Revista Română de Interacțiune Om-Calculator 7 (1) 2014, 1-20

Parser de dependențe pentru limba română realizat pe baza parserelor pentru alte limbi romanice

Iulia Maria Florea[1,2], Traian Rebedea[1,2], Costin-Gabriel Chiru[1]

[1]Universitatea Politehnica din Bucureşti, Facultatea de Automatică şi Calculatoare, Splaiul Independenței, Nr. 313, 060042 Bucureşti, România

[2]TeamNet International, Splaiul Independenței, Nr. 319, 060044 Bucureşti, România
E-mail: iulia.florea@cti.pub.ro, traian.rebedea@cs.pub.ro, costin.chiru@cs.pub.ro

Rezumat. Determinarea dependențelor sintactice între cuvintele dintr-o frază reprezintă o sarcină importantă în domeniul procesării limbajului natural, fiind utilă pentru o varietate de aplicații, printre care traducerea automată, extragerea şi clasificarea opiniilor din texte, aplicațiile de tip întrebare-răspuns şi altele. Lucrarea de față reprezintă un prim pas pentru crearea semi-automată a unui corpus adnotat cu dependențe sintactice pentru limba română, îmbogățit cu informații despre tipul cuvintelor şi a relațiilor dintre ele. În lipsa unui parser sintactic sau de dependențe creat (antrenat) special pentru limba română, acest corpus este necesar pentru a obține rezultate mai bune în aplicațiile lingvistice care au nevoie de arbori de dependențe. Pentru aceasta, am plecat de la două tipuri de parsere existente foarte cunoscute, primul antrenat pentru limba franceză şi al doilea pentru limba spaniolă, care au fost modificate pentru a analiza frazele în limba română. Rezultatele obținute prin această metodă sunt explicate şi comparate cu cele întoarse de către un analizor antrenat pentru limba română, pe un corpus de dimensiuni medii.

Cuvinte cheie: prelucrarea limbajului natural, parsarea de dependențe, adaptare parsere, extragere caracteristici, sintaxă

1. Introducere

Procesarea limbajului natural (PLN) oferă una dintre modalitățile de a face interacțiunea om-calculator (IOC) mai interesantă şi mai accesibilă. De exemplu, recunoaşterea scrisului de mână sau a vorbirii sunt integrate în diverse aplicații software folosite pe scară largă. Datorită evoluției tehnicilor de învățare automată şi a dezvoltării aplicațiilor din domeniul PLN, parsarea de dependențe a devenit o parte importantă a procesării limbajului, fiind esențială aplicațiilor de dimensiuni mai mari şi a celor care fac prelucrări mai complexe. În contextul IOC, dependențele sintactice sunt

utile pentru realizarea unor interfețe multi-modale mai complexe, precum a sistemelor de tip întrebare-răspuns sau a agenților conversaționali.

În PLN, prin parsare (eng. *parsing*) se înțelege, în general, obținerea unui arbore care prezintă relațiile dintre cuvintele unei fraze analizate. Există două tipuri de arbori de parsare: de parsare sintactică şi de dependențe, care pun accentul pe legăturile dintre cuvinte. Primul tip se referă la structura frazei şi poate fi obținut folosind gramatici independente de context (eventual probabilistice), în timp ce al doilea tip arată relațiile gramaticale, cum ar fi atributele sau complementele care pot fi regăsite într-o propoziție. Figurile 1.a şi 1.b, extrase din Marneffe şi Manning (2008), prezintă câte un exemplu din fiecare tip de arbore de parsare, pentru a accentua diferențele. Fraza analizată este în limba engleză: *„Bell, based in Los Angeles, makes and distributes electronic, computer and building products”*.

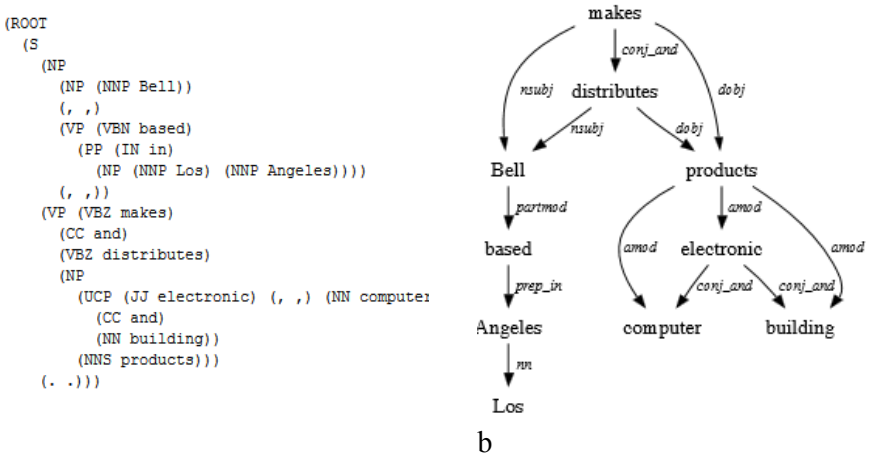

Figura 1. Exemple de arbori de parsare construiți pentru aceeaşi frază: a) arbore de parsare sintactic, b) arbore de dependențe (preluați din Marneffe şi Manning (2008)).

Relațiile dintre cuvinte sunt importante în domeniul prelucrării limbajului natural. De exemplu, aplicațiile de traducere automată care folosesc arbori de dependență, structura frazei şi algoritmi de învățare automată obțin rezultate mai bune decât metodele anterioare (Alshawi et al., 2000).

Un alt exemplu unde relațiile dintre cuvintele unei fraze pot ajuta la îmbunătățirea performanțelor obținute este în extracția informațiilor, în special a entităților numite (eng. *named entities*). De exemplu, o cercetare în domeniul extracției entităților numite din biologie (Fundel et al., 2007)

confirmă faptul că rezultatele au fost îmbunătățite ca urmare a folosirii parsării de dependențe.

Analiza frazei poate fi utilizată, de asemenea, în aplicații de determinare a polarității opiniilor. Părerile pozitive sau negative despre persoane, locuri, organizații sau alte entități numite pot fi, de asemenea, determinate cu ajutorul arborilor de dependențe (Boiy şi Moens, 2009). În plus, jocurile pe calculator pot fi îmbogățite cu aplicații de înțelegere a limbajului care folosesc dependențe (Gorniack, 2007).

Din păcate, pentru limba română nu există niciun corpus de dimensiuni suficient de mari, adnotat cu relațiile sintactice între cuvinte, care să poată fi folosit pentru a antrena parsere de dependențe sau sintactice sau pentru a dezvolta parsere alternative. O aplicație semi-automată este o modalitate de a dezvolta mai uşor un astfel de corpus. Întrucât performanţele unui parser de dependențe statistic cresc odată cu dimensiunea corpusului folosit pentru antrenare, este necesar ca acest corpus să conțină cât mai multe propoziții din categorii de texte cât mai diferite.

Lucrarea de față prezintă o aplicație concepută pentru crearea semi-automată a unui corpus adnotat pentru limba română, folosind parsere antrenate pentru alte limbi romanice, cum ar fi franceza sau spaniola. Două tipuri de parsere sunt analizate: un exemplu de parser sintactic, reprezentat de Stanford Parser, şi un parser de dependențe, MaltParser. De asemenea, este prezentată o modalitate de a compara diferențele obținute prin adaptarea a două tipuri de parsere antrenate pe limbi similare. Parserul construit pornind de la limba spaniolă foloseşte tipuri de legături între cuvinte specifice limbii spaniole, care sunt însă similare cu cele existente în limba română.

2. Alte aplicații similare

2.1 Extragerea caracteristicilor pentru parsare

Potrivit cercetărilor efectuate de Ryan McDonalds et al. (2005), există mai multe caracteristici esențiale care por fi utile pentru a descrie tipul unei dependențe. Fiecare dintre ele este văzută ca o relație părinte-copil şi poate fi descrisă folosind următoarele caracteristici principale şi combinații ale acestora:

a) caracteristici de bază (unigrame):

- cuvântul părinte
- parte de vorbire a părintelui;
- cuvântul copil;
- partea de vorbire a copilului.

b) caracteristici de nivel doi (bigrame):

- se pot folosi (complet sau parțial) următoarele informații pentru a construi caracteristici de tip bigramă: cuvântul părinte, partea de vorbire a părintelui, cuvântul copil, partea de vorbire a copilului.

c) caracteristicile părților de vorbire dintre părinte şi copil:

- partea de vorbire a părintelui, o parte de vorbire dintre ele, partea de vorbire a copilului.

d) părțile de vorbire ale cuvintelor apropiate:

- partea de vorbire a părintelui împreuna cu partea de vorbire a cuvântului premergător/următor părintelui;
- partea de vorbire a copilului, împreună cu partea de vorbire a cuvântului premergător/următor copilului.

e) caracteristici legate de tipul dependenței:

- tipul dependenței între părinte şi copil;
- direcția dependenței.

În plus, informații cu privire la direcția dependenței (stânga, dacă copilul se găseşte în partea stângă a părintelui în fraza originală, respectiv dreapta, altfel) şi distanța dintre cele două cuvinte sunt, de asemenea, informații utile. Pot fi folosite şi lematizoare (sau eliminarea sufixelor) pentru a elimina inflexiunile unui cuvânt (cum ar fi diferențele dintre timp, număr sau gen). Pentru fiecare dependență, toate caracteristicile de mai sus sunt reținute pentru antrenarea unui parser statistic de dependențe.

2.2 Corpusul adnotat cu dependențe în limba română

Pentru dezvoltarea aplicației prezentate în această lucrare, inițial s-a pornit de la un corpus de dimensiuni reduse, distribuit online, parte a proiectului RORIC-LING (http://www.phobos.ro/roric/), inclus în proiectul BALRIC-LING. Acesta a fost dezvoltat pentru limbile română şi bulgară, cu scopul

de a atrage atenția cu privire la aplicațiile ştiințifice şi industriale care pot fi dezvoltate în domeniul PLN. Acesta proiect oferă, printre altele, resurse lingvistice şi adnotări centrate pe cuvinte, corpusuri şi etichete la nivel de expresie şi frază, etc.

În cadrul acestui proiect se regăseşte şi un corpus adnotat cu dependențele între cuvinte. Astfel, propozițiile adnotate conțin toate cuvintele din frază, împreună cu indicele fiecărui cuvânt, cuvântul de care depinde şi tipul legăturii. Părțile de vorbire şi tipurile de dependențe sunt dependente de limba propoziției.

Părțile de vorbire din limba română

Corpusul conține doar informații de bază despre părțile de vorbire, fără a oferi detalii despre alte caracteristici ale lexemelor (de exemplu, gen, număr sau caz). Există nouă părți de vorbire utilizate pentru adnotare, care includ substantive, adverbe, prepoziții, cifre, etc. De asemenea, există unele părți de vorbire împărțite în subtipuri:

- pronumele sunt împărțite în două tipuri, cel reflexiv fiind separat;
- trei tipuri de verbe: unul auxiliar şi două tipuri principale;
- două tipuri de conjuncții: coordonatoare şi auxiliare;
- patru tipuri de articole: posesive, hotărâte, nehotărâte şi demonstrative;
- două tipuri de adjective: obținute din verbe la participiu şi altele.

Tipurile de dependențe pentru limba română

Următoarele tipuri de dependențe specifice limbii române sunt utilizate în acest corpus:

- **atribute**: în acest caz, cuvântul părinte este întotdeauna un substantiv. Copilul poate fi un substantiv, verb, adverb sau adjectiv.
- **subiect**: în general, subiectul poate fi un substantiv sau orice altă parte de vorbire cu rol de subiect (de ex., un pronume sau numeral).
- **complement**: în acest caz, mai multe tipuri de obiecte pot fi găsite în limba română:
  - complement direct;
  - complement indirect;

  - complemente circumstanțiale de timp, loc sau mod;
  - complement circumstanțial de agent, de obicei, acesta este un substantiv sau un înlocuitor (pronume sau numeral).
- **nume predicative**: sunt legate de verbe de stare (de ex., „a fi") şi părțile de vorbire care pot fi nume predicative sunt, în general, adjective, substantive sau orice înlocuitor.
- **alte relații**: demonstrativă (copilul este un pronume demonstrativ), reflexivă, prepozițională, conjuncțională, hotărâtă, comparativă, nehotărâtă, negativă, posesivă şi auxiliară.

2.3 Adnotarea părților de vorbire în limba română

În prima fază a procesului de adnotare a dependențelor, frazele au fost etichetate cu informații despre părțile de vorbire folosind un serviciu web dezvoltat pentru limba română. Acesta a fost realizat de către Institutul de Cercetare pentru Inteligenta Artificială (ICIA) „Mihai Drăgănescu" şi se bazează pe Qtag (Mason, 1998). Astfel, resursele lexicale s-au obținut folosind două corpusuri paralele (unul englez-român şi altul francez-român), împreună cu fraze adnotate automat.

Algoritmul de etichetare este pur probabilistic. Cuvântul actual este citit şi apoi este căutat într-un dicționar. Dacă nu a fost găsit, posibila parte de vorbire este „ghicită". În caz contrar, probabilitatea pentru fiecare etichetă posibilă este determinată folosind părțile de vorbire găsite în dicționar, împreună cu probabilitatea ca o anumită etichetă să urmeze altor două etichete (trigrame de părți de vorbire). Partea de vorbire finală este determinată după variația probabilităților contextuale.

Ieşirea constă din două seturi de etichete, primul cuprinzând principalele caracteristici ale unui cuvânt, iar celălalt conținând atributele aplicabile în funcție de tipul morfologic al cuvântului. O listă de caracteristici selectate pentru fiecare parte de vorbire se găseşte în Tufiş (1998).

Figura 2 prezintă etichetele determinate pentru următoarea propoziție: *„Zilele următoare, vremea se va menține la temperaturi în limitele specifice perioadei"*. Informațiile obținute în urma etichetării sunt tupluri separate prin caracterul '|':

- prima parte este cuvântul real;
- a doua este forma de bază, fără inflexiuni de gen, număr sau timp;

- următoarea informație este partea de vorbire principală;
- ultima parte oferă informațiile de flexionare.

```
Zilele|zi|NPRY|Ncfpry următoare|următor|AN|Afpf--n ,|,|COMMA|COMMA vremea|vreme|NSRY|Ncfsry
se|sine|PXA|Px3--a--------w va|vrea|VA3S|Va--3s menține|menține|VN|Vmnp la|la|S|Spsa
temperaturi|temperatură|NPN|Ncfp-n în|în|S|Spsa limitele|limită|NPRY|Ncfpry specifice|specific|APN|Afpfp-n
perioadei|perioadă|NSOY|Ncfsoy .|.|PERIOD|PERIOD
```

Figura 2. Propoziție în limba română etichetată folosind *POS tagger*-ul de la ICIA

3. Implementarea soluțiilor propuse

3.1 Reguli de dependențe şi euristici

Pentru parserul construit pe baza celui pentru limba franceză, am definit un set de reguli pentru a îmbunătăți precizia de etichetare şi pentru accelerarea procesului de analiză. Pentru fiecare parte de vorbire, în loc de a încerca să aplicăm orice etichetă pentru o dependență, am limitat mulțimea de tipuri posibile. Astfel, am adăugat următoarele reguli, în funcție de partea de vorbire a copilului:

- în cazul în care acesta este o conjuncție, dependența este relație conjunctivă;
- dacă acesta este o prepoziție, dependența este relație prepozițională;
- dacă este un articol nehotărât, dependența este o relație nehotărâtă;
- dacă acesta este un verb, atunci este rădăcină a arborelui de dependențe sau este legat de rădăcina arborelui prin conjuncții, în cazul frazelor;
- un substantiv poate fi doar un atribut, complement, subiect sau nume predicativ;
- un adjectiv poate fi doar un atribut şi, în general, depinde de un substantiv din apropiere;
- un adverb poate fi un atribut sau un complement şi de cele mai multe ori depinde de un verb;
- o conjuncție introduce de cele mai multe ori o nouă propoziție, iar verbele din propozițiile coordonate sau subordonate ar trebui să depindă de ea.

Există, de asemenea, reguli legate de partea de vorbire a părintelui:

- dacă acesta este un substantiv, atunci relația dintre cuvinte este atribut, tipul acesteia depinzând de partea de vorbire a copilului. Atributele în limba română pot fi alte substantive (în cazul direct considerat apoziție), adjective, verbe sau adverbe.
- dacă este un verb, dependența este un complement. Nu există nici o regulă exactă şi tipul acestuia depinde de sensul cuvântului copil.
- dacă este un pronume demonstrativ, atunci aceasta este o relație demonstrativă;
- dacă este un pronume reflexiv, atunci este o relație reflexivă;
- dacă este un cuvânt auxiliar pentru adjective comparative, atunci este o relație comparativă;
- dacă este un cuvânt auxiliar negativ, atunci este o relație negativă.

De exemplu, în cazul frazei din Figura 2 (*„Zilele următoare, vremea se va menține la temperaturi în limitele specifice perioadei”*) se vor obține următoarele relații:

- relația *„la temperaturi”* este prepozițională;
- adjectivul *„următoare”* este atribut care determină substantivul *„zilele”*;
- substantivele *„zilele”*, *„vremea”*, *„temperaturi”* şi *„perioadei”* pot fi atribute, complemente, subiecte sau nume predicative, iar partea de propoziție va fi determinată în funcție de caz, formă (articulată sau nearticulată) şi cuvântul pe care îl determină.
- grupul verbal *„se va menține”* va fi la rădăcina arborelui de dependențe rezultat.

3.2 Adaptarea Stanford Parser pentru limba română

Parserul sintactic de la Stanford (http://nlp.stanford.edu/software/lex-parser.shtml) a fost antrenat şi pe texte în limba franceză, însă poate determina doar un arbore de parsare sintactic pentru această limbă, pe baza gramaticilor probabilistice independente de context. În plus, acesta recunoaşte doar partea de vorbire principală, fără caracteristici suplimentare. O parte din clasele de vorbire recunoscute de parser sunt: N (substantiv), A (adjectiv), V (verb), ADV (adverb), P (prepoziție), D (determinant), C

(conjuncție), I (interjecție), CL, PRO (diferite tipuri de pronume), PUNCT (punctuație) și ET (cuvânt străin).

Transformarea părților de vorbire dintr-o frază în limba română către părți de vorbire din limba franceză include nu doar modificări de bază (pentru o parte de vorbire din limba română există un echivalent în limba franceză recunoscut de parser), dar și transformarea într-o altă parte de vorbire (de ex., nu există un echivalent pentru numeralul din română în părțile de vorbire utilizate de Stanford Parser pentru franceză) și adăugarea de cuvinte necesare, cum ar fi articole hotărâte sau nehotărâte.

Algoritmul este dezvoltat folosind patru faze:

- transformarea părților de vorbire din română în franceză;
- crearea arborilor de parsare în limba franceză;
- obținerea dependențelor dintre cuvinte pentru limba franceză;
- etichetarea dependențelor astfel obținute.

În prima fază, părțile de vorbire din limba română trebuie modificate pentru a fi recunoscute de către parserul sintactic francez. La început, prepozițiile și determinanții sunt ignorați. Există trei tipuri de cuvinte: cele care trebuie să fie precedate de un articol (substantivele hotărâte, adjectivele și pronumele), cele care pot fi precedate de un articol sau o prepoziție și cele care nu au nevoie de nimic în fața lor.

Pentru fiecare cuvânt, în funcție de informațiile furnizate de către POS tagger-ul românesc, putem afla dacă poate exista o prepoziție sau un articol în fața lui. În cazul în care acesta poate exista, considerăm că determinanții pot fi chiar în fața cuvântului în fraza originală sau, în cazul unui substantiv, orice articol poate fi găsit în fața adjectivelor care îl preced. O altă regulă de gramatică pe care am observat-o este că prepozițiile au mai multe șanse să fie găsite în fața articolelor. Un exemplu relevant al acestei abordări ar fi determinarea prepozițiilor. În cazul în care cuvântul curent este un substantiv care poate avea o prepoziție în fața lui, se presupune că orice prepoziție ar putea fi doar în fața substantivului, a unui determinant sau a unui adjectiv care poate preceda cuvântul curent. Deci, când vom ajunge la un alt cuvânt, cum ar fi un verb sau adverb, ne oprim din căutarea de prepoziții. Astfel, vom analiza fiecare parte de vorbire din fața substantivului curent și dacă este un adjectiv sau un determinant, trecem peste el la cuvântul care îl precede. Continuăm să căutăm prepoziții până

ajungem la un alt cuvânt important. Dacă vom găsi mai mult de o prepoziție, le păstrăm pe toate într-o stivă, pe care apoi o scriem de la ultimul şi primul cuvânt inserat.

De asemenea, se iau în considerare diferențele lingvistice, cum ar fi poziția articolelor hotărâte şi importanța articolelor în limba franceză (orice substantiv corect într-o frază trebuie să fie precedat de un articol). În cazul în care articolul este necesar, dar nu există în fraza originală din limba română, din cauza diferențelor de limbă, un marcaj suplimentar va fi adăugat în fața cuvântului.

Apoi, se trece la faza de parsare sintactică de către parserul francez. Acesta a fost antrenat pe un corpus francez şi arborii de parsare sunt obținuți folosind gramatici probabilistice independente de context. Doar cuvintele din fraza originală sunt păstrate şi orice semn suplimentar, adăugat în etapa anterioară, va fi eliminat după ce este obținut arborele de parsare. Figura 3 prezintă un exemplu de arbore de parsare, obținut folosind Stanford Parser pentru limba franceză pentru propoziția aflată în partea de sus a imaginii.

```
“Pe Drumul National 12 A, care leagă Miercurea Ciuc de Comănești, în dreptul Pasului Frumoasa, mai multe tiruri au rămas în pantă.“
(ROOT
 (SENT
  (PP (P Pe)
   (NP (N Drumul)
    (AP (A National))
    (Srel
     (NP (PRO 12))
     (VN (V A) (PUNCT ,))
     (NP (PRO care)))
    (VPpart
     (VN (V leagă))
     (NP
      (MWN (N Miercurea_Ciuc) (P de) (N Comănești) (PUNCT ,)))
     (PP (P în_dreptul)
      (NP (N Pasului) (N Frumoasa))))))
  (PUNCT ,)
  (NP (ADV mai) (D multe) (N tiruri))
  (VN (V au) (V rămas))
  (PP (P în)
   (NP (N pantă)))
  (PUNCT .)))
```

Figura 3. Exemplu de arbore de parsare obținut folosind Stanford Parser pentru limba franceză

Pentru a transforma arborii de parsare în arbori de dependențe, am modificat unele reguli şi euristici, folosite pentru analiză în limba spaniolă, prezentate de Gelbukh et al. (2007):

- dacă relația conține doar un element, atunci acela este rădăcina;
- dacă relația conține conjuncții coordonate, prima este rădăcina; orice tip de coordonare introduce o nouă propoziție principală;
- dacă relația conține un pronume relativ, atunci acesta este rădăcina, iar pronumele relativ introduce o propoziție auxiliară;
- dacă relația/modelul conține un verb auxiliar, urmat de un verb la participiu, atunci verbul la participiu este rădăcina. Pot exista cel mult două cuvinte între cele două verbe. Acest lucru se întâmplă în cazul timpurilor compuse, cum ar fi viitorul în limba română.
- dacă modelul conține un verb la infinitiv, atunci acesta este rădăcina;
- dacă modelul conține un verb principal, atunci acesta este rădăcina - verbul este întotdeauna ales rădăcină a frazei, aşa că trebuie luat în considerare ca fiind cel mai important cuvânt din partea analizată;
- dacă modelul conține un verb auxiliar şi orice alt verb, atunci verbul auxiliar nu este niciodată rădăcina, acest lucru este valabil în special în cazul timpurilor compuse, atunci când verbul la participiu este de preferat să fie ales în calitate de parte importantă;
- dacă primul element este un articol, atunci acesta nu este rădăcină; articolele determină întotdeauna cuvântul principal cel mai apropiat;
- în cazul grupurilor nominale, dacă modelul conține un substantiv, atunci acesta este rădăcina - într-un grup nominal există întotdeauna un substantiv şi cel puțin un determinant, cum ar fi un adjectiv;
- în cazul în care rădăcina nu a fost găsită deja, vom alege un substantiv, un adjectiv sau un adverb ca parte importantă a structurii, în funcție de caz, forma articulată sau nearticulată (rădăcina va fi mai degrabă un substantiv articulat) sau, dacă acesta este un adverb, primul care apare în structură.

În ceea ce priveşte regulile de mai sus, ordinea de evaluare este foarte importantă, pentru că atunci când o regulă este potrivită pentru o parte din propoziție, este adăugată o dependență. Am considerat regulile cele mai relevante ca fiind cele legate de verbe auxiliare şi principale deoarece

acestea sunt, de asemenea, cele mai importante cuvinte din propoziție (verbele sunt mai aproape de rădăcina arborelui). Apoi, regulile care implică substantive şi locuțiuni substantivale sunt selectate, deoarece acestea sunt aproape de rădăcină în structura arborescentă, având mai multe funcții sintactice, cum ar fi diferitele tipuri de subiect sau diferite complemente.

De exemplu, pentru propoziția din Figura 3 (*„Pe drumul național 12 A, care leagă Miercurea Ciuc de Comănești, în dreptul Pasului Frumoasa, mai multe tiruri au rămas în pantă."*), se vor obține următoarele relații:

- în grupul de cuvinte *„Pe drumul național"* se va alege drept rădăcină cuvântul *„drumul"*. Prepoziția *„pe"* şi substantivul *„național"* vor determina rădăcina.
- în grupul nominal *„mai multe tiruri"*, conform regulilor de mai sus, rădăcina va fi substantivul *"tiruri"*, adjectivul *„multe"* îl va determina şi va avea funcția de atribut, iar adverbul *„mai"* va fi legat de adjectiv.
- în grupul verbal *"au rămas"*, se va aplica regula specific verbelor auxiliare şi va rezulta că rădăcina construcției este verbul la participiu *„rămas"*.

Folosind aceste reguli, am modificat arborele de parsare original, creând un altul, în care cuvintele mai importante pe un nivel superior în structura arborescentă. Algoritmul porneşte de la frunze şi avansează până când se ajunge la rădăcină şi se determină un cuvânt de care depind toate celelalte din frază. Pentru găsirea de dependențe, am folosit acelaşi algoritm ca cei de la Stanford. Deoarece gramatica limbii române nu are reguli stricte în ceea ce priveşte ordinea cuvintelor, atât dependențe proiective cât şi non-proiective sunt posibile. De Marneffe şi Manning (2008) prezintă mai multe abordări pentru algoritmul de determinare de dependențe, dar noi am considerat că cea mai bună opțiune este pentru cazul non-proiectiv. Un arbore de parsare proiectiv implică faptul că muchiile reprezentând dependențele nu se intersectează dacă se păstrează ordinea cuvintelor. Acest lucru este, în general, adevărat pentru limba engleză, dar nu se poate spune acelaşi lucru pentru limbile romanice.

Ideea algoritmului este de a menține o listă de cuvinte care au fost întâlnite până la momentul curent şi încă o listă pentru cuvintele care nu au încă un părinte. Pentru cuvântul curent, trebuie să se verifice dacă există un cuvânt în lista de cuvinte fără părinte care poate depinde de el. Dacă nu există, algoritmul caută în lista de cuvinte pentru a afla dacă există în ea un

posibil părinte pentru cuvântul curent. Dacă este aşa, este creată o nouă legătură. Altfel, cuvântul curent va fi adăugat la lista celor fără părinte şi nici dependență nu va fi creată în această etapă.

Dacă graful de dependențe obținut respectă regulile generale ale unui arbore, la final va fi doar un cuvânt în lista celor fără părinte: rădăcina. Celelalte dependențe trebuie să fie conectate la rădăcină, într-un mod direct sau indirect. Pentru a avea un arbore format corect, nu trebuie să existe niciun nod separat. Căutarea prin toate cuvintele din listă înseamnă că toate dependențele pot fi traversate şi dependențe non-proiective pot fi obținute.

După ce se obțin dependențele, acestea trebuie să fie etichetate. În primul rând, există câteva reguli care pot fi urmate pentru obținerea tipurilor corecte de dependențe:

- dacă cuvântul principal este un substantiv şi copilul este substantiv, adjectiv, adverb sau un verb, eticheta de dependență poate fi doar de atribut şi tipul acesteia depinde de partea de vorbire a copilului;
- dacă rădăcina este un verb, iar copilul este un substantiv în cazul nominativ, atunci copilul este subiect sau complement direct;
- în caz contrar, pentru fiecare dependență, vom crea un set de caracteristici, cum s-a menționat în secțiunea 3.1.

În primul rând, partea de vorbire a cuvintelor analizate trebuie să fie luată în considerare. Pentru fiecare dintre ele, există o mulțime de etichete permise (de ex., prepozițiile pot fi găsite doar în relații auxiliare sau prepoziționale). Aceste etichete posibile sunt extrase din propozițiile adnotate. Atât copilul cât şi părintele din noua dependență trebuie să aibă aceeaşi parte de vorbire cu cele din dependențele analizate anterior. Noul set este comparat cu fiecare set de dependențe din corpusul adnotat pentru a o găsi pe cea mai asemănătoare (cel care are cel mai mare număr de trăsături comune) şi tipul aceleia va deveni şi tipul noii dependențe. Figura 4 prezintă un exemplu de dependențe obținute după rularea metodei propuse în această secțiune.

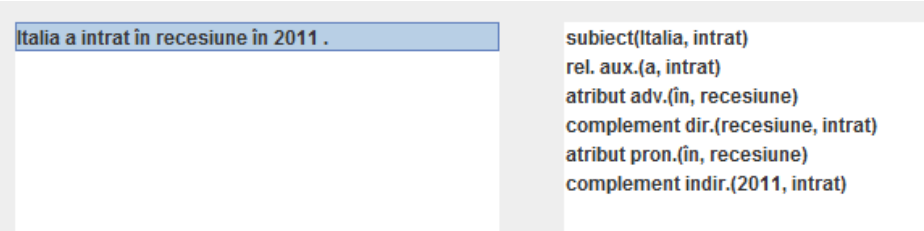

Figura 4. Exemplu de dependențe adnotate folosind Stanford Parser pentru limba franceză

3.3 Adaptare MaltParser pentru limba română

Pentru acest experiment, MaltParser a fost configurat pentru a utiliza algoritmul bazat pe arce (Nivre et al., 2006) şi clasificatoare liniare din pachetul LIBLINEAR (Fanetal, 2008) pentru a anticipa următoarele tranziții. Acesta foloseşte un corpus adnotat pentru limba spaniolă, antrenat pe articole din ziare.

Părțile de vorbire din limba română sunt modificate pentru a se potrivi cu cele din spaniolă, recunoscute de MaltParser. Fiecare parte de vorbire din limba română a fost legată de un corespondent din spaniolă sau adaptată pentru una similară, dacă aceasta nu avea niciun corespondent. Toate informațiile necesare au fost luate de la POS tagger-ul pentru limba română.

În ceea ce priveşte modificările pentru limba spaniolă, există caracteristici suplimentare, care conțin detalii auxiliare. Cea mai mare parte din aceste informații sunt furnizate de către POS tagger-ul pentru română, dar au existat, de asemenea, mai multe modificări care au fost făcute pentru ca parserul să poată rula pe fraza dată la intrare, în afară de redenumirea informațiilor despre cuvinte.

Diferențele de sintaxă între limbile spaniolă şi română care au fost luate în considerare în cadrul procesului de translatare a părților de vorbire din română în spaniolă sunt următoarele:

- verbele semi-auxiliare spaniole sunt mapate în verbele copulative şi auxiliare din română;
- nu există informații furnizate de POS tagger-ul pentru limba română cu privire la clasificarea numelor proprii (aceasta se referă la substantive proprii care pot fi clasificate ca persoană, organizație, locuri şi altele);

- POS tagger-ul din limba română nu face diferența dintre pronumele interogative şi relative. Pe acestea le-am separat în funcție de poziția cuvântului în frază. Dacă indicele cuvântul în frază este mai mic de trei, este mai probabil să fie pronume interogativ. În caz contrar, dacă cuvântul este la mijlocul frazei, acesta este un pronume relativ;
- dacă sunt necesare informații suplimentare cerute de MaltParser care nu sunt furnizate de tagger-ul din limba română, atunci aceste informații au fost omise. Este posibil să se omită unele atribute şi atunci se va transmite valoarea '0' în locul lor;
- toate tipurile de timpuri trecute, existente în limba română, sunt mapate în timpurile trecute existente în spaniolă;
- pronumele hotărâte sunt mapate în pronume nehotărâte;
- numai anumite caracteristici sunt disponibile pentru toate tipurile de pronume, iar cele mai multe dintre ele sunt necesare pentru pronumele personale;
- unele tipuri de determinanți, care nu sunt recunoscuți de parserul spaniol, sunt mapate în articole;
- clasificarea semantică a substantivelor proprii a fost ignorată aşa cum a fost menționat anterior;
- gradul substantivului a fost, de asemenea, ignorat ca urmare a lipsei de informații furnizate de POS tagger;
- genul neutru român este mapat în genul comun spaniol;
- modul infinitiv al verbelor din limba română este echivalent cu participiul din limba spaniolă;
- numărul pronumelui invariabil este ignorat.

Tipurile de dependențe din spaniolă sunt următoarele: subiect, modificator (ține locul atributelor si a diferitelor tipuri de complemente), complement direct şi indirect, negare, conjuncții, precum şi alte dependențe inexistente în limba română (Gelbukh et al., 2005). Figura 5 prezintă un exemplu de arbore de dependențe obținut folosind metoda propusă.

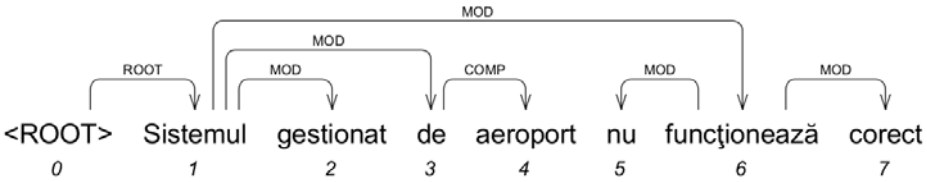

Figura 5. Arbore de dependențe obținut folosind modelul spaniol din cadrul MaltParser

4. Rezultatele experimentelor

Un prim set de experimente a fost realizat pe 10% din frazele din corpusul românesc adnotat prezentat în cadrul secțiunii 2.2. Am comparat rezultatele obținute cu cele ale unui parser românesc, dezvoltat la Universitatea "Alexandru Ioan Cuza" din Iaşi, accesibile prin intermediul unui serviciu web (şi disponibil online la adresa http://nlptools.infoiasi.ro/WebFdgRo/, dar despre care nu am găsit nici un articol publicat). Aceste rezultate sunt prezentate în Tabelul 1. Se pot observa rezultate mai bune în cazul folosirii parserului de la Stanford, datorită caracteristicilor extrase din corpusul adnotat. Pe de altă parte, parserul spaniol nu este influențat de propozițiile de test. Rezultatele în acest caz depind numai de asemănările dintre sintaxa frazei din spaniolă şi română. Totuşi, se observă faptul că parserul de dependențe antrenat special pentru limba română are rezultate mult mai bune decât ambele variante prezentate în secțiunile anterioare.

Tabelul 1. Rezultatele experimentelor făcute pe corpusul adnotat

Parser	Acurateţe
UAIC Parser	85%
Stanford Parser	73%
Malt Parser	62%

Cele mai multe dependențe recunoscute în mod corect de către ambele parsere descrise în cadrul lucrării sunt atributele şi cuvintele subordonate. Pe de altă parte, conjuncțiile, pronumele relative sau verbele subordonate pot depinde de verbul principal sau de alte cuvinte şi acest lucru este specific algoritmului de parsare. De exemplu, MaltParser este antrenat să asocieze conjuncțiile subordonatoare verbelor principale şi verbele subordonate sunt legate de aceste conjuncții, în timp ce parserul UAIC arată

dependențe între verbul subordonat şi principal, iar cuvântul de legătură este atașat la verbul subordonat.

În continuare am încercat să determinăm performanțele fiecărui parser de dependențe în funcție de tipul frazei analizate. Astfel, în Tabelul 2 sunt prezentate rezultatele obținute de cele trei parsere (UAIC şi cele două adaptate pentru limba română), atât pentru fraze simple, cât şi pentru fraze compuse. Se poate observa că rezultatele adaptării Stanford Parser din franceză în română se degradează substanțial, pe când cele ale parserului construit de către UAIC, precum şi ale Malt Parser sunt similare pentru fraze compuse şi simple. Explicația constă în faptul că Stanford Parser este de fapt un analizor sintactic pentru limba franceză, iar regulile introduse nu funcționează eficient în cadrul frazelor compuse din limba română.

Tabelul 2. Dependențe etichetate corect în funcție de tipul frazei

Parser	Fraze simple	Fraze compuse
UAIC Parser	77%	75%
Stanford Parser	67%	51%
Malt Parser	57%	59%

Următorul set de teste a fost rulat pentru a analiza, pentru cele mai importante părți de vorbire, cât de bine sunt legate de alte cuvinte. Datorită tipurilor de dependențe şi a regulilor introduse, specifice limbii române, parserul francez are o precizie mai mare pentru etichetarea dependențelor în cazul general. Prepozițiile, conjuncțiile şi unele tipuri de pronume sunt mai uşor de etichetat, deoarece există un singur tip de legătură permis pentru ele, dar în cazul propozițiilor subordonate, ele pot fi asociate unui verb greşit. Pe de altă parte, substantivele sunt mai greu de etichetat, datorită multiplelor funcții sintactice care acestea le pot avea. Tabelul 3 prezintă precizia asociată fiecărei părți de vorbire.

În cazul adjectivelor, testele au fost efectuate pe fraze scurte, simple, care conțin cel puțin un adjectiv. Ele sunt uşor de asociat cuvântului corect, un substantiv sau un verb de stare. Datorită euristicilor folosite, parserul francez asociază, de asemenea, tipul corect de dependență, în cele mai multe cazuri. Parserul bazat pe spaniolă consideră că adjectivul este un modificator, aceasta fiind abordarea corectă, dar mai generală şi, de asemenea, echivalentă cu alte dependențe din română.

Tabelul 3. Rezultatele parserelor adaptate în funcție de părțile de vorbire implicate în dependențe

Părți de vorbire	Tipuri de dependențe (%)	
	Stanford Parser - franceză	Malt Parser - spaniolă
Adjective	90.1	88.2
Adverbe	87.5	84.3
Conjuncții	55.3	60.7
Substantive	58.2	56.8
Verbe predicative	88.5	70.5
Prepoziții	64.1	65.2
Pronume	73.1	72.3

Verbele predicative sunt, de asemenea, uşor de determinat. Ele sunt, cel mai adesea, rădăcina frazei, în special în cazul unor propoziții simple. În alte expresii în care există mai multe verbe, se pot face greşeli atunci când se încearcă găsirea verbului principal şi asocierea lui la verbele subordonate.

Conjuncțiile sunt asociate verbelor şi, în cazul propozițiilor coordonate, ele pot fi asociate cu orice verb. Pe de altă parte, în cazul introducerii expresiilor subordonate, fiecare conjuncție trebuie să fie legată la verbul principal. Aceasta este o problemă pentru parsere, pentru că acestea nu au fost antrenate în acest context.

Substantivele sunt mai greu de asociat oricărei părți de vorbire, deoarece acestea pot avea roluri diferite în propoziții. Ele pot fi subiecte, complemente sau atribute, şi nu există nici o regulă pentru a le asocia mai uşor cu alte părți de vorbire.

5. Concluzii

Lucrarea de față prezintă o aplicație ce poate fi utilizată pentru dezvoltarea semi-automată a unui corpus adnotat cu dependențe sintactice pentru limba română. Pentru corectarea adnotărilor greşite introduse de către adnotarea automată propusă, prin folosirea parserelor dezvoltate pentru limbile romanice înrudite, este necesară folosirea unor aplicații de adnotare manuală de către lingvişti. O astfel de aplicație pe care autorii au folosit-o cu succes este Brat (http://brat.nlplab.org/), care dispune de o interfață grafică web în care pot fi încărcate fişiere pre-adnotate care apoi sunt modificate.

Pentru aceasta, în cadrul lucrării am analizat rezultatele obținute în urma adaptării parserelor sintactice deja existente, antrenate pe două limbi romanice: spaniola şi franceza. Pentru a obține o adnotare de dependențe în limba română, în primul rând am modificat informațiile despre părțile de vorbire din limba română în limba destinație şi am folosit două tipuri de parsere: unul de structură a frazei şi altul bazat pe dependențe. Rezultatele obținute au fost comparate folosind mai multe criterii.

Arborii de parsare sunt mai puțin influențați de limbaj, mai ales în cazul unei limbi cu puține reguli în ceea ce priveşte ordinea cuvintelor. Deoarece franceza şi româna sunt destul de asemănătoare, o mare parte din regulile de sintaxă sunt aplicabile ambelor limbi. Acest lucru înseamnă că o gramatică independentă de context obținută din corpusul de antrenare din limba franceză este parțial corectă pentru limba română. Mai mult, euristicile care modifică arborii de parsare sunt, de asemenea, adaptați la sintaxa limbii. Acest lucru face parsarea mai relevantă pentru limba română. Regulile şi euristicile îmbunătățesc, de asemenea, procesul de analiză, asociind fiecărei părți de vorbire un set mic de posibile dependențe. Experimentele au dovedit că se obțin rezultate mai bune în cazul frazelor scurte şi părțile de vorbire clasificate cel mai bine sunt adjectivele.

Pe de altă parte, parserele bazate pe fraze adnotate sunt mai dependente de limbă. Există un model care este învățat pornind de la caracteristici care conțin cuvinte întregi şi părțile de vorbire corespunzătoare. De asemenea, nu există nici o posibilă influență externă a rezultatelor. Acestea se bazează numai pe similitudinile lingvistice, furnizarea de informații exacte cu privire la caracteristicile auxiliare ale părților de vorbire şi adaptarea unora la categoria cea mai asemănătoare din cea de-a doua limbă.

De asemenea, rezultatele sunt diferite în funcție de părțile de vorbire. Adjectivele, de exemplu, sunt cel mai adesea asociate substantivului corespunzător, mai ales în cazul unor propoziții simple. Pe de altă parte, substantivele pot avea o varietate de roluri în frază şi acestea sunt mai dificil de etichetat corect.

În concluzie, performanțele obținute prin adaptarea parserelor existente pentru alte limbi romanice, precum franceza şi spaniola, sunt mai slabe decât rezultatele singurului parser public disponibil pentru limba română în acest moment. Pe de altă parte, folosirea acestora poate fi o soluție pentru crearea semi-automată (prin corectarea rezultatelor întoarse de către lingvişti) a unui corpus de mari dimensiuni adnotat cu dependențe pentru

limba română. Serviciul web de parsare pus la dispoziție de către UAIC nu poate fi folosit pentru adnotarea unor volume mari de texte, aşa cum ar fi necesar pentru o aplicație comercială. În plus, în momentul de față nu există un corpus liber adnotat cu dependențe sintactice pentru limba română, însă acesta ar putea fi creat prin metoda prezentată în acest articol.

Bibliografie

Alshawi, H., Douglas, S., & Bangalore, S. Learning dependency translation models as collections of finite-state head transducers. *Comput. Linguist., 26(1)*, pp. 45-60, 2000.

Boiy, E., & Moens, M.-F. A machine learning approach to sentiment analysis in multilingual Web texts. *Information Retrieval, 12(5)*, pp. 526-558, 2009.

Fundel, K., Küffner, R., Zimmer, R., & Miyano, S. RelEx–Relation extraction using dependency parse trees. *Bioinformatics, 23(3)*, pp. 365-371, 2007.

Gelbukh, A., Torres, S., & Calvo, H. (2005). Transforming a Constituency Treebank into a Dependency Treebank. *Procesamiento del Lenguaje Natural 35*, pp. 145-152, 2005.

Gorniak, P, & Roy, D. Probabilistic Grounding of Situated Speech using Plan Recognition and Reference Resolution. *Proceedings of the International Conference on Multimodal Interfaces (ICMI 2005)*, pp. 138-143, 2005.

de Marneffe, M.C., & Manning, C.D. *Stanford typed dependencies manual*, 2008.

Mason, O. *QTag – A Portable Probabilistic Tagger*. Available online at http://www-clg.bham.ac.uk/QTAG, 1997.

McDonald, R., Pereira, F., Ribarov, K., & Hajic, J. Non-projective Dependency Parsing using Spanning Tree Algorithms. *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pp. 523-530, 2005.

Nivre, J., Hall, J., & Nilsson, J. MaltParser: A Data-Driven Parser-Generator for Dependency Parsing. *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC2006)*, pp. 2216-2219, 2006.

Tufiş, D. Tagging Romanian Texts: a Case Study for QTAG, a Language Independent Probabilistic Tagger – Romanian POS Tagger. *Proceedings of the First International Conference on Language ressources and Evaluation (LREC1998)*, 1998.