

Pentru unele fenomene lingvistice specifice limbii române nu am găsit o relație pertinentă în *iulaLSPdep* sau în *UD* şi am decis să creăm noi nume pentru aceste relații: *dblclitic* – pentru dublarea prin clitic a complementului direct sau indirect; *pmod* – pentru complementul circumstanțial realizat ca grup prepozițional şi pentru atributul prepozițional; *poss* – pentru dativul posesiv; *possclitic* – pentru cliticul în dativ, cu sens posesiv, care dublează un dativ posesiv realizat printr-un subtantiv; *post* – pentru prepozițiile care apar în postpoziție față de centru; *sc* – pentru a pune în relație centrul unei subordonate de conjuncția care introduce subordonata respectivă; *secobj* – pentru obiectul secundar.

## 5. Adnotarea automată şi corectarea manuală a corpusului

Urmând îndeaproape procedura descrisă de (Arias et al. 2014), am adnotat automat corpusul nostru folosind parserul statistic disponibil liber MaltParser[4] (Nivre and Hall, 2005) – cu un model statistic delexicalizat extras din treebank-ul pentru limbă spaniolă IULA LSP[5] (Marimon and Bel, 2014) – şi am corectat manual arborii rezultaţi.

Anterior, echipa IULA[6] folosise cu succes această strategie pentru a grăbi crearea unei bănci de arbori pentru limba catalană, motivaţi de argumente precum: 1. similaritatea tipologică între limbile catalană şi spaniolă; 2. un scor LAS (Labelled Attachment Score, Scorul de Ataşare Etichetată) foarte bun (94%) obţinut pentru modelul spaniol atunci când este utilizat pe propoziţii în limba spaniolă; 3) facilitatea oferită de MaltParser de a construi modele controlând diverse trăsături, ca de exemplu excluderea informaţiei lexicale şi utilizarea exclusivă a etichetelor morfo-sintactice. Cum fac parte din aceeaşi familie de limbi romanice, am considerat că putem utiliza similaritatea tipologică între spaniolă şi română pentru a reduce munca manuală substanţială pe care o adnotare de la zero ar presupune-o. În plus, Florea et al. (2014) raportează rezultate încurajatoare (acurateţe de peste 60%) pentru experimente de adaptare a unor parsere dedicate unor limbi romanice (franceză şi spaniolă) la limba română.

Pentru a putea aplica modelul de limbă spaniolă în analiza automată, a fost necesar să realizăm automat corespondenţa între setul de etichete morfo-sintactice[7] utilizate în corpusul nostru şi cel folosit de IULA[8] (ambele derivate din specificaţiile EAGLES) şi, de asemenea, să convertim fişierele noastre din formatul .xml în formatul CONLL utilizat de MaltParser.

Înainte de etapa de corectare, am transferat automat din setul de etichete sintactice IULA în setul nostru de etichete de dependenţe tot ce s-a putut transfera ne-ambiguu. Etichete precum SPEC sau MOD (cu mai mult de o

[4] http://www.maltparser.org/

[5] http://www.iula.upf.edu/recurs01_tbk_uk.htm

[6] http://www.iula.upf.edu/indexuk.htm

[7] http://nl.ijs.si/ME/V4/msd/html/index.html, Specificaţiile morfosintactice MultText East pentru limba română

[8] http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-es.html

etichetă echivalentă în setul românesc) au fost lăsate spre dezambiguizare în etapa de corectare.

Pentru o vizualizare arborescentă a propozițiilor analizate şi facilitarea corectării lor, am folosit editorul grafic XML yEd şi câteva scripturi Perl de import/export către/din XML oferite de echipa IULA. yEd (vezi Figura 1) oferă posibilitatea ştergerii sau adăugării de noduri noi, ştergerii, adăugării sau re-etichetării muchiilor arborelui, diverse moduri de vizualizare a structurii grafului (ierarhic, arborescent, radial, organic etc.), focalizare sau defocalizare (engl. Zoom in/out).

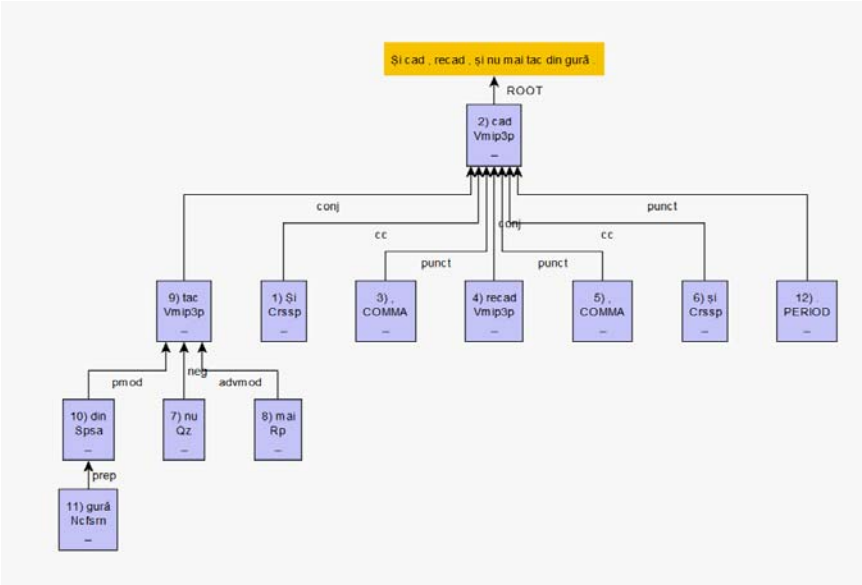


*Figura 1: Reprezentarea arborescentă cu yEd a adnotării sintactice a propoziției: „Şi cad, recad, şi nu mai tac din gură."*

După corectarea a 500 de arbori sintactici (din sub-corpusul jurnalistic), am antrenat un model de limbă română pe aceste propoziții, folosind modul de procesare *learn*[9] (din engleză, „învață"). O practică comună atunci când se utilizează instrumente statistice de prelucrare a limbajului este optimizarea acestora pe un anumit corpus: calcularea şi fixarea parametrilor modelului statistic astfel încât să producă cele mai bune rezultate posibile

[9] Pentru analiza sintactice propriu-zisă, se folose□te modul de procesare *parse*.

(în termenii unor măsuri statistice) pe un anumit set de date. Utilizând MaltOptimizer – un instrument disponibil liber, dezvoltat special pentru optimizarea analizorului MaltParser –, am antrenat un model statistic de limbă română optimizat pe aceiaşi 500 de arbori.

O evaluare pe datele de antrenare, deşi nerecomandată într-un proces de evaluare corect ştiinţific, ne-a confirmat intuiţia că modelul optimizat va fi mai bun decât cel ne-optimizat chiar şi pentru alte tipuri de text (literar, juridic, academic etc. versus jurnalistic), astfel că am folosit acest model pentru a analiza sintactic următoarea tranşă de propoziţii (am ales 500 de propoziţii din sub-corpusul beletristic). Corectura manuală a primelor 150 de propoziţii din această tranşă s-a realizat cu semnificativ mai puţine eforturi decât pentru primele 500 de propoziţii, lucru care va fi reflectat şi în rezultatele evaluărilor statistice prezentate în următoarea secţiune şi care recomandă deja înlocuirea modelului statistic spaniol cu cel optimizat pe limba română.

## 6. Evaluarea rezultatelor experimentelor

Am utilizat pentru evaluare instrumentul disponibil liber MaltEval, implementat pe baza scripturilor eval.pl şi eval07.pl, create pentru evaluările din competiţiile CoNLL[10] 2006 şi CoNLL 2007, dedicate analizei sintactice cu dependenţe (Nillson and Nivre, 2008). Aceste competiţii au consacrat Label Atacchment Score (LAS) – care reprezintă raportul dintre numărul de cuvinte cu centre şi etichete corect idenificate şi numărul total de cuvinte din propoziţie – drept măsură a performanţei analizei sintactice, dar instrumentul MaltEval oferă şi posibilitatea calculării altori măsuri: LA(numărul de cuvinte cu etichete corecte raportat la numărul total de cuvinte din propoziţie), UAS (numărul de cuvinte cu centru corect identificat raportat la numărul total de cuvinte din propoziţie), AnyRight (numărul de cuvinte care au centrul, eticheta sau pe amândouă corect identificate raportat la numărul total de cuvinte din propoziţie) etc.

[10] http://ifarm.nl/signll/conll/

## 6.1. Prima etapă de evaluare.

Într-o etapă inițală, la începutul muncii de corecție, am efectuat evaluarea rezultatelor pe primele 100 de propoziții corectate (parte din sub-corpusul jurnalistic).

Tabelul 2. Evaluarea rezultatelor pentru primele 100 de propoziții corectate.

| Metrică | Scor |
|---|---|
| LAS | 0,216 |
| LA | 0,417 |
| UAS | 0,514 |
| AnyRight | 0,715 |

Scorurile din Tabelul 1 indică faptul că un număr mare de corecții manuale trebuie aplicate, dar o rată de eroare este inerentă procesului de adnotare statistic, cu atât mai mult cu cât folosim un model delexicalizat antrenat pe altă limbă. Din experiența de corectare, multe dintre erori se datorează faptului că am creat un set de etichete de dependențe mai rafinat decât setul IULA, iar altele sunt consecința principiilor de analiză diferite adoptate de noi (vezi secțiunea 4).

Totuşi, luând în considerare scorul AnyRight (0,715) şi experiența concretă de lucru la corectura manuală, am apreciat că analiza sintactică automată furnizată de modelul statistic pe limba spaniolă oferă o bază solidă, fiind de preferat muncii de construire de la zero a arborelui de dependențe echivalent fiecărei propoziții.

## 6.2. A doua etapă de evaluare

Potrivit experimentelor cu perechea de limbi catalană-spaniolă, după corectarea primelor 1000 de propoziții şi antrenarea unui model statistic lexicalizat pe acestea, s-a observat o creştere de aproximativ 7% a scorului LAS (Arias et al, 2014), de la 0,790 la 0,864 pentru propoziții cu lungimea de 18-19 cuvinte. Deoarece nu dispunem nici de suficient timp (crearea nucleului de treebank este restricționată la durata unui stagiu post-doctoral de 16 luni), nici de suficiente resurse umane, am decis să grăbim antrenarea unui model, după corectarea a doar 500 de propoziții. Aşa cum am menționat în Secțiunea 4, au fost antrenate două modele: optimizat (denumit în continuare *Ro-opt-500*) şi neoptimizat (denumit în continuare *Ro-neopt-500*). Cu aceste două modele şi cu vechiul model de limbă spaniolă au fost analizate sintactic automat 100 de propoziții noi, iar după corectura lor

manuală (care a fost luată drept corpus de referință) au fost evaluatele performanțele de adnotare în termenii scorului LAS. După cum se poate observa în tabelul de mai jos, creşterea scorului LAS este mai spectaculoasă decât în cazul experimentului catalano-spaniol, lucru explicabil prin faptul că modelul statistic spaniol (antrenat pe 30.000 de propoziții) obținea deja rezultate performante aplicat pe texte în limba catalană: un scor de 0,790, mult mai mare decât cel de aproximativ 0,260 obținut pe limba română. Pentru scopurile noastre, o creştere de 35% a scorului LAS după doar 500 de propoziții adnotate, reprezintă un salt foarte util, reflectat de altfel în uşurarea muncii de corectură manuală.

*Tabelul 3: Evaluarea modelelor statistice optimizat şi neoptimizat antrenate pe primele 500 de propoziții corectate manual.*

| **Modelul statistic** | **Scorul LAS** |
|---|---|
| Ro-neopt-500 | 0,469 |
| Ro-opt-500 | **0,547** |
| Spaniol | 0,202 |

## Concluzii

Considerăm că abordarea aleasă – de a adnota automat cu un model statistic pe o limbă cu sintaxă asemănătoare şi de a corecta manual această adnotare – este cea mai potrivită pentru a obține în timp scurt o resursă importantă şi de calitate. Rezultatele obținute până acum, evaluate prin metrica LAS, sunt încurajatoare şi credem că în scurt timp (după corectarea a 1000 de arbori sintactici) vom putea raporta o nouă creştere substanțială a acestui scor. În final, resursa noastră va număra 5000 de propoziții analizate sintactic.

Deoarece între timp a luat avânt o inițiativă de aliniere completă la setul UD a 3 treebankuri româneşti (RacaiRoTb, treebank-ul dezvoltat de Perez (2014) şi un treebank în curs de dezvoltare de către C. Mărănduc în cadrul unei teze de masterat), vom face transferul etichetelor din treebank-ul nostru la un set de etichete UD. Chiar dacă România nu este partener oficial în proiectul UD, intenționăm să promovăm şi să punem la dispoziția comunității de cercetare treebank-ul nostru, RACAI-RoTb, prin intermediul platformei UD.

## Mulțumiri

Această lucrare a fost realizată în cadrul proiectului "Cultura română şi modele culturale europene: cercetare, sincronizare, durabilitate", cofinanțat de Uniunea Europeană şi Guvernul României din Fondul Social European prin Programul Operațional Sectorial Dezvoltarea Resurselor Umane 2007-2013, contractul de finanțare nr.POSDRU/159/1.5/S/136077.

## Referințe

Abeille, A (ed.). *Treebanks. Building and Using Parsed Corpora.* Kluwer Academic Publishers, 2003.

Arias, B., Bel, N., Fomicheva, M., Larrea, I., Lorente, M., Marimon, M., Mila, A., Vivaldi, J. and Padro, M., 2014. Boosting the creation of a treebank, *Proceedings of LREC 2014*, Reykjavik, Iceland

Bick, E., Greavu, A., 2010. A Grammatically Annotated Corpus of Romanian Business Texts, *Proceedings of Multilinguality and Interoperability in Language Processing with Emphasis on Romanian*, Editura Academiei Romane, 169-183.

Brants, S., Dipper, S., Eisenberg, P., Hansen, S., König, E., Lezius, W., Rohrer, C., Smith, G. and Uszkoreit H. (2004).TIGER: Linguistic Interpretation of a German Corpus. *Journal of Language and Computation*, 2004 (2), 597-620.

Călăcean, M., Nivre, J., 2009. A Data-Driven Dependency Parser for Romanian, *Proc. Seventh International Workshop on Treebanks and Linguistic Theories*, 65-76.

Chiang, D.( 2007). Hierarchical Phrase-Based Translation. *Computational Linguistics*, MIT Press.

Colhon, M., Simionescu, R. (2012). Deriving a statistical syntactic parsing from a treebank. *Proceedings of the 2Nd International Conference on Web Intelligence, Mining and Semantics*, 8 pages.

Garside, R., Leech, G., Váradi, T., *Manual of Information for the Lancaster Parsed Corpus*. Lancaster University, 1992.

Filip, F. G., Leiviska, K. ( 2009). Large-scale complex systems. *Springer Handbook of Automation ( S. Y. Nof, Ed. )* , Springer , Dordrecht, p. 619-.638

Florea, I.M., Rebedea, T., Chiru, C.G. Parser de dependențe pentru limba română realizat pe baza parserelor pentru alte limbi romanice. *Revista Romana de Interactiune Om-Calculator 7(1)*, 1-20, 2014.

Hajič, J., Hajičová, E., Pajas, P., Panevová, J., Sgall, P., Vidová Hladká, B., 2001. Prague Dependency Treebank 1.0 (Final Production Label), *CD-ROM*, CAT: LDC2001T10, ISBN 1-58563-212-0, Linguistic Data Consortium.

Hristea, F., Popescu, M. (2003). A Dependency Grammar Approach to Syntactic Analysis with Special Reference to Romanian. F. Hristea şi M. Popescu (Ed.), *Building*

*Awareness in Language Technology*, Editura Universităţii din Bucureşti, 9-16.

Ion, R., Irimia, E., Ştefănescu, D. and Tufiş, D., 2012. ROMBAC: The Romanian Balanced Annotated Corpus, *Procedings of LREC 2012* Istanbul, Turkey.

Marcu D., Wong, W., 2002. A Phrased-Based, Joint Probability Model for Statistical Machine Translation, *Proceedings Of the Conference on Empirical Methods in Natural Language Processing (EMNLP 02),* Philadelphia, PA, July, 133-139.

Marimon M., Bel, N. Dependency structure annotation in the IULA Spanish LSP Treebank. *Language Resources and Evaluation*. Amsterdam: Springer Netherlands, 2014.

Mărănduc C. şi Perez. A.-C. 2015. A Romanian dependency treebank, *paper accepted at CICLing 2015*, Cairo, 14-20 April.

Nilsson, J., Nivre, J., 2008. MaltEval: An Evaluation and Visualization Tool for Dependency Parsing, *Proceedings of LREC 2008*, Marrakesch, Morocco.

Nivre, J., Hall, J. 2005. Maltparser: A language-independent system for data-driven dependency parsing, *Proceedings of the 4th Workshop on Treebanks and Linguistic Theories (TLT)*, 137-148.

Perez, A.-C. *Resurse lingvistice pentru prelucrarea limbajului natural*. PhD thesis, "Al. I Cuza" University, Iasi, 2014.

Popescu, M. (2003). Dependency Grammar Annotator. F. Hristea şi M. Popescu (coord.), *Building Awareness in Language Technology*, Bucureşti, Editura Universităţii din Bucureşti, 17-34.

Seretan, V., Wehrli, E., Nerima, L., Soare, G., 2010. FipsRomanian: Towards a Romanian Version of the Fips Syntactic Parser, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10),* Valletta, Malta

Skut, W., Krenn B., Brants Th., Uszkoreit, H., 1997. An Annotation Scheme for Free Word Order Languages, *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP-97)*. Washington, DC, USA

Taylor, A., Mitchell, M., Santorini, B., 2003. *The PENN Treebank: An Overview*, in Abeille (2003), 6-22.

Trandabăţ, D., Irimia, E., Barbu Mititelu, V., Cristea, D., Tufiş, D. *The Romanian Language in the Digital Age. Limba română în era digitală*. In White Papers Series (Rehm, Georg and Uszkoreit, Hans). Springer-Verlag, Berlin, Heidelberg, 2012.

Och, F.-J., Ch. Tillmann, Ney, H., 1999. Improved Alignment Models for Statistical Machine Translation. *Proceedings of the Joint Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora* (EMNLP/VLC 99), College Park, MD, June, 20–28.

Yamada K., Knight, K., 2002. A Decoder for Syntax-based Statistical MT. *Proceedings Of the 40th Annual Conf. of the Association for Computational Linguistics*, Philadelphia, PA, July, 303-310.

Zens R., Ney H., 2004. Improvements in phrase-based statistical machine translation. *Proceedings of HLT-NAACL*, 257-264.