Nucleu de bancă de arbori sintactici pentru limba română

Autor: **Elena I. IRIMIA**

Lucrare realizată în cadrul proiectului *"Cultura română şi modele culturale europene - cercetare, sincronizare, durabilitate"*, cofinanțat din FONDUL SOCIAL EUROPEAN prin Programul Operațional Sectorial pentru Dezvoltarea Resurselor Umane 2007 – 2013, Contract nr. POSDRU/159/1.5/S/136077.

* * *

ISBN 978-973-167-322-6

CUPRINS

TABLE OF CONTENTS

REZUMAT

Într-o epocă în care tehnologia informaţiei digitale devine din ce în ce mai complex interconectată cu toate aspectele vieţii umane, limbajul natural, în calitatea sa fundamentală de transmiţător de informaţie, este menit digitalizării. Pentru a supravieţui în societatea informaţională a viitorului, pentru ca vorbitorii săi nativi să se poată bucura neîngrădit de avantajele progresului tehnologic în viața publică și privată, la standardele la care au acces alți cetățeni europeni, limba română are nevoie de resurse și instrumente electronice dedicate. Acest suport tehnologic îi poate asigura integrabilitatea în complexele aplicații inteligente, mobile și web, care au devenit indispensabile.

Proiectul descris în lucrarea de faţă este doar un pas dintr-o strategie amplă de integrare a limbii române în spațiul digital european. Limba română are un dramatic deficit tehnologic de recuperat în raport cu limbile care dispun de sprijin avansat (cea mai avantajată între acestea fiind engleza): resursele și instrumentele lingvistice dezvoltate sunt limitate atât cantitativ cât și calitativ.

Utilizarea corpusurilor electronice, de către lingviști și ingineri din domeniul PLN deopotrivă, are deja o istorie de zeci de ani, în special în context internațional. Deși aplicațiile bazate pe prelucrarea și modelarea limbajului natural au fost inițial bazate pe reguli construite prin efortul susținut al cercetătorilor lingviști, cu timpul au luat avânt metodele statistice care funcționează extrăgând automat modele lingvistice din corpusuri electronice de mari dimensiuni. Inițial, modelele statistice se bazau pe text neprocesat și adnotat, dar cu timpul au apărut abordări care presupun adnotarea prealabilă a textului înainte de învățarea modelelor, la diferite niveluri lingvistice: la început doar la nivel morfo-lexical, ulterior la nivel sintactic şi chiar semantic. În context internațional, pentru multe aplicații din PLN, integrarea informației sintactice a condus la creșterea performanței față de algoritmii bazați doar pe informație morfologică sau față de cei ne-supervizați. Exemplificând doar pentru Traducerea Automată Statistică, diverși autori au raportat reducerea ratei erorilor atunci când au experimentat cu modele sintactice, încă de la începutul anilor 2000. În România însă facem abia primii paşi către valorificarea informaţiei sintactice în aplicaţii de Traducere Automată: un studiu din 2012 descria o metodă de extragere a unor șabloane de traducere din texte paralele Română-Engleză, adnotate cu constituenți sintactici, dar nu mergea mai departe la utilizarea șabloanelor pentru îmbunătățirea calității traducerii. Din perspectiva lingvisticii teoretice, existența unui corpus adnotat la nivel morfo-lexical şi sintactic oferă posibilitatea căutărilor avansate: înlănțuiri de cuvinte, înlănțuiri de etichete morfologice și chiar lanțuri de relații sintactice. Pe baza rezultatelor găsite se pot susține, completa sau ajusta teoriile lingvistice.

Pentru a asigura suportul tehnologic necesar nivelului de analiză sintactică a limbii, tradițional, eforturile de cercetare s-au îndreptat în două direcții: dezvoltarea de corpusuri analizate sintactic (eng. treebank, sau bancă de arbori) și dezvoltarea de analizoare sintactice (eng. parser). Primele corpusuri analizate sintactic au fost banca de arbori Lancaster (LPC, eng. Lancaster Parsed Corpus) şi banca de arbori Penn TreeBank. Realizate în anii 90, au constituit modele de urmat pentru numeroase alte proiecte asemănătoare precum băncile de arbori germane NEGRA, TIGER, corpusurile scrise sau vorbite TüBa, realizate la Tübingen pentru limbile germană, engleză și japoneză, banca de arbori cehească Prague Dependency Treebank, pentru a le enumera doar pe cele mai importante. Comunitatea dezvoltatorilor și utilizatorilor de bănci de arbori este numeroasă și activă. Anual se ține, în diverse locuri din Europa, un eveniment științific (International Workshop on Treebanks and Linguistic Theories), ajuns la a unsprezecea ediție, în care sunt prezentate ultimele realizări în domeniu.

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. Rezultatul proiectului este o bancă de 4042 de arbori (i.e. de propoziții adnotate), a căror lungime medie este de nouă cuvinte: evident, un corpus cu propoziții scurte. Autorii au evitat cazurile lingvistice problematice prin includerea exclusiv a propozițiilor, nu și a frazelor. Frazele au fost segmentate în propoziții, fiecare dintre acestea fiind analizate separat, manual. O altă bancă de arbori pentru limba română (nefinalizată şi inaccesibilă când am început această cercetare) este anunțată în 2014. Adnotarea cu relații specifice gramaticii de dependențe s-a făcut tot manual, cu ajutorul unei interfețe special dezvoltate (TreeAnnotator), şi a fost încheiată în 2015. Au rezultat 4.500 de propoziţii, cu o lungime medie de 37 de cuvinte pe propoziţie, (un total de 115.000 cuvinte), acoperind mai multe stiluri funcţionale şi perioade istorice.

În lipsa unui treebank de mari dimensiuni pentru limba română disponibil pentru antrenarea unui model statistic şi în perspectiva adnotării sintactice a corpusului computațional de referință pentru limba română contemporană CoRoLa, am decis să ne concentrăm eforturile pe dezvoltarea unui **nucleu de treebank** care să fie cât mai reprezentativ, oferind un model la scară redusă al tiparelor sintactice din limba română.

Am ales drept formalism pentru adnotarea sintactică gramatica de dependenţe (GD), care oferă o analiză ergonomică, fiind bazată pe corespondenţe unu-la-unu între cuvintele din propoziţie şi nodurile arborelui de analiză corespunzător propoziţiei. În plus, legăturile de dependenţe sunt mult mai aproape de relațiile semantice, deschizând drumul către următorul nivel de analiză a limbajului. De asemenea, analiza automată cu dependenţe are loc mult mai facil, având la bază parcurgerea cuvânt cu cuvânt a propoziţiei şi acceptarea sau ataşarea acestora

la arbore unul câte unul, fără a aştepta până când structura de constituenţi a unui anumit grup sintactic este completă pentru a ataşa întregul grup. Pentru formalismele care permit structuri de dependenţe non-proiective, GD oferă posibilitatea unui tratament adecvat al limbilor cu topică variabilă, cum este cazul limbii române. O trecere în revistă cronologică a principiilor şi evoluţiei gramaticii de depedenţe se regăseşte în Capitolul 2 a lucrării.

Am preconizat la începutul acestui proiect că resursa va avea dimensiuni limitate (5.000 de propoziţii), dar va fi caracterizată prin reprezentativitate şi diversitate, acoperind cât mai multe şabloane sintactice din limba română şi oferind o bază solidă pentru crearea unui model statistic de analiză sintactică. Pentru a capta în resursa noastră cât mai multe fenomene sintactice din limba română, aceasta trebuie să includă propoziţii din domenii şi stiluri funcţionale diverse. De aceea, am selectat propoziţiile de adnotat din ROMBAC, un corpus românesc balansat dezvoltat la ICIA. Criteriul de selecţie folosit, frecvenţa verbelor în ROMBAC, ne garantează că avem de a face cu structuri sintactice des întrebuinţate în limbă, asigurând astfel reprezentativitatea resursei noastre.

Pe baza informației morfo-lexicale din ROMBAC, am putut identifica automat verbele predicative și calcula frecvențele acestora în corpus. Ne-am concentrat pe cele mai frecvente 500 de verbe din fiecare dintre cele 5 secțiuni ale corpusului și am extras din ROMBAC câte 1.000 de propoziții din fiecare secțiune, astfel încât fiecare dintre cele 500 de verbe frecvente să apară în cel puțin două propoziții din fiecare domeniu. Cele 5.000 de propoziţii extrase astfel din ROMBAC vor reprezenta corpusul de lucru în continuare (treebank-ul). Propoziţiile selectate trebuie să aibă o lungime cuprinsă între 10 şi 40 de cuvinte şi cel puţin un verb predicativ în structură.

Pentru a compensa costurile mari de timp şi efort necesare îndeplinirii scopului enunţat, am urmărit automatizarea a cât mai multe dintre etapele proiectului. În comunitatea de cercetare sunt practicate două strategii de dezvoltare a unui treebank: 1) adnotarea manuală de la zero (sau pornind de la adnotarea morfo-sintactică) a propoziţiilor folosind un instrument grafic pentru facilitarea acesteia şi 2) adnotarea automată folosind instrumente disponibile (statistice sau bazate pe reguli) și corectarea manuală ulterioară a soluțiilor furnizate de acestea. Am optat pentru a doua strategie bazându-ne pe rezultatele pozitive obținute în experimente asemănătoare de către echipe de cercetare internaţionale şi naţionale. Exploatând similaritatea tipologică între limbile română, spaniolă și catalană, am reprodus procedura folosită de o echipă de cercetare de la IULA, institutul spaniol la care am desfăşurat stagiul de mobilitate internaţională prilejuit de bursa postdoctorală. Echipa IULA adnotase anterior un treebank catalan folosind un model statistic spaniol. Similar, am adnotat corpusul nostru cu analizorul sintactic MaltParser antrenat

pe treebank-ul de limbă spaniolă IULA LSP și am corectat rezultatele obținute. Pentru corectura manuală am folosit instrumentrul yEd , care dispune de o interfață grafică intuitivă.

O astfel de adnotare croslingvistică este posibilă deoarece MaltParser oferă opțiunea antrenării de modele statistice de-lexicalizate, bazate exclusiv pe secvențe de etichete morfo-sintactice, și nu pe cuvinte. Ne-am bazat pe faptul că cele două limbi implicate, româna și spaniola, împart șabloane sintactice instanțiate prin secvențe de părți de vorbire similare.

Pentru a menține consistența adnotării, am decis să pornim, într-o primă etapă, cu prima jumătate a corpusului de adnotat în care am inclus propoziții de lungime cuprinsă între 10 și 20 de cuvinte, și să lăsăm propozițiile mai lungi, și implicit mai complexe sintactic, pentru adnotare și corectare într-o etapă secundară. Fiecare secțiune a corpusului a fost împărțită astfel în două tranșe a câte 500 de propoziții: în prima etapă se corectează prima tranșă, cu propoziții mai scurte, din fiecare secțiune, iar în cea de-a doua etapă se corectează propozițiile de lungime mai mare rămase. Ipoteza este că, procedând în acest mod, ne vom concentra în prima parte pe familiarizarea cu principiile de corectare, aplicându-le pe propoziții mai scurte, care să pună mai puține probleme de corectare; corectura din etapa a doua va fi mai facilă, deoarece fiecare dintre seturile secundare de propoziții corectate corespunzătoare unei anumite secțiuni din text și unui anumit stil literar (jurnalistic, beletristic, academic, științific și juridic) va beneficia de un model statistic de adnotare antrenat pe datele similare din seturile corectate în prima etapă.

Am început adnotarea automată cu un set de 500 de propoziții din sub-corpusul jurnalistic folosind modelul statistic de-lexicalizat de limbă spaniolă. Am optat să începem cu stilul jurnalistic datorită intuiției că modelul statistic obținut va fi unul destul de divers atât sintactic cât și lexical (nu controlat și specific, cum ar fi fost un model antrenat pe sub-corpusul medical sau juridic, de exemplu); în același timp, datorită particularităților stilistice, ne-am așteptat ca procesul de corectură să fie mai facil decât cel al unui text beletristic, în care un limbaj figurativ poate pune probleme de interpretare sintactică și semantică chiar și unui adnotator experimentat.

Am decis antrenarea unui model lexicalizat pe limba română după doar 500 de propoziții corectate, intuind că modelul obținut va avea deja performanțe mai bune decât cel spaniol, lucru confirmat de evaluările efectuate. Am repetat procedura de reantrenare după corectura a 500 de propoziții din fiecare sub-corpus, adăugând de fiecare dată la corpusul de antrenare ultimele propoziții corectate. Ciclul de lucru este: 1) adnotare cu modelul statistic cel mai performant la dispoziție; 2) corectura setului de propoziții adnotat la pasul 1; 3) adăugarea setului corectat la corpusul de antrenare și re-antrenarea unui model extins, mai performant decât precedentul. Fiecare tranşă de propoziţii a fost corectată manual de către doi adnotatori umani, un specialist informatician şi un specialist lingvist. Adeseori, aceştia au comunicat între ei pentru a conveni asupra cazurilor de adnotare problematice.

Toate resursele şi instrumentele folosite pentru ducerea proiectului la bun sfârşit sunt descrise detaliat în Capitolul 3 al lucrării, în timp ce modul de lucru, atât pentru construirea corpusului de adnotat pe baza ROMBAC cât şi pentru adnotarea sa automată cu MaltParser şi corectarea manuală, este prezentat în Capitolul 4.

Pentru evaluarea rezultatelor, am folosit măsuri şi instrumente consacrate în domeniu. Competițiile CoNLL 2006 și CoNLL 2007, dedicate analizei sintactice cu dependențe și devenite repere de evaluare a performanței parserelor, au dezvoltat propriile scripturi Perl de evaluare, pe baza cărora s-a construit ulterior în Java instrumentul MaltEval, întrebuinţat de noi. Pe întreg parcursul proiectului au avut loc diverse tipuri de evaluări ale rezultatelor modelului statistic antrenat cu MaltParser, iniţial pe corpusul spaniol IULA LSP şi ulterior pe propoziţiile corectate românești acumulate. Rezultatele şi interpretările noastre asupra rezultatelor acestor evaluări se regăsesc în Capitolul 5. **Evoluția performanței de adnotare a acestui model este grăitoare, de la un scor LAS de 0,58 pentru prima antrenare la unul de 0,87 pentru ultima.** De altfel, dificultatea muncii de corectare a scăzut în mod evident pe parcursul procesului. În acest moment, cu un model statistic care reduce substanțial munca de corectare manuală, este fezabilă perspectiva extinderii treebank-ului dezvoltat dincolo de limita de 5.000 de propoziții pe care ne-am propus-o, mai ales că se urmărește integrarea nivelului de analiză sintactică în corpusul computațional de referinţă pentru limba română contemporană, CoRoLa, proiect prioritar al Academiei Române.

De asemenea, într-o etapă ulterioară finalizării acestui proiect, intenţionăm să folosim metodologia evaluării distorsionate pe întreg treebank-ul pentru a identifica eventualele erori de adnotare umană și a le corecta. Chiar dacă posibilitatea existenței acestui tip de eroare în treebank-ul nostru a fost redusă datorită implicării în munca de corectare a doi specialişti (cel de-al doilea revizuind munca de corectare a primului), metodologia menționată ne poate ajuta să eliminăm complet eroarea din nucleul de treebank pe care l-am dezvoltat.

Principalele contribuţii ale acestui proiect sunt:

- Dezvoltarea unui nucleu de bancă de arbori pentru limba română divers şi reprezentativ, alcătuit din 5.000 de propoziţii analizate sintactic automat cu relaţii de dependenţă şi corectate manual de către doi specialişti lingvişti;
- Dezvoltarea unui set de relaţii de dependenţă specific limbii române dar aliniabil standardelor internaţionale în domeniu;
- Dezvoltarea unui ghid de adnotare cu exemple corespunzător setului de relaţii de dependenţe stabilit;
- Antrenarea unui model statistic de limbă română cu performanţe bune în raport cu dimensiunea corpusului de antrenare (0,87 scor LAS pentru 4500 de propoziţii de

antrenare); folosit cu instrumentul statistic MaltParser, acest model poate servi la adnotarea ulterioară a altor corpusuri de limbă română.

ABSTRACT

In an era when digital information technology becomes more and more complexly intertwined with all aspects of human life, the natural language, in its fundamental role of information transmitter, is bound to digitalization. For a language to survive in the future information society and for its native speakers to freely enjoy the technological progress in their private and public life, there is an imperative need of technologies dedicated to its understanding, processing and generation. A proper technological support can secure its integration in complex intelligent applications, both web and mobile, that became so compulsory.

The project we describe is just a step in a broad strategy whose purpose is the Romanian language integration in the European digital space. There is a dramatic technological deficit for the Romanian language to overcome in relation to the languages that benefit of advance support (with English being the most advantaged): digital linguistic resources and tools developed for Romanian are limited, both quantitatively and qualitatively.

The using of electronic corpora, by both linguists and engineers, has a history of decades, especially in an international context. Although the applications dealing with processing and modelling the natural language were initially based on rules, constructed with sustained effort by linguists, in time, statistical methods were developed, that automatically extract linguistic models from big electronic corpora. Initially, statistical models were based on raw texts (unprocessed and un-annotated), but later appeared approaches based on a prior linguistic annotation of the data, at different levels: part-of-speech tagging, parsing, semantic annotation, etc. Internationally, the integration of syntactical information in NLP applications lead to better performances, in comparison with algorithms based only on morpho-lexical information or with the un-supervised algorithms. For example, since 2000, in the field of Statistical Machine Translation, different authors reported a reduced error rate when experimenting with syntactic models. Instead, in Romania we are just doing the first steps in using syntactic information in MT applications: a study from 2012 describes a method for extracting translation patterns from parallel Romanian-English texts annotated with syntactic constituents, but it does not go further to the using of these patterns in improving the quality of the translation.

From the theoretical linguistics' perspective, a corpus annotated at morpho-lexical and syntactical level offers the possibility of advanced searching: word chains, part-of-speech chains, even syntactic labels chains. On the results of these searching, linguists can adjust or complete their linguistic theories.

To assure the technological support for the syntactic analysis of a specific language, traditionally, research efforts focused on two directions: developing syntactically annotated

corpora (treebanks) and developing software tools for automatic syntactic annotation (parsers). The first syntactically analysed corpora were the Lancaster treebank (LPC, Lancaster Parsed Corpus) and the Penn Treebank. Developed in the '90, these treebanks were followed by numerous other similar projects: German treebanks NEGRA and TIGER, the written and spoken TüBa corpora for German, English and Japanese, the Czech Prague Dependency Treebank, to mention only the most important ones. The community of treebank developers and users is numerous and active. Annually, the International Workshop on Treebanks and Linguistic Theories presents the latest developments in the field.

The interest for developing a treebank for Romanian started with the one designed in the RORIC-LING project: 4.042 trees with a medium length of nine words. This was obviously an inadequate resource, since the authors excluded from it the longer sentences: they actually split all the sentences into clauses and analysed them separately, thus avoiding problematic linguistic cases. Another Romanian treebank (which was unfinished and inaccessible when we started our project) was announced at the end of 2014. The annotation with dependency relations was done manually, using a dedicated annotation interface (TreeAnnotator) and was finished in 2015. 4.500 sentences resulted, with a medium length of 37 words, covering different functional styles and historic periods.

In the absence of a big treebank for Romanian (available for training a statistical model) and in the prospect of syntactically annotating the computational reference corpus for contemporary Romanian (CoRoLA, **Co**rpus of **Ro**manian **La**nguage, under development as a priority project of the Romanian Academy), we embarked on the task of developing a **core of a treebank**, aimed to be representative and to offer a scale model of the syntactic patterns in Romanian.

The formalism chosen for annotation is the Dependency Grammar (DG) that offers an ergonomic analysis, being based on one-to-one correspondences between the words in the sentence and the nodes in the corresponding dependency tree. Moreover, the dependency links are a step further to semantic relations, paving the way to the next level of analysis for the text: the semantic level. Also, the dependency parsing is done easier, being based on covering the sentence word by word and accepting and attaching the words to the tree one by one (without having to wait for the constituency structure of a certain syntactic phrase to be completed to attach the whole phrase). For the formalisms that allow non-projective dependency structures, DG offers the possibility of adequate treatment of the relatively free word order languages (like Romanian). A chronological survey of the principles and the evolution of the DG can be found in Chapter 2 of our study.

We foresaw a resource modest in dimension (5.000 sentences), but diverse and representative for the Romanian language, covering as many of the syntactic patterns in Romanian as possible and offering a solid base for the creation of a statistical model for syntactic analysis. Therefore, the treebank must include sentences form different domains and functional styles. To assure this, we selected them from ROMBAC, a Romanian balanced corpus with five sub-sections: prose, journalism, academic, medical, juridical. The selection criterion is the frequency of the main verbs in ROMBAC, which guarantees that we deal with syntactic structures that are frequently used in the language assuring our resource's representativeness.

Based on the morpho-lexical annotation in ROMBAC, we could automatically identify the main verbs and compute their frequency in the corpus. We focused on the 500 most frequent verbs in each of the 5 sections of ROMBAC and we extracted 1.000 sentences from each section, so that each of the 500 frequent verbs occurs at least in two sentences. The 5.000 selected sentences (which count more than 10 and less than 40 words and have at least a main verb in the structure) represent our working corpus.

To reduce the time and effort costs, we wanted to automatize the annotation work as much as possible. Two strategies of treebank development are possible: 1) manual annotation from the scratch using a graphic editor to facilitate the work; 2) automatic annotation using available tools (rule-based or corpus-based) and manual correction of the automatic annotation errors. We opted for the second strategy, mainly because similar experiments conducted by international and national research teams proved to be successful. Using the typological similarity between Romanian, Catalan and Spanish, we re-enacted the procedure designed by a research team from IULA, the Spanish research centre that we visited during the international mobility stage offered by the post-doctoral scholarship. Previously, the team has been annotating a Catalan treebank using a Spanish statistical model. Similarly, we annotated our corpus using MaltParser with a model trained on IULA LSP corpus and we manually corrected the results. For the manual correction, we used the yEd instrument, with many user-friendly facilities.

Such a cross-linguistic annotation was possible because MaltParser offers the opportunity to train de-lexicalised statistical models, based only on POS tags and not words. Our assumption was that the two languages involved, Spanish and Romanian, share syntactical patterns instantiated through similar parts-of-speech.

To maintain the consistency, we started the annotation with shorter sentences and postponed the longer and more complex sentences to be annotated when we accumulated more experience in the manual correction and when the statistical model was performing better. Each section of the corpus was split in two sets of 500 sentences: the first set, containing shorter

sentences was to be annotated in the first stages of the project, while we become familiar with the correction principles; moreover, the correction of the second set, containing longer sentences,

will be facilitated by a more complex statistical model, already trained on data from each domain in the corpus (from the first corrected sets).

We started the annotation with a set of 500 sentences form the journalistic sub-corpus, using the de-lexicalised Spanish statistical model. We opted to start with the journalistic style because in this way we will obtain quicker a more diverse statistical model (not controlled and specific, as a model trained on the medical or juridical sections would have been); moreover, due to the stylistic particularities of these section, we expected the correction process to be less complicated than in the case of a belletrist text from the prose sub-section, whose figurative language can face even an experimented human annotator with syntactic and semantic ambiguities.

We decided to train a Romanian lexicalised statistical model right after the correction of the first 500 sentences, guessing that the obtained model will already have better performances than the Spanish one when used on new Romanian sentences (the guess was confirmed by subsequent evaluations). We repeated the training of the statistical model after each 500 corrected sentences, adding them to the previously corrected one in the training corpus. The working cycle is: 1) annotation with the best statistical model available; 2) correction of the sentences annotated at step 1); 3) adding the new corrected set at the training corpus and re-training an extended model, better than the previous one.

All the sentences were corrected by two human annotators, an informatician and a linguist. Often, the two specialists communicated to agree on a problematic case. In the future, we intend to use tecnhiques for the automatic identification of errors to correct any of the errors that escaped the humans annotators' vigillence.

All the resources and tools used in this project are described in detail in Chapter 3, while the working strategy, both for the selection of the corpus and for its automatic annotation and manual correction, is presented in Chapter 4.

To evaluate the results of the automatic annotation process, we used measures and tools already established in the field. The CoNLL 2006 and CoNLL 2007 competitions dedicated to dependency parsing, that became reference terms for the parsers' evaluation, designed their own evaluation Perl scripts. On the basis of these scripts was later developed the java instrument MaltEval, which we used in our evaluations. During the project, different types of evaluations were conducted: the results and our interpretation of these results are presented in Chapter 5. The evolution of the model's performance is significant, from a LAS score of 0,58 for the first Romanian model to a score of 0,87 for the last evaluation, with a model of 4.500 sentences.

Actually, from the human annotator perspective, the difficulty of the correction work considerably decreased along the process. At this point, having a statistical model that significantly reduces the correction work, the perspective of extending the core treebank is feasible, especially in the context of CoRoLa developing and the aim of introducing in CoRoLa the syntactic analysis level.

The most important contributions of this project are:

- The development of a core of a treebank for Romanian, diverse and representative, comprising 5.000 dependency parsed sentences, manually corrected by two linguists;
- The development of a set of dependency relations specific to the Romanian languages but easy to align to the international standards;
- The development of an annotation guide comprising various examples for each of the relations in the developed set;
- The training of a good Romanian statistical model taking into account the training corpus dimensions (a LAS score of 0,87 for 4.500 training sentences); this model can be used to annotate with MaltParser other Romanian corpora.

CUPRINS

CAPITOLUL 1

INTRODUCERE

1.1. Contextul general

Într-o epocă în care tehnologia informaţiei digitale devine din ce în ce mai complex interconectată cu toate aspectele vieţii umane, limbajul natural, în calitatea sa fundamentală de transmiţător de informaţie, este menit digitalizării. Pentru a supravieţui în societatea informaţională a viitorului, pentru ca vorbitorii săi nativi să se poată bucura neîngrădit de avantajele progresului tehnologic în viața publică și privată, la standardele la care au acces alți cetățeni europeni, limba română are nevoie de resurse și instrumente electronice dedicate. Acest suport tehnologic îi poate asigura integrabilitatea în complexele aplicații inteligente, mobile și web, care au devenit indispensabile.

Proiectul descris în lucrarea de faţă este doar un pas dintr-o strategie amplă de integrare a limbii române în spațiul digital european. Comisia Europeană are ca prioritate dezvoltarea unei Piețe Digitale Unice (Digital Single Market), dar, în același timp, rămâne fidelă strategiei sale de promovare a multilingvismului în societatea europeană. În acest sens, în aprilie 2015 a avut loc la Riga un summit european dedicat Pieței Digitale Unice Multilingve, la care România a participat și unde s-a angajat la producerea și promovarea de tehnologii digitale pentru înlăturarea barierelor lingvistice.

Limba română are un dramatic deficit tehnologic de recuperat în acest domeniu în raport cu limbile care dispun de sprijin avansat (cea mai avantajată între acestea fiind engleza): resursele și instrumentele lingvistice dezvoltate sunt limitate atât cantitativ cât și calitativ (vedeți studiul "Limba română în era digitală" (Trandabăţ et al., 2012), elaborat în cadrul proiectului METANET, într-o serie de studii dedicate disponibilităţii şi utilizării tehnologiei limbajului pentru 31 de limbi europene). Totuși, anterior acestui studiu și de atunci încoace, multe eforturi individuale, instituționale sau prin colaborarea mai multor instituții au avut loc în direcția micșorării acestor diferențe tehnologice. O enumerare a acestor eforturi se regăsește în studiul META-NET menționat.

La Institutul de Cercetări pentru Inteligență Artificială "Mihai Drăgănescu" (ICIA), în cadrul grupului de lucru pentru Prelucrarea Limbajului Natural (PLN), cercetările sunt concentrate în mai multe direcții, dintre care cele mai importante vor fi enumerate în continuare:

1) Dezvoltarea wordnetului românesc, RoWordnet (Tufiş şi Cristea, 2002, Barbu Mititelu et al., 2014) – o ontologie lexicală monolingvă aliniată printr-un index interlingual la Princeton

Wordnet (Wordnetul original, a cărui dezvoltare a început în 1985), și, prin acesta, la o rețea globală de wordneturi, cunoscută sub numele de Global Wordnet – a debutat la începutul anilor 2000 în cadrul proiectului internațional BalkanNet și continuă și astăzi. RoWordnet este o resursă esențială în dezvoltarea a numeroase aplicații monolingve și multilingve, precum dezambiguizarea semantică, sistemele de traducere automată, sistemele întrebare-răspuns, etc. Echipa ICIA îl dezvoltă continuu, în direcția celorlalte interese de cercetare ale sale: de exemplu, pentru o vreme ne-am concentrat exclusiv pe implementarea unor sinseturi pentru verbe, datorită preocupării pentru crearea de cadre de subcategorizare pentru acestea și utilizarea cadrelor pentru dezvoltarea unui analizor sintactic (en., parser) pentru limba română.

2) Traducerea automată este o altă preocupare importantă a cercetărilor, susținută și de participarea la proiectul internațional ACCURAT (Analysis and evaluation of Comparable Corpora for Under Resourced Areas of machine Translation) în perioada 2010-2012. Scopul acestui proiect a fost dezvoltarea de metodologii și tehnologii prin care corpusuri comparabile de mari dimensiuni să fie exploatate pentru creșterea performanțelor aplicațiilor de traducere automată prin metode statistice (Tufiş et al., 2013a). Alte direcții de cercetare abordate au fost dezvoltarea unui sistem de traducere automată bazat pe exemple (Irimia, 2009), dezvoltarea unui sistem de traducere automată pentru limbaj vorbit (Tufiş et al. 2013b), dezvoltarea de corpusuri paralele care servesc drept resurse de antrenare pentru traducătoare statistice, oferirea online, spre utilizare în scopuri de cercetare, a unui sistem de traducere statistic fiabil şi performant, pentru perechi de limbi precum engleză-română, germană-română, spaniolă-română.

3) De asemenea, ICIA este angajat, împreună cu Institutul de Informatică Teoretică din Iași, într-un program prioritar ale Academiei Române: realizarea unui corpus computațional de referință pentru limba română contemporană, denumit CoRoLa (Barbu Mititelu şi Irimia, 2014). Acesta va fi o colecție de texte în format digital (scrise și orale) de dimensiune mare (cinci sute de milioane de cuvinte). Adnotate cu metainformații – precum autor, data publicării, etc. – și cu date lingvistice – precum părți de vorbire, forma din dicționar a cuvântului adnotat, etc. – documentele vor fi disponibile liber online, spre consultare şi valorificare în scopuri de cercetare. CoRoLa va incorpora, iniţial, şi o secţiune adnotată sintactic (aproximativ 10.000 de arbori de dependenţe sintactice), ce va fi utilizată ulterior pentru antrenarea unui model statistic şi adnotarea unei părţi mai mari a corpusului folosind un analizor sintactic statistic.

Utilizarea corpusurilor electronice, de către lingviști și ingineri din domeniul PLN deopotrivă, are deja o istorie de zeci de ani, în special în context internațional. Deși aplicațiile bazate pe prelucrarea și modelarea limbajului natural au fost inițial bazate pe reguli construite prin efortul susținut al cercetătorilor lingviști, 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 ultimi ani au câştigat teren metodele hibrid, care combină cunoştinţe linvistice explicite cu metode de extragere automată a cunoştinţelor implicit codificate în corpusurile electronice.

Deşi iniţial modelele statistice se bazau pe text neprocesat şi adnotat, cu timpul au apărut abordări care presupun adnotarea prealabilă a textului înainte de învăţarea modelelor, la diferite niveluri lingvistice: la început doar la nivel morfo-lexical, ulterior la nivel sintactic şi chiar semantic. În context internaţional, pentru multe aplicaţii din PLN, integrarea informaţiei sintactice a condus la creşterea performanţei faţă de algoritmii bazaţi doar pe informaţie morfologică sau faţă de cei ne-supervizaţi. Exemplificând doar pentru Traducerea Automată Statistică, diverşi autori au raportat reducerea ratei erorilor atunci când au experimentat cu modele sintactice, încă de la începutul anilor 2000 (Och et al, 1999, Marcu şi Wong, 2002, Yamada şi Knight, 2002). În România însă facem abia primii paşi către valorificarea informaţiei sintactice în aplicaţii de Traducere Automată: un studiu din 2012 descria o metodă de extragere a unor şabloane de traducere din texte paralele Română-Engleză, adnotate cu constituenţi sintactici, dar nu mergea mai departe la utilizarea şabloanelor pentru îmbunătăţirea calităţii traducerii (Colhon, 2012).

Din perspectiva lingvisticii teoretice, existenţa unui corpus adnotat la nivel morfo-lexical şi sintactic oferă posibilitatea căutărilor avansate: înlănţuiri de cuvinte, înlănţuiri de etichete morfologice şi chiar lanţuri de relaţii sintactice. Pe baza rezultatelor găsite se pot susţine, completa sau ajusta teoriile lingvistice. De exemplu, pentru limba engleză, Sampson (2003) ilustrează cum studiile pe un corpus adnotat la nivel sintactic au scos în evidenţă faptul că propoziţiile de tipul subiect-verb intranzitiv sunt mult mai puţin frecvente decât se susţinea în anumite manuale lingvistice.

Pentru a asigura suportul tehnologic necesar nivelului de analiză sintactică a limbii, tradiţional, eforturile de cercetare s-au îndreptat în două direcţii: dezvoltarea de corpusuri analizate sintactic (eng. treebank, sau bancă de arbori[1]) şi dezvoltarea de analizoare sintactice (eng. parser).

[1]Denumirea sugestivă de bancă de arbori se datorează faptului că fiecare propoziţie analizată sintactic poate fi reprezentată grafic sub forma unui arbore: în noduri sunt cuvintele propoziţiei, iar arcele reprezintă relaţiile sintactice dintre cuvinte.

1.2. Stadiul internaţional şi naţional al cercetării în domeniu

Primele corpusuri analizate sintactic au fost banca de arbori Lancaster (LPC, eng. Lancaster Parsed Corpus, Garside et al., 1992) şi banca de arbori 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 și japoneză (http://www.sfs.uni-tuebingen.de/en/ascl/resources/corpora.html), banca de arbori cehească 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 bănci de arbori 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 românești, menţionate mai jos) numără doar câteva mii de propoziții. În cazul corpusurilor mari, performanțele se datorează unor echipe de lucru numeroase, cuprinzând atât informaticieni, cât și lingviști, care au înțeles importanța științifică, culturală și strategică a unei astfel de resurse și au investit uneori aproape un deceniu în atingerea acestui scop. Comunitatea dezvoltatorilor și utilizatorilor de bănci de arbori este numeroasă și activă. Anual se ține, în diverse locuri din Europa, un eveniment științific (International Workshop on Treebanks and Linguistic Theories), ajuns la a unsprezecea ediție, în care sunt prezentate ultimele realizări în domeniu.

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 și Popescu, 2003). Rezultatul proiectului este o bancă 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 și 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 eșuează în a reflecta, de exemplu, cazurile în care un argument verbal se realizează ca subordonată. Formalismul gramatical utilizat este gramatica de dependențe iar propozițiile reflectă stilul jurnalistic. Autorii au dezvoltat și o interfață grafică de adnotare (Popescu, 2003), care porneşte

de la text complet neadnotat, fără nici un fel de informație morfo-lexicală. Un alt rezultat al acestui proiect este un inventar de relații sintactice de dependență pentru limba română (Hristea și Popescu, 2003).

O altă bancă de arbori pentru limba română (nefinalizată şi inaccesibilă când am început această cercetare) este anunțată în Perez (2014). Adnotarea cu relații specifice gramaticii de dependențe s-a făcut tot manual, cu ajutorul unei interfețe special dezvoltate (TreeAnnotator), şi a fost încheiată în 2015 (Mărănduc şi Perez, 2015). Au rezultat 4.500 de propoziţii, cu o lungime medie de 37 de cuvinte pe propoziţie, (un total de 115.000 cuvinte), acoperind mai multe stiluri funcţionale şi perioade istorice: traducerile în limba română pentru FrameNet-ul[2] englezesc şi pentru romanul 1984 al lui George Orwell, texte beletristice româneşti, documente din Wikipedia şi din Acquis-ul Comunitar, texte politice etc. Aceasta este o resursă dezvoltată cu preocupare pentru reprezentarea complexităţii sintactice a limbii române.

Un alt corpus românesc (jurnalistic) adnotat la nivel sintactic este raportat în Bick şi Greavu (2010). Adnotarea se face cu un parser (VISL[3]) a cărui gramatică a fost scrisă prin adaptarea celei pentru limba italiană. Formalismul gramatical adoptat în VISL este gramatica de constrângeri (Constrained Grammar, (Karlsson 1990; Karlsson et al., eds, 1995)). Corpusul (de peste 21 de milioane de cuvinte) poate fi vizualizat prin căutări efectuate la adresa http://corp.hum.sdu.dk/cqp.ro.html.

Câteva încercări de creare a unor analizoare sintactice automate pentru limba română au avut loc de asemenea: Călăcean şi Nivre (2009) au antrenat MaltParser[4] pe treebank-ul dezvoltat de Hristea şi Popescu (2003) iar Seretan et al. (2010) au adaptat analizorul bazat pe reguli Fips[5] pentru limba română. Cele două parsere nu sunt disponibile pentru descărcare şi integrare în alte aplicaţii, ci doar pentru utilizare online.

1.3. Scopul şi obiectivele cercetării de faţă

În secţiunea precedentă am enumerat iniţiativele de dezvoltare de resurse şi instrumente pentru analiza sintactică a limbii române. Rezultatele acestora sunt fie insuficiente, cantitativ sau calitativ, fie inaccesibile pentru utilizare în mod independent, pentru adnotarea de noi resurse. Am menţionat de asemenea necesitatea introducerii nivelului de analiză sintactică în

[2] https://framenet.icsi.berkeley.edu/fndrupal/

[3] http://beta.visl.sdu.dk/visl/pt/parsing/automatic/

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

[5] http://www.latl.unige.ch/)

instrumentele şi aplicaţiile din PLN pentru limba română şi intenţia incorporării unui sub-corpus analizat sintactic în corpusul de referinţă CoRoLa. În lipsa unui treebank de mari dimensiuni pentru limba română disponibil pentru antrenarea unui model statistic şi în perspectiva adnotării sintactice a corpusului CoRoLa, am decis să ne concentrăm eforturile pe dezvoltarea unui **nucleu de treebank** care să fie cât mai reprezentativ, oferind un model la scară redusă al tiparelor sintactice din limba română.

Am ales drept formalism pentru adnotarea sintactică gramatica de dependenţe, care oferă o analiză ergonomică, fiind bazată pe corespondenţe unu-la-unu între cuvintele din propoziţie şi nodurile arborelui de analiză corespunzător propoziţiei. O trecere în revistă cronologică a principiilor şi evoluţiei gramaticii de depedenţe se regăseşte în **Capitolul 2**.

Am preconizat la începutul acestui proiect că resursa va avea dimensiuni limitate (5.000 de propoziţii, cu dimensiuni cuprinse între 10 şi 40 de cuvinte), dar va fi caracterizată prin fiabilitate şi diversitate, acoperind cât mai multe şabloane sintactice din limba română şi oferind o bază solidă pentru crearea unui model statistic de analiză sintactică. Treebankul de 5.000 de propoziții obținut va facilita astfel adnotarea sintactică de calitate pentru corpusul de referinţă CoRoLa. Pentru a capta în resursa noastră cât mai multe fenomene sintactice din limba română, aceasta trebuie să includă propoziţii din domenii şi stiluri funcţionale diverse. De aceea, am selectat propoziţiile de adnotat din ROMBAC, un corpus românesc balansat dezvoltat la ICIA (Ion et al., 2012).

Pentru a compensa costurile mari de timp şi efort necesare îndeplinirii scopului enunţat, am urmărit automatizarea a cât mai multe dintre etapele proiectului. În comunitatea de cercetare sunt practicate două strategii de dezvoltare a unui treebank: 1) adnotarea manuală de la zero (sau pornind de la adnotarea morfo-sintactică) a propoziţiilor folosind un instrument grafic pentru facilitarea acesteia şi 2) adnotarea automată folosind instrumente disponibile (statistice sau bazate pe reguli) și corectarea manuală ulterioară a soluțiilor furnizate de acestea. Am optat pentru a doua strategie bazându-ne pe rezultatele pozitive obținute în experimente similare (Arias et al., 2014, Florea et al., 2014). Exploatând similaritatea tipologică între limbile română, spaniolă și catalană, am reprodus procedura folosită în (Arias et al, 2014) de adnotare a unui treebank catalan folosind un model statistic spaniol. Astfel, am adnotat corpusul nostru cu analizorul sintactic MaltParser antrenat pe treebank-ul de limbă spaniolă IULA LSP[6] (Marimon şi Bel, 2014) și am corectat rezultatele obținute. O astfel de adnotare croslingvistică este posibilă deoarece MaltParser oferă opțiunea antrenării de modele statistice de-lexicalizate, bazate

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

exclusiv pe secvențe de etichete morfo-sintactice, și nu pe cuvinte. Ne-am bazat pe faptul că cele două limbi implicate, româna și spaniola, împart șabloane sintactice instanțiate prin secvențe de părți de vorbire similare. În exemplul de mai jos puteţi observa că cele două propoziţii, traduceri recipoce în română și spaniolă, corespund unor secvențe de părți de vorbire similare (diferențele sunt marcate cu caractere italice).

> Marți[adv] ,[punct] miniștrii[subst] desemnați[adj] s[pron]- au[aux] prezentat [verb] în_fața [prep] Parlamentului [subst] pentru [prep] a[aux] primi [verb] votul [subst] de [prep] învestitură [subst].
>
> Martes[adv] ,[punct] *los[det]* ministros[subst] designados[adj] se[pron] han[aux] presentado[verb] ante[prep] *el[det]* Parlamento[subst] para[prep] recibir[verb] *el[det]* voto[subst] de[prep] investidura[subst].

Pentru corectura manuală am folosit instrumentrul yEd[7], care dispune de o interfață grafică intuitivă. Pentru evaluarea rezultatelor, am folosit măsuri și instrumente consacrate în domeniu. Competițiile CoNLL[8] 2006 și CoNLL 2007, dedicate analizei sintactice cu dependențe și devenite repere de evaluare a performanței parserelor, au dezvoltat propriile scripturi Perl de evaluare, pe baza cărora s-a construit ulterior în Java instrumentul MaltEval (Nillson and Nivre, 2008).

Toate resursele şi instrumentele menţionate sunt descrise detaliat în **Capitolul 3**, în timp ce modul de lucru, atât pentru construirea corpusului de adnotat pe baza ROMBAC cât şi pentru adnotarea sa automată cu MaltParser şi corectarea manuală, este prezentat în **Capitolul 4**.

Procesul de adnotare a corpusului a fost substanțial facilitat de un stagiu de mobilitate desfășurat la Institut Universitari de Lingüística Aplicada (IULA) al universității Pompeu Fabra din Barcelona. Am avut astfel ocazia de a colabora cu o parte din echipa din spatele experimentului redactat în (Arias et al, 2014), membri activi în proiectul de dezvoltare a treebank-urilor pentru spaniolă și catalană de la IULA. Aceştia au pus la dispoziție atât expertiză, cât și resurse și instrumente concrete, după cum va reieși din secțiunile următoare.

Pe întreg parcursul proiectului au avut loc diverse tipuri de evaluări ale rezultatelor modelului statistic antrenat cu MaltParser, inițial pe corpusul spaniol IULA LSP şi ulterior pe propoziţiile corectate românești acumulate. Rezultatele şi interpretările noastre asupra rezultatelor acestor evaluări se regăsesc în **Capitolul 5**.

[7] http://www.yworks.com/en/products/yfiles/yed/

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

CAPITOLUL 2

FORMALISMUL GRAMATICII DE DEPENDENŢE

2.1. O scurtă istorie a gramaticii de dependențe

Originile formalismului gramaticii de dependenţe (GD, eng. Dependency Grammar) au fost identificate (Krujiff, 2002) în antichitate, când se crede că a fost scrisă prima gramatică de acest tip, gramatica formală a limbii sanscrite a lui Pāṇini, datată în intervalul 350-250 î.e.n. Mai târziu, gramaticieni precum Apollonius (200 e.n) sau Priscianus (500 e.n.), au fost precursorii conceptului de dependenţă prin noţiuni precum *specificarea semantică* (funcţia anumitor cuvinte este aceea de a clarifica sau circumscrie semnificaţia altor cuvinte) sau *asimetria relaţiilor* dintre cuvinte (de exemplu, un adverb are nevoie de un verb pe care să-l modifice, în timp ce un verb nu are neapărat nevoie să fie modificat de un adverb).

Sub influenţa gramaticienilor şi logicienilor antichităţii, dar şi a contactului tot mai susţinut cu limba arabă, în a cărei gramatică dependenţa sintactică era deja un concept fundamental, în evul mediu şi-a făcut loc în lingvistica europeană conceptul de dependenţă, "*dependentia*", definită de cărturari latini în funcţie de *"determinatio"* (introdus de Boethius în secolul 6 cu referire la cuantificatori)*:* dacă A *determină* pe B, atunci B este *dependent* de A. Noţiuni bazate pe relaţia de dependenţă se regăsesc şi în gramaticile modistice medievale, în special în operele lui Martin de Dacia sau Thomas de Erfurt. Deşi conceptul de dependent intră în umbră în secolele modernităţii, cel de determinant se păstrează şi este completat de alte noţiuni precum *subordonata dependentă*, *modificarea* şi *modificatorii*, *complementul*.

2.1.1. Tesnière

Noţiunea modernă de gramatică de dependenţe este atribuită lingvistului francez Lucien Tesnière (1959) şi este datată în 1939, studiile sale fiind publicate post-mortem. La baza acestei noţiuni stă ideea că între cuvintele unei propoziţii există *relaţii binare asimetrice* şi că acest set de relaţii constituie structura sintactică a propoziţiei. Asimetria relaţiei de dependenţă conduce la distincţia de tip centru/dependent între cuvintele care intră în relaţie (în terminologia lui Tesnière, *régissant/subordonné*). În accepţiunea actuală a gramaticii de dependenţe, fiecare cuvânt dintr-o propoziţie depinde de un (singur) alt cuvânt din aceeaşi propoziţie, cu excepţia cuvântului care este *rădăcina* propoziţiei (sau elementul central, elementul principal) care nu depinde de nici un cuvânt

Dar pentru Tesnière, unitatea de bază a relaţiei de dependenţă nu este cuvântul, ci nucleul, o categorie aparte de cuvinte în care intră doar cuvintele complete (fr. Tesnière "*pleines*"), sau cuvintele conţinut: verbele, substantivele, adjectivele şi adverbele. Cuvintele funcţionale (fr. Tesnière "*vides*") sunt cooptate în nucleul cuvintelor complete pe care le determină. Reprezentarea pe care a preferat-o Tesnière pentru relaţiile sintactice este cea de "stemă" (vezi Figura 2.1). Centrul este aşezat deasupra dependentului, numerele din parantezele pătrate se referă la argumentele din reprezentarea logică, iar cuvintele funcţionale sunt reprezentate înaintea unei bare verticale ce le separă de nucleul care le incorporează:

Nucleul este elementul de bază al teoriei lui Tesnière.

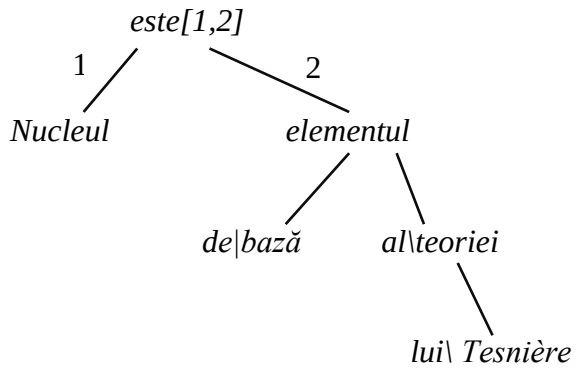


Figura 2.1 Reprezentarea în formă de stemă (preferată de Tesnière) pentru analiza sintactică cu dependenţe a propoziţiei: *Nucleul este elementul de bază al teoriei lui Tesnière.*

În teoria lingvistică a lui Tesnière, dependenţa (denumită de fapt conexiune, fr. *connexion*) reprezintă doar una din cele trei tipuri de relaţie sintactică identificate în limbă, la care se adaugă translatarea (fr. *translation*, numită şi transfer) şi joncţiunea (fr. *junction*).

Relaţia de transfer se stabileşte între cuvinte funcţionale şi cuvinte conţinut căror prezenţa cuvintelor funcţionale le schimbă categoria lexicală pentru a putea intra în relaţii de dependenţă care în mod normal nu le sunt accesibile. De exemplu, în construcţia "teoria lui Lucien", articolul "lui" intră în relaţia de transfer cu substantivul propriu "Lucien", permiţându-i acestuia din urmă să modifice substantivul "teoria", funcţionând ca un adjectiv. Conceptul de translatare sau transfer a fost puternic criticat de lingvişti, care l-au acuzat pe Tesnière de confuzie între categorii şi funcţii gramaticale. Weber (1996) combate aceste critici, susţinând că translatarea nu reprezintă o schimbare reală de categorie gramaticală, iar potenţialul de conectare a elementului translatat ca centru faţă de dependenţii săi rămâne identic. Weber vede translatarea ca pe un mijloc de a extinde clasele de valenţă, permiţând elementului translatat să completeze o valenţă la care altfel nu ar avea acces.

Joncţiunea este relaţia care leagă elemente aflate pe acelaşi nivel sintactic, în care nici unul nu poate fi văzut ca depinzând de celălalt sau celelalte elemente. Ea se stabileşte între elemente coordonate, care au acelaşi centru sau sunt centre ale aceluiaşi dependent, şi rezolvă fenomenul sintactic de coordonare care constituie o problemă serioasă în teoriile gramaticilor de dependenţe actuale, ce exclud joncţiunea. Aceste teorii sunt forţate să includă elemente de constituenţă sau relaţii de non-dependenţă pentru tratarea coordonării. În exemplul din Figura 2.2, "Joncţiunea" şi "translaţia" nu intră într-o relaţie de dependenţă (cu un termen centru şi celălalt dependent) şi trebuie să stea pe acelaşi nivel în analiza sintactică:

Joncţiunea şi translatarea sunt relaţii sintactice.

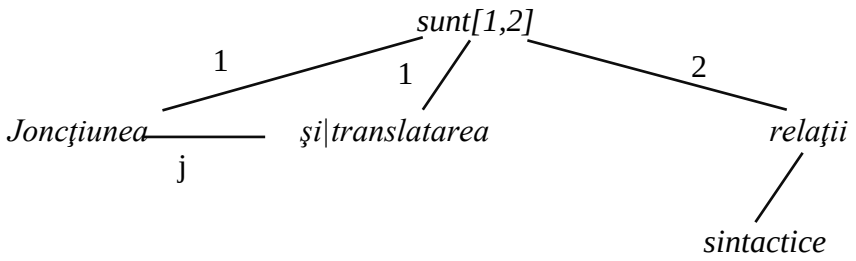


Figura 2.2 Reprezentarea în formă de stemă pentru analiza sintactică cu dependenţe a propoziţiei: *Joncţiunea şi translatarea sunt relaţii sintactice.*

Unul dintre cele mai dezbătute subiecte din domeniu, modul în care un anumit formalism tratează conflictul dintre ordinea cuvintelor în propoziţie şi ordinea elementelor în structura sintactică corespunzătoare propoziţiei, este gestionat de Tesnière prin separarea clară între ordinea lineară (fr. *ordre linéaire*) a şirurilor de cuvinte de suprafaţă şi ordinea structurală (fr. *ordre structurale*), bazată pe o reţea de relaţii gramaticale, situată pe un nivel abstract, independent de cel de suprafaţă. Sintaxa trebuie să se ocupe cu studiul ordinii structurale, în timp ce ordinea lineară ar trebui delegată morfologiei şi fonologiei. Ca o consecinţă a acestei separări iniţiale postulate de Tesnière, ordinea cuvintelor nu are un rol important în gramatica cu dependenţe, ceea ce avantajează în mod deosebit limbile cu topică mai liberă, printre care se numără şi limba română. Din acest punct de vedere, formalismul gramaticii cu dependenţe este mult mai potrivit decât, de exemplu, al descrierii din gramaticile de Guvernare şi Legare (eng. Government and Binding, GB), care trebuie să includă mişcări de topicalizare complexe pentru a gestiona ordinea cuvintelor în propoziţie.

2.1.2. Hays şi Gaifman

Deşi a avut mult mai puţini susţinători decât formalismul gramaticilor de constituenţi care se dezvolta în paralel, gramatica de dependenţe a beneficiat de primele încercări de formalizare în anii ’60, când Hays (1964) şi Gaifman (1965) i-au studiat proprietăţile matematice. Dar înainte de a discuta concluziile acestui studiu, trebuie să menţionăm axiomele introduse de Robinson (1970):

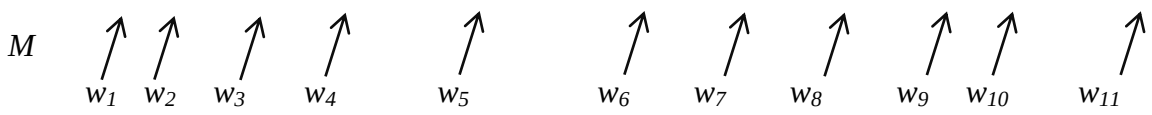
1. Unul şi numai unul dintre elemente este independent.
2. Toate celelalte elemente depind în mod direct de un alt element.
3. Nici un element nu depinde în mod direct de mai mult de un element.
4. Dacă A depinde în mod direct de B şi un alt element C intervine între ele (în ordinea lineară a şirului de cuvinte în propoziţie), atunci C depinde direct de A sau de B sau de alt element care intervine între ele şi care nu este C.

Din primele trei axiome se poate deduce concluzia că graful asociat analizei sintactice cu dependenţe a unei propoziţii este de fapt un *arbore*, a cărui *rădăcină* nu depinde de nici un alt element al propoziţiei. Condiţia trei este cea de *centru unic* pentru fiecare dependent, înglobată de cele mai multe dintre variantele GD. Cea de-a patra axiomă, numită astăzi condiţia de proiectivitate a arborelui, interzice intersecţia muchiilor într-un arbore de dependenţe. Această condiţie are efect asupra corespondenţei dintre ordinea nodurilor în arbore şi cea a cuvintelor în propoziţie, fiind foarte dificil de satisfăcut de limbile care nu au topică fixă. De altfel, Tesnière nu a impus această condiţie şi multe dintre teoriile GD moderne au renunţat la ea, deoarece privează GD de cel mai important avantaj al său, compatibilitatea cu limbile cu topică relativ sau complet liberă.

Pe baza primelor 3 axiome, Debusmann (2000) descrie formal gramatica de dependenţe după cum urmează:

Fie R o relaţie binară de dependenţe definită pe mulţimea W a cuvintelor dintr-o propoziţie, $R \subset W \times W$. O funcţie M stabileşte corespondenţa între elementele mulţimii W şi mulţimea efectivă a cuvintelor din propoziţie, ca în exemplul de mai jos:

Hays şi Gaifman studiază proprietăţile matematice ale gramaticii cu dependenţe .



Proprietăţile lui R sunt:

1. $\forall w_1, w_2, \dots, w_{k-1}, w_k \in W: \langle w_1, w_2 \rangle \in R \dots \langle w_{k-1}, w_k \rangle \in R: w_1 \neq w_k$ (aciclicitate)
2. $\exists! w_1 \in W: \forall w_2 \in W, \langle w_1, w_2 \rangle \notin R$ (existenţa şi unicitatea rădăcinii)

3. $\forall w_1, w_2, w_3 \in W: \langle w_1, w_2 \rangle \in R \wedge \langle w_1, w_3 \rangle \in R \rightarrow w_2 = w_3$ (proprietatea de centru unic)

Din aciclicitate rezultă şi proprietatea de asimetrie, cea care definea anterior relaţia de dependenţă:

$$\forall w_1, w_2 \in W: \langle w_1, w_2 \rangle \in R \rightarrow \langle w_2, w_1 \rangle \notin R$$

Din asimetrie decurge ireflexivitatea:

$$\forall w_1 \in W: \langle w_1, w_1 \rangle \notin R$$

Revenind la Hays (1964), acesta a formalizat o regulă de dependență ca pe o specificare asupra valenței unei anumite unități sintactice. În viziunea lui Hays, gramatica de dependențe, asemenei celei de constituenți, folosește două alfabete, unul terminal (un lexicon sau o listă de morfeme) și unul ne-terminal, alcătuit dintr-o listă de nume sau simboluri asociate tipurilor de grupuri sintactice (eng. *phrases*). O funcție de atribuire realizează corespondența între elemente ale alfabetului terminal și elemente ale alfabetului ne-terminal. O regulă de dependență se definește pe alfabetul ne-terminal și constă dintr-un simbol ne-terminal care guvernează un număr finit de simboluri ne-terminale dependente:

$$(x_{j_1}, x_{j_2}, \dots, *, \dots, x_{j_n})$$

În reprezentarea de mai sus, x_i este elementul care guvernează (centrul) iar $x_{j_1}, x_{j_2}, \dots, *, \dots, x_{j_n}$ este o listă de *n* elemente dependente, care indică valența centrului. O astfel de regulă ordonează implicit elementele dintre paranteze, unde * este poziția guvernantului în lista de dependenți. Regula $x_i(*)$ indică faptul că x_i poate fi o frunză în arborele de dependențe, în timp ce $*(x_i)$ înseamnă că x_i poate fi rădăcina arborelui.

Hays exemplifică cu o regulă specifică limbii engleze:

$V_\propto(N_{p_1}, *, N, D_\beta)$, unde $V_\propto$ este o clasă de verbe, N_{p_1} o clasă de substantive la numărul plural, *N* o clasă de substantive iar D_β o clasă de adverbe. Un șir de cuvinte precum "Children eat candy neatly" (ro. "Copiii mănâncă bomboane cu grijă") corespunde regulii enunțate.

Gramaticile de dependențe Hays-Gaifman (GDHG) îndeplinesc toate axiomele lui Robinson, inclusiv pe cea de proiectivitate. Hays (1964) demonstrează de asemenea că GDHG sunt slab echivalente (eng. "weakly equivalent") cu Gramaticile Independente de Context (GIC, eng. Context-Free Grammars), în sensul că au același alfabet terminal și, pentru fiecare șir din acest alfabet, fiecare structură atribuită de oricare dintre cele două gramatici corespunde unei structuri atribuite de cealaltă gramatică (dar structurile nu sunt neapărat identice). Astfel, condiția de proiectivitate, obligatorie în GIC, se impune încă o dată ca și consecință a definiției GDHG și face ca acest tip de gramatici să fie privite în comunitatea de cercetare doar ca niște variante notaționale ale GIC.

Ca reacție, Duchier (1999) propune o GD non-proiectivă, în care restricțiile de ordine a cuvintelor sunt specificate separat, prin secvențe de categorii gramaticale, ca în exemplul *Seq(det(w),adj(w),n(w))*, care constrânge un determinator să preceadă un adjectiv care la rândul său precede substantivul pe care îl modifică.

2.1.3. Mel'čuk

În paralel, în anii ’80, Mel'čuk introduce în Statele Unite ale Americii tradiţia gramaticii de dependenţe, care stătea la baza sintaxei în cele mai multe dintre ţările vorbitoare de limbi slave (Hudson, 1990). Mel'čuk (1988) declară că nu şi-a propus o formalizare a gramaticii de dependenţe, nu şi-a propus să dea definiţii, ci să construiască un instrument care, “lăsând la o parte erorile simple sau inconsistenţele, poate fi evaluat doar în termeni de eficacitate sau naturaleţe, dar nu în termeni de adevăr sau fals”[9]. În schimb, el face o clasificare a tipurilor de dependenţe, o distincţie între dependenţa *morfologică, sintactică* şi *semantică*, sugerând că alte tipuri de dependenţe, precum legăturile anaforice, pot fi recunoscute în limbă. În legătură cu dependenţa morfologică, Mel'čuk (1988) observă că:

1) orice limbă are categorii gramaticale invariabile morfologic, ceea ce conduce la apariţia unor discontinuităţi în lanţul de dependenţe; din acest motiv, el califică dependenţa morfologică drept un “tip marginal” de dependenţe.
2) o dependenţă morfologică poate fi simetrică (ex. în sintagma ”două fete”, numeralul este dependent de substantiv cu privire la genul său, în timp ce substantivul este dependent de numeral cu privire la număr);
3) un cuvânt poate fi dependent morfologic de mai multe cuvinte (condiţia de centru unic este încălcată);

Dependenţa semantică, spre deosebire de cea morfologică, este universală, aplicându-se tuturor cuvintelor din propoziţie (cu câteva excepţii), şi este unilaterală (sau asimetrică). Totuşi, asemeni dependenţei morfologice, permite mai mulţi guvernanţi pentru acelaşi termen. Adeseori, dependenţa morfologică şi cea semantică merg în sensuri diferite, ca atunci când, de exemplu, un substantiv determină morfologic genul, numărul şi cazul articolului său în timp ce, din punct de vedere semantic, articolul este cel care determină subsantivul.

Dependenţa sintactică completează celelalte două tipuri, asigurând conectivitatea tuturor elementelor din propoziţie. Ea este asimetrică şi satisface condiţia de centru unic. Deşi nu încearcă definirea formală a dependenţei sintactice, Mel'čuk (1988) propune mai multe criterii

[9] Traducerea noastră pentru citatul: „... leaving aside simple errors and inconsistencies, it can be evaluated solely in terms of expediency or naturalness, not in terms of truth or falsity.”

pentru a identifica dacă două elemente sunt conectate printr-o relaţie de dependenţă (criteriul de corespondenţă lineară şi cel de corespondenţă prozodică), pentru a identifica direcţia relaţiei de dependenţă –sau care este centrul şi care este dependentul în cazul a două elemente conectate – (criteriul rolului sintactic, criteriul punctului de contact morfologic, criteriul omisibilităţii şi cel al predictabilităţii) sau pentru a identifica tipul relaţiei de dependenţă (criteriul contrastului semantic în perechi minimale, criteriul substituibilităţii reciproce a arborilor sau criteriul repetabilităţii dependenţei).

În apărarea formalismului gramaticii de dependenţe, Mel'čuk (1988) combate criticile considerate nejustificate precum:

1) existenţa în limbă a dublei dependenţe: în exemplul, "Demonstraţia s-a dovedit dificilă", "dificilă" poate fi considerat ca fiind dependent atât de verb, cât şi de substantiv. Mel'čuk susţine că dependenţa faţă de substantiv nu este una de natură sintactică, ci de natură semantică. El argumentează şi exemplifică cu situaţii din alte limbi în care participiile depind morfologic de grupul nominal care constituie subiectul fără a fi dependente sintactic de acesta. Soluţia este separarea clară între nivelele morfologic, semantic şi sintactic;
2) controversa legată de dependenţa mutuală între subiect şi verbul principal din propoziţie: este de asemenea nejustificată în viziunea lui Mel'čuk, pentru care rădăcina este verbul principal iar dependenţa acestuia de subiect este de natură morfologică, nu sintactică.

2.1.4. Alte şcoli importante în GD

Alte şcoli lingvistice importante care au contribuit la dezvoltarea GD sunt:

1) şcoala germană, reprezentată de Helbig (1992) şi Engel (1994, 1996). Contribuţia lui Helbig este esenţială pentru dezvoltarea teoriei valenţei, centrală în GD, în timp ce Engel a redactat o gramatică practică de referinţă bazată pe GD şi mai multe studii teoretice;
2) şcoala finlandeză, care a debutat cu studii introductive şi istorice asupra GD de autori precum Korhonen (1977) sau Tarvainen (1981), dar a continuat cu dezvoltări originale, precum sistemul Functional Dependency Parsing Language (Karlsson, 1990) şi analizorul cu dependenţe dezvoltat de Järvinen şi Tapanainen (1997);
3) şcoala de la Praga (Sgall et al., 1986, Hajičová et al., 1995) este cea care a contribuit cel mai mult la cercetarea dependenţelor semantice, adresând semantica Montague, şi este de asemenea cea care a optat explicit pentru un sistem de reprezentare multistratal: descrierea funcţională generativă la nivelul funcţional de suprafaţă şi

reprezentarea tecto-gramaticală la nivelul funcţional de adâncime, care aduce categoriile funcţionale în nucleul semantic şi foloseşte o theta-teorie redusă pentru valenţele verbelor.

2.1.5. Distincții și variațiuni în GD

Un alt tip de distincție între relațiile de dependențe este cea propusă de Nikula (1986) între relațiile din construcțiile *exocentrice* și cele din construcțiile *endocentrice* (concepte care își au originea în studiile lui Bloomfield (1933)). O construcție gramaticală este endocentrică dacă are aceeași funcție lingvistică cu unul dintre elementele sale și exocentrică dacă nu îndeplinește acest criteriu. De exemplu, relația între adjectiv și substantiv într-un grup nominal se stabilește într-o construcție endocentrică, deoarece substantivul poate înlocui grupul nominal fără să dezmembreze structura sintactică, în timp ce relația dintre prepoziție și substantiv într-un grup prepozițional se stabilește într-o construcție exocentrică, prepoziția nefiind în măsură să înlocuiască grupul prepozițional.

Această distincție este legată de cea dintre relațiile centru-complement și centru-modificator din teoriile sintactice contemporane. În timp ce aceste tipuri de relații au o analiză clară în gramatica de dependențe, pentru alte tipuri de structuri, precum construcții cu verbe auxiliare, articole, prepoziții, conjuncții subordonatoare, teoriile din GD nu se pun de accord cu tratamentul acestora. De exemplu, unele dintre teorii consideră verbul auxiliar drept centru pentru verbul lexical, alte teorii fac alegerea opusă, în timp ce o altă categorie de teorii consideră că relațiile din complexul verbal nu sunt dependențe (Nivre, 2005). Divergența de analiză pentru structurile cu cuvinte funcționale se datorează existenței a două tipuri de criterii în selecția centrului: *criterii sintactice* și *criterii semantice*. De exemplu, cele mai multe versiuni de DG tratează prepoziția drept centru al grupului prepozițional, conform logicii criteriului sintactic, în timp ce altele consideră că dependența este de fapt semantică, între verb și (de exemplu) substantivul din grupul prepozițional, în timp ce propoziția nu are decât un rol de dependent față de acest substantiv.

Un alt punct de inflexiune în teoria GD este dat de anumite formalisme care consideră relația de dependență insuficientă pentru analiza sintactică a limbajului natural, cum de altfel o considera și Tesnière, care o completase cu relațiile de translatare și joncțiune. Hellwig (1986, 2003), Mel'čuk (1988) şi Hudson (1990) exploatează în paradigmele pe care le construiesc posibilitatea de a permite o formă redusă de analiză de constituenți, îndeosebi pentru tratamentul coordonării.

Teoriile din GD variază și cu privire la inventarul de tipuri de relații de dependențe pe care le admit. Cele mai multe adoptă fie un set mai mult sau mai puțin detaliat de funcții

gramaticale de suprafață (subiect, obiect direct, obiect indirect, modificator nominal, etc.) fie un set de tipuri de roluri semantice, provenind din tradiția relațiilor tematice (agent, pacient, scop, etc.). Alternativ, Mel'čuk (1988), folosește indici numerici pentru dependenții din cadrul de valență și etichete descriptive pentru celelalte tipuri de dependenți. Opțiunea de a nu eticheta relațiile de dependențe este şi ea comună în sistemele practice de analiză sintactică.

2.1.6. Analiză sintactică automată cu dependențe

Pornind de la tradiţia teoretică a GD, s-au implementat, cu diferite grade de fidelitate teoretică, sisteme computaţionale de analiză sintactică a limbajului natural, sau analizoare sintactice (eng. *parser*). Astfel de sisteme produc reprezentări ce conţin noduri lexicale conectate prin arcuri de dependenţă, etichetate sau nu cu tipuri de dependenţe. Asemeni altor tipuri de aplicaţii din PLN, parserele au fost dezvoltate la început pe baza regulilor gramaticale iar ulterior s-a trecut la abordări bazate pe date (pe corpus) şi eventual la abordări hibrid.

Algoritmii de analiză sintactică cu dependenţe bazaţi pe reguli gramaticale pot fi clasificaţi în următoarele categorii:

1) algoritmi care, pornind de la echivalenţa GD cu GIC, sunt foarte asemănători cu cei deja folosiţi pentru GIC: de ex, Hays (1964) propune un algoritm de programare dinamică de jos în sus (eng. bottom-up) similar algoritmului CKY (Kasami, 1965, Younger, 1967); mai recent, Sleator şi Temperly (1991,1993), Lombardo şi Lesmo (1996), Barbero et al., 1998, propun şi ei algoritmi derivaţi din CKY sau Early (1970);
2) algoritmi bazaţi pe analiză eliminatoare: pentru o anumită propoziţie de analizat, reprezentări sintactice care nu sastisfac anumite constrângeri sunt eliminate până când se ajunge la o listă de reprezentări valide (Karlsson, 1990; Karlsson et al., 1995, Maruyama, 1990, Harper şi Helzerman, 1995, Tapanainen şi Jarvinen, 1997; Jarvinen şi Tapanainen, 1998, Duchier, 1999, 2003).
3) algoritmi bazaţi pe o GD simplificată şi o strategie de analiză deterministă (Convington, 2001): în versiunea sa cea mai utilizată, această strategie presupune parcurgerea propoziţiei de la stânga la dreapta şi încercarea de a lega fiecare cuvânt curent ca centru sau dependent pentru cuvântul precedent. Gramatica simplificată presupune doar o funcţie booleană care să specifice dacă, pentru o pereche de cuvinte w_1 şi w_2, w_1 poate fi centru pentru w_2. Algoritmul lui Convington (2001) are o complexitate $O(n^2)$ şi poate fi adaptat pentru limbi cu topică liberă, flexibilă sau fixă. Pe baza algoritmului lui Convington, Nivre (2003), (Obrebski, 2003) și Kromann (2004) au dezvoltat propriile strategii de analiză sintactică cu dependenţe.

Primele încercări de analiză sintactică cu dependențe bazată pe date au fost de fapt strategii bazate pe gramatici care foloseau modele probabilistice extrase din corpusuri doar pentru dezambiguizare (Caroll şi Charniak, 1992). Mai târziu, Eisner (2000) dezvoltă modele probabilistice de analiză sintactică pe care le reuneşte sub noţiunea de gramatică bilexicală. Toate modelele lui Eisner sunt distribuţii de probabilitate comună pe etichete morfo-sintactice, cuvinte şi legături de dependenţe. Modelul C, cel care produce rezultatele cele mai bune conform Eisner (1996) este definit astfel:

$$P(tw_1, \dots, tw_n, links) = \prod_{i=1}^{n} P(lc_i|tw_i)P(rc_i|tw_i)$$

unde tw_i este al i-lea cuvânt adnotat morfo-sintactic din propoziţie, lc_i este copilul stâng al cuvântului *i*, rc_i este copilul drept al cuvântului *i*.

Probabilitatea de generare a fiecărui copil este condiţionată de cuvântul centru adnotat morfo-sintanctic şi de eticheta morfo-sintactică a copilului precedent (copiii stângi se generează de la dreapta la stânga; într-o reprezentare de tip (centru, dependent) al relaţiei de dependenţă, dependentul este copilul drept al centrului, iar centrul este copilul stâng al dependentului):

$$P(lc_i|tw_i) = \prod_{j=1}^{m} P(tw_{lc_{j,i}}|t(lc_{j-1,i}), tw_i)$$

$$P(rc_i|tw_i) = \prod_{j=1}^{m} P(tw_{rc_{j,i}}|t(rc_{j-1,i}), tw_i)$$

unde $lc_{j,i}$ este al j-lea copil stâng al cuvântului i iar $t(lc_{j-1,i})$ este eticheta morfo-sintactică a copilului stâng precedent (*j-1*) (analog $rc_{j,i}$ şi $t(rc_{j-1,i})$ pentru copiii drepţi).

Samuelson (2000) este primul care propune un model probabilistic cu dependenţe etichetate şi care permite structuri ne-proiective. Acest model (care nu a fost niciodată implementat) conţine două procese stocastice, unul de sus în jos (eng. top-down) care generează structura arborelui de dependenţe şi altul de jos în sus care generează şirul de suprafaţă dată fiind structura arborelui. Wang şi Harper (2004) implementează o extensie a modelului CDG (Constrained Dependency Grammar, Maruyama, 1990) cu un model probabilistic generativ cu dependenţe etichetate şi obţin rezultate performante.

În paralel, se dezvoltă modele complet discriminative de învăţare inductivă combinate cu strategii de analiză deterministă, care nu mai implică deloc gramatici formale:

1) Kudo şi Matsumoto (2000), Yamada şi Matsumoto (2003) folosesc maşini de vectori de suport (eng. support vector machines) pentru a antrena clasificatoare care prezic

următoarea acţiune a analizorului determinist construind arbori de dependenţe fără etichete;

2) Nivre et al. (2004) propune o analiză inductivă pentru a produce reprezentări etichetate cu tipuri de deptendenţă, folosind tehnica de învăţare bazată pe memorie.

2.1.7. Avantajele gramaticii de dependenţe

Deşi formalismul gramaticilor de constituenţi a avut un loc central şi tradiţional în teoriile lingvistice, gramaticile de dependenţe au câştigat teren, în special în lingvistica computaţională. Potrivit lui Convington (2001), GD oferă avantajul minimalismului (fiecare nod din structură corespunde unui cuvânt din propoziţia analizată, cu excepţia nodului rădăcină care este un nod artificial şi reprezintă întreaga propoziţie), ceea ce face ca structurile obţinute să ocupe mai puţin spaţiu de reprezentare iar prelucrarea lor cu instrumente informatice să fie mai uşoară. În plus, legăturile de dependenţe sunt mult mai aproape de relaţiile semantice, deschizând drumul către următorul nivel de analiză a limbajului. De asemenea, analiza automată cu dependenţe are loc mult mai facil, având la bază parcurgerea cuvânt cu cuvânt a propoziţiei şi acceptarea sau ataşarea acestora la arbore unul câte unul, fără a aştepta până când structura de constituenţi a unui anumit grup sintactic este completă pentru a ataşa întregul grup.

Pentru formalismele care permit structuri de dependenţe non-proiective, GD oferă posibilitatea unui tratament adecvat al limbilor cu topică variabilă, cum este cazul limbii române.

Nivre (2005) concluzionează că există un compromis între expresivitatea reprezentării sintactice pe care o oferă gramaticile de constituenţi şi facilitatea analizei sintactice automate şi a stocării concise a datelor pe care o oferă gramaticile cu dependenţe, dar că acestea din urmă sunt "suficient de expresive pentru a fi utile în sistemele de prelucrare a limbajului natural dar şi suficient de restricţionate pentru a permite analiză automată completă cu înaltă acurateţe şi eficienţă".

2.1.8. Gramatica de dependenţe câştigă teren

În ultimii 10 ani, interesul pentru gramatica de dependențe a crescut în comunitatea de cercetare, în special pentru că un număr tot mai mare de limbi, printre care și limbi cu topică liberă sau relativ liberă, au primit atenție în cercetarea și industria PLN; preocuparea pentru GD se reflectă în:

1. organizarea de conferințe regulate dedicate acestui formalism: Conferința Internațională pentru Lingvistica Dependenței, Depling 2011, 2013, 2015[10];

[10] http://depling.org/dependency.php

2. organizarea de competiții pentru sisteme dedicate unor probleme din domeniu: competiția CoNLL[11] 2006/2007 pentru analiză sintactică cu dependențe în context multilingv; competiția CoNLL 2008/2009 pentru analiza comună a dependențelor sintactice și semantice; SANCL 2012, competiția pentru analiza sintactică a web-ului organizată de Google; SemEval 2014/2015: analiza cu acoperire largă a dependențelor semantice (recuperarea relațiilor predicat-argument pentru toate cuvintele conținut);
3. dezvoltarea de instrumente de analiză sintactică și de resurse adnotate cu relații de dependențe (treebank-uri)
4. inițiativa de standardizare Universal Dependencies (UD, ro. Dependențe Universale), care își propune unificarea și coordonarea croslingvistică a adnotării cu dependențe sintactice a corpusurilor. Obiectivele sale principale sunt 1) un inventar universal de tipuri de relații de dependență bazat pe Universal Stanford Dependencies (De Marneffe et al., 2014) și 2) instrucțiuni menite să asigure consistența adnotării pentru construcții similare din limbi diferite dar, în același timp, să recunoască și să includă relații gramaticale specifice anumitor limbi. Ca orice proiect de standardizare, UD este esențial pentru că facilitează cercetarea croslingvistică și dezvoltarea de tehnologie multilingvă bazată pe sintaxă.

2.2. Gramatica utilizată pentru adnotare

Am ales pentru adnotarea resursei noastre formalismul gramaticii de dependenţe pentru toate avantajele enumerate la sfârșitul secțiunii anterioare. Setul de etichete folosit (denumit în continuare ROdep) a fost obținut prin îmbinarea a două seturi pe care le-am avut la dispoziție, la care am adăugat etichete noi pentru relații din gramatica românească ce nu aveau corespondent în nici unul din cele două seturi. De asemenea, principiile de adnotare, precum folosirea criteriului sintactic sau a criteriului semnatic pentru alegerea centrului, sunt în primul rând în concordanţă cu principiile gramaticii româneşti tradiţionale.

Deoarece am implicat în procesul de adnotare automată un model statistic antrenat pe corpusul IULA LSP, munca de corectare manuală a început pe un set de propoziții adnotate cu etichetele folosite în acest corpus (denumite în continuare etichete iulaLSPdep). Astfel, avea sens să dezvoltăm pentru limba română un set de etichete în care să integrăm etichetele iulaLSPdep, pentru a ne ușura munca de corectare manuală. În același timp, este foarte important să producem o adnotare sintactică în concordanță cu normele internaționale, pentru a facilita

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

utilizarea resursei noastre în proiecte multilingve viitoare. De aceea, ne-am îndreptat către inițiativa de standardizare croslingvistică a metodologiei de adnotare sintactică UD, de unde am împrumutat un important număr de etichete, în special pentru adnotarea fenomenelor de discurs, care erau ne-adnotate în iulaLSPdep. Dar cel mai important aspect de urmărit a fost respectarea principiilor gramaticii românești și evidențierea clară a relațiilor sintactice specifice limbii române.

În continuare, vom detalia şi exemplifica tipurile de relaţii utilizate în adnotarea noastră, grupându-le în funcţie de provenienţa lor. În toate exemplele din această lucrare, interpretarea tripleţilor cu care vom reprezenta relaţia de dependenţă este următoarea: primul termen este dependentul relaţiei, al doilea termen este centrul iar ultimul termen este eticheta relaţiei de dependenţă. De exemplu, în (i, auzit, *posclitic*), cliticul "i" depinde de verbul "auzit" iar relaţia dintre ei este *posclitic*.

2.2.1. Relații introduse pentru a ne conforma gramaticii limbii române

Am introdus etichete noi, motivate de următoarele decizii:

- Clasificarea cliticelor pronominale: de dublare (*dblclitic*), posesiv (*posclitic*), reflexiv şi reciproc (ambele ca *reflclitic*).

  Exemplu:

  1. Lui Ion i-am auzit vocea. (clitic posesiv**)**

  (i, auzit, *posclitic*)

  2. I-am spus lui Ion să vină. (cliticul de dublare)

  (I, spus, *dblclitic*) (Ion, spus, *iobj*)

  3. M-am răzgândit. (cliticul reflexiv)

  (M-, răzgândit, *reflclitic*)

  4. Copiii se împing și se bat. (cliticul reciproc)

  (se, împing, *reflclitic*), (se, bat, *reflclitic*)

- Relaţia *poss* a fost introdusă pentru etichetarea complementului în dativ ce exprimă posesia.

  Exemplu:

  Mi-am luat haina.

  (Mi, luat, *poss*)

- Diferenţierea între două argumente ale verbului care apar în aceeași propoziție în cazul acuzativ: obiectul direct (*dobj*, identificat, cel mai adesea, prin posibilitatea dublării prin clitic și prin prepoziția „pe") și obiectul secundar (*secobj*).

  Exemplu :

L-au învăţat pe Ion un dans.

(L, învăţat, *dblclitic*), (Ion, învăţat, *dobj*), (un dans, învăţat, *secobj*)

- Relația *post* leagă de centru o prepoziție dependentă care apare după el în ordinea liniară a propoziției (în postpoziție).

  Exemplu:

  El este teribil de timid.

  (de, teribil, *post*)

- Atunci când un dependent intră într-o relație ternară (sintactică și semantică) cu verbul și cu un nominal (subiect sau obiect), acesta este legat de verb și primește eticheta *spe* (element predicativ suplimentar).

  Exemplu:

  I-am văzut pe copii împreună.

  (împreună, văzut, *spe*)

- Elementele corelative primesc eticheta *correl*:

  Exemplu:

  A ascultat-o fie$_1$ pe Maria, fie$_2$ pe Ana.

  (fie$_1$, fie$_2$, *correl*), (fie$_2$, pe Maria, *cc*).

- În propoziţiile subordonate introduse de conjuncţii subordonatoare, centrul verbal din subordonată este legat de acestea prin relaţia *sc*:

  Exemplu:

  Ştiu că ai mult de lucru.

  (că, ştiu, *dobj*), (ai, că, *sc*)

În tabelul de mai jos prezentăm, comparativ, inventarul de relații de dependență folosit pentru analiza limbii române, pe cel folosit pentru limba spaniolă (iulalLSPdep) și pe cel utilizat în proiectul UD, având ca etalon setul ROdep (elemente din UD care, din diverse motive, nu au correspondent în ROdep nu apar în tabel):

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	CONJ	cc
compound	-	compound
conj	COORD, 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	MPAS	-
pmod	MOD, PP-LOC, PP-DIR	-
pobj	OBLC	-
poss	-	-
possclitic	-	-
post	-	-
pred	PRD, ATR	root, xcomp, ccomp
prep	COMP	case
punct	PUNCT	punct
reflclitic	MPRON, MIMPERS	expl
remnant	SUBJ-GAP, COMP-GAP, MOD-GAP	remnant
reparandum	-	reparandum
root	ROOT	root
sc	-	-
secobj	-	iobj

spe	OPRD	xcomp
subj	SUBJ	nsubj, nsubjpass, csubj, cubjpass
voc	VOC	vocative
xcomp	OPRD	xcomp

Tabelul 2.1. Tablou comparativ al relațiilor sintactice în ROdep, iulaLSPdep și UD

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

- în UD, prepozițiile și conjuncțiile nu pot fi centre de grup sintactic, ci doar determinanți: prepozițiile sunt legate prin relația *case* de nominalul pe care-l însoțesc, conjucțiile subordonatoare sunt *mark* față de centrul propoziției subordonate.
- în iulaLSPdep, prepozițiile sunt centre de grup sintactic, 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 considerarea 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.

2.2.2. 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 față de o propoziție cu verbul la diateza activă) și *csubjpass* (realizat ca subordonată față de 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 propozițional, se folosește pentru ambele cazuri aceeași etichetă de relație: *ccomp*;

- *BYAG* devine la noi *agc* (complement de agent): În UD, el nu se marchează printr-o relație specială, ci ca *nmod*, cu un determinant legat prin *case* (prepoziția care îl introduce). Deși element cu ocurență opțională în propoziție, complementul de agent apare în cadrul de subcategorizare al verbului centru al propoziției, motiv pentru care am convenit să-l marcăm diferit de alți modificatori substantivali;

- *OBLC* devine la noi *pobj* (obiect prepozițional): acesta este un determinant obligatoriu al predicatului, care are ca centru o prepoziție selectată de acesta. Statutul de complinire obligatorie a predicatului ne determină să-l tratăm diferit de grupurile prepoziționale care funcționează ca modificatori (și care sunt analizate ca *pmod*, vezi mai jos). Imposibilitatea prepoziției de a fi centru de grup în UD face ca acestei relații să nu îi corespundă vreo relație în UD;

- *COMP* devine la noi *prep* (complementul prepoziției în grupul prepozițional);

- *OPRD* devine la noi *spe* (predicativul suplimentar).

3. rafinarea relațiilor prea generale:

- *AUX*: am marcat diferit verbele auxiliare în funcție de diateza la care se află verbul pe care îl însoțesc: pentru diateza activă am folosit relația *aux*, iar pentru diateza pasivă am folosit relația *auxpass*, după modelul oferit de UD;

- *MOD*: în setul iulaLSPdep, desemnează orice tip de modificator (element a cărui apariție în propoziție este facultativă), indiferent de partea de vorbire prin care se realizează; după model UD, am ales să distingem între: *nmod* (modificator realizat ca substantiv sau pronume), *advmod* (modificator realizat ca adverb) și *appos* (apoziția). În plus față de UD, adnotăm și *pmod*, un modificator realizat ca grup prepozițional;

- *SPEC* (specificator) are ca echivalenți în adnotarea noastră *amod* (modificator adjectival) și *det* (determinatorii, în cazul nostru doar articolele), după modelul oferit de UD.

4. unificarea unor relații:

- am considerat că diferențierea între *PP-LOC* (complement circumstanțial de loc) și *PP-DIR* (complement circumstanțial de loc, ce indică direcția) nu este justificată dacă nu se fac și 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 și 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", ro. „a fi") și *PRD* (pentru numele predicative ale celorlalte verbe copulative). În spiritul lingvisticii românești, am decis să nu folosim etichete diferite pentru aceeași funcție sintactică, indiferent de regentul ei, așadar am folosit relația *pred*;
- în funcție de valoarea sa, reflexivă sau impersonală, în iulaLSPdep se folosesc două relații pentru pronumele reflexiv: *MPRON* și *MIMPERS*. Ambele cazuri, precum și utilizarea cu voloare reciprocă a aceluiași pronume, sunt acoperite de relația *reflclitic* în limba română.

2.2.3. 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 românești.

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 și 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 și conjuncția care însoțește fenomenul de dublare a subiectului;
2. altele adnotează fenomene morfologice: *auxpass* – auxiliarul de pasiv; *mwe* – termeni multicuvânt; *name* – nume de persoane și 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 greșit separate în text; *list* – pentru liste de elemente de același fel (de ex., adrese, numere de telefon etc.); *discourse* – în special pentru interjecții și cuvinte de umplutură (Ăăăă..., păi); *parataxis* –

pentru împletirea vorbirii directe cu intervențiile naratorului, pentru propoziții incidente; *reparandum* – pentru disfluențe în vorbirea directă; *foreign* – pentru secvențe de cuvinte străine.

Înainte de etapa de corectare, am transferat automat din setul de etichete sintactice IULA în setul nostru de etichete de dependențe tot ce s-a putut transfera ne-ambiguu. Etichete precum *spec* sau *mod* (cu mai mult de o etichetă echivalentă în setul românesc) au fost lăsate spre dezambiguizare în etapa de corectare. Pentru unele dintre etichete, transferarea nu presupune nimic mai mult decât scrierea cu minuscule în loc de majuscule: în setul nostru, singura etichetă scrisă cu majuscule este cea care marchează rădăcina, *ROOT*.

IULAdep	**ROdep**
SUBJ	subj
DO	dobj
IO	iobj
OBLC	pobj
BYAG	agc
ATR	pred
PRD	pred
OPRD	spe
PP-LOC	pmod
PP-DIR	pmod
VOC	voc
COMP	prep
NEG	neg
COORD	conj
CONJ	cc
PUNCT	punct
unknown	dep
AUX	aux

Tabelul 2.2. Corespondența între etichetele din iulaLSPdep și etichetele din ROdep care au fost transferate automat

Într-o etapă ulterioară finalizării proiectului, ne propunem să realizăm o variantă a resursei noastre complet aliniată la UD, alăturându-ne inițiativei de standardizare și deschizând calea pentru colaborări croslingvistice. În acest scop, vom renunța la unele dintre deciziile de adnotare luate pentru a fi în spiritul gramaticii limbii române:

1) Vom rafina relația *subj*, căreia îi corespund în UD, așa cum am menționat anterior, relațiile: *nsubj*, *nsubjpass*, *csubj* și *csubjpass*;

   Exemple:

   El citește.

   (el, citește, *nsubj*)

   Cântecul a fost compus de interpret.

(cântecul, compus, *nsubjpass*)

Cine aleargă după doi iepuri nu prinde nici unul.

(aleargă, prinde, *csubj*)

(cine, aleargă, *nsubj*)

Cine a încălcat legea a fost pedepsit de instanţă.

(încălcat, pedepsit, *cubjpass*)

(cine, încălcat, *nsubj*)

2) Vom adopta tratamentul prepoziţiilor şi elementelor subordonatoare din UD, adică acestea nu vor mai fi centru de grup sintactic, ci elemente dependente de cuvintele conţinut;
3) Vom adopta tratamentul obiectului secundar din UD: aici, în cazul în care un verb are două argumente în acuzativ, cel animat este legat de verb ca obiect indirect, iar cel inanimat ca obiect direct;

   Bunica i-a învăţat pe copii o poezie.

   (pe, învăţat, *iobj*)

   (copii, pe, *prep*)

   (poezie, învăţat, *dobj*)
4) Vom adopta tratamentul verbelor copulative din UD:
   - Relaţia dintre numele predicativ şi verbul copulativ a fi este *cop*, cu numele predicativ ca centru al relaţiei, doar în cazul verbului "a fi"; numele predicativ devine astfel *ROOT*, dacă se află în propoziția principală, sau, dacă se găsește într-o subordonată, devine centrul acesteia.

   Exemplu:

   Maria este fericită.

   (este, fericită, *cop*)

   (fericită, *, *ROOT*)

   - Toate celelalte verbe copulative sunt centre pentru numele predicative, care devin *xcomp* pentru verb;

   Exemplu:

   Maria a devenit ingineră.

   (ingineră, devenit, *xcomp*)

   - Subordonata predicativă a verbului copulativ „a fi" este analizată ca *ccomp*, aceasta fiind singura situaţie în care copulativul „a fi" este centru.

   Exemplu:

Noi suntem cum ne ştii.

(suntem, *, *ROOT*)

(ştii, suntem, *ccomp*)

O descriere a fenomenelor sintactice româneşti în formalismul UD a fost încărcată pe site-ul iniţiativei de standardizare şi poate fi studiată la adresa: http://universaldependencies.github.io/docs/ro/dep/index.html.

2.2.4. Descrierea detaliată a setului final de etichete ROdep și a principiilor de adnotare

În continuare vom detalia şi exemplifica anumite aspecte ale setului final de etichete ROdep. Pentru facilitarea unei adnotări consecvente, a fost redactat un ghid de adnotare care şi-a propus să ilustreze cu exemple fiecare tip de instanţiere (în funcţie de partea de vorbire) a fiecărui tip de relaţie de dependenţă. Acest ghid este reprodus în Anexa 1 şi este foarte util pentru înţelegerea principiilor de adnotare, completând secţiunea de faţă.

2.2.4.1. Rădăcina

Rădăcina unei propoziţii poate fi: un verb, un adjectiv sau o interjecţie. În absenţa unei părţi de vorbire de acest tip, rădăcina va fi centrul grupului sintactic dominant. Rădăcina intră în relația *ROOT* cu un nod artificial în arbore, care în resursa noastră conține întreaga propoziție analizată.

2.2.4.2. Legarea propozițiilor în frază

La nivel de frază, propoziţiile sunt separate fie prin punctuaţie (virgulă, punct şi virgulă, două puncte, etc.), fie prin conjucţii sau elemente relative.

Conjuncțiile pot fi coordonatoare sau subordonatoare.

- **Coordonarea**

Primul conjunct este centru pentru toţi ceilalţi conjuncţi, ca şi pentru conjuncție. Conjucţia coordonatoare (*și, sau* etc.) intră în relaţia *cc* cu primul conjuct. Conjuncţii stabilesc relaţia *conj* cu primul element al coordonării în ordine lineară.

Exemplu: Ion, Maria și Iulia

(Maria, Ion, *conj*)

(și, Ion, *cc*)

(Iulia, Ion, *conj*)

O conjuncție coordonatoare poate să apară și la începutul unei propoziții. Aceasta este etichetată tot *cc* și depinde de rădăcina propoziției. De fapt, este vorba de o coordonare care se extinde pe mai multe propoziții. Nu putem atașa conjuncția primului conjunct, deoarece se

găsește în altă propoziție, așa că o atașăm primului conjunct disponibil în propoziția curentă: predicatul.

Exemplu:

Și au salutat gazda.

(şi, salutat, *cc*)

Propoziţiile coordonate sunt tratate ca orice alte elemente aflate în raport de coordonare.

Exemplu:

Ion a sosit, dar Maria întârzie.

(dar, sosit, *cc*)

(întârzie, sosit, *conj*)

- **Subordonarea**

Conjuncţiile subordonatoare (al căror unic rol în fraze este acela de a marca relaţia de subordonare) sunt centre ale propoziţiilor subordonate pe care le introduc şi stabilesc relaţiile potrivite (în funcţie de tipul de subordonată pe care o introduc) în propoziţia principală. Ele intră în relaţia *sc* cu centrul verbal din propoziţia subordonată.

Pronumele, adjectivele şi adverbele relative pot apărea în:

- întrebări directe, când nu au un rol subordonator:

  Exemplu:

  Cine vine?

  (cine, vine, *subj*)

- întrebări indirecte şi propoziţii subordonate relative (obligatorii sau facultative în frază), unde au un rol dublu: pe de o parte marchează relaţia de subordonare, pe de altă parte sunt fie argumente ale verbului fie adjuncţi în propoziţia subordonată.

  Exemplu:

  Aș vrea să știu cine vine.

  (vine, știu, *dobj*)

  (cine, vine, *subj*)

Datorită celui de-al doilea rol, am decis să nu le tratăm ca elemente subordonatoare. Relativele vor intra într-o relaţie de dependenţă în interiorul subordonatei pe care o introduc, în timp ce verbul din subordonata respectivă va intra în relaţie directă de dependenţă cu elementul regent.

2.2.4.3. Tratamentul complexului verbal

- Auxiliariale sunt dependente de verbul ne-predicativ şi legate de acesta prin relaţia *aux*.

- Auxiliarul pasiv („a fi”) intră în relaţia *auxpass* cu verbul ne-predicativ.
- Prepoziţia care marchează forma de infinitiv a verbului (“a”) intră în relaţia *mark* cu acesta.
- Conjuncţia specifică modului subjonctiv, “să”, intră în relaţia *mark* cu verbul numai dacă apare în propoziţia principală (Ex.: “Să mâncăm împreună!”). Pentru rolul său de conjuncţie subordonatoare vezi secţiunea precedentă Subordonarea.
- Negaţia **(**“nu”**)** este ataşată de verb prin relaţia *neg*.
- Cliticul pronominal de dublare este ataşat centrului prin relaţia *dblclitic*.
- Cliticul reflexiv este ataşat verbului prin relaţia *reflclitic*.
- Cliticele reflexive folosite în construcţii pasive sunt ataşate centrului verbal prin relaţia *passmark*.
- Cliticele cu semnificaţie posesivă sunt ataşate centrului verbal prin relaţia *possclitic* dacă şi numai dacă complementul posesiv este deja ataşat verbului; altfel, cliticul este legat de verb prin relaţia *poss*.
- Vocativele sunt ataşate verbului prin relaţia *voc*.

2.2.4.4. Structura argumentală a centrului verbal

Relaţiile de dependenţă din structura argumentală a centrului verbal sunt:

- **subiectul** *subj* – poate fi exprimat prin substantiv, pronume, numeral, adverb, propoziţie subordonată (cu verb la un mod predicativ sau nepredicativ).
- **obiectul direct** *dobj* – poate fi exprimat prin substantiv, pronume, numeral, propoziţie subordonată (cu verb la un mod predicativ sau ne-predicativ). Când obiectul direct este definit, este realizat printr-un grup prepoziţional ce are ca centru prepoziţia “pe”. În acest caz, relaţia dintre verb şi prepoziţie este *dobj*, iar substantivul, pronumele sau numeralul intră în relaţia *prep* cu prepoziţia. Pentru tratamentul cliticului de dublare al obiectului direct, vedeţi mai sus.
- **obiectul indirect** *iobj* – poate fi exprimat prin substantiv, pronume, numeral, propoziţie subordonată (cu verb la un mod predicativ sau ne-predicativ). Uneori obiectul indirect se realizează printr-un grup prepoziţional guvernat de prepoziţia “la” sau “pentru”. În acest caz, relaţia dintre verb şi prepoziţie este *iobj*, iar substantivul, pronumele sau numeralul intră în relaţia *prep* cu prepoziţia. Pentru tratamentul cliticului de dublare al obiectului indirect, vedeţi mai sus.
- **obiectul secundar** *secobj* – când un verb are două argumente în cazul acuzativ, cel animat (identificat prin prepoziția “pe” și/sau dublarea prin clitic) este obiect

direct, iar cel inanimat este obiect secundar. Obiectul secundar poate fi realizat printr-un substantiv, un pronume sau un numeral.

- **obiectul prepozițional** *pobj* – anumite verbe, adjective, substantive și interjecții selectează un argument prepozițional; prepoziția intră în relația *pobj* cu centrul care o selectează iar complementul prepoziției intră în relația prep cu prepoziția.

  Exemplu:

  Mă tem de împrejurări.

  (de, tem, *pobj*)

  (împrejurări, de, *prep*)

  Atunci când argumentul prepozițional este realizat printr-o propoziție subordonată, aceasta poate fi:

  - O subordonată relativă, când prepoziția încă apare:

    Exemplu:

    Mă tem de ce poți face.

    (de, tem, *pobj*)

    (poți, de, *prep*)

  - O subordonată introdusă de o conjuncție subordonatoare, când prepoziția dispare:

    Exemplu:

    Mă tem că nu aud bine.

    (că, tem, *pobj*)

    (aud, că, *sc*)

- **complementul de agent** *agc* – în limba română, acesta este un grup sintactic prepozițional, guvernat de prepozițiile "de" sau "de către". Prepoziția stabilește relația *agc* cu centrul (un verb pasiv sau un adjectiv), în timp ce substantivul sau pronumele care o urmează intră în relația *prep* cu prepoziția.
- **predicativul** *pred* – aceasta este relația care se stabilește între verbul copulativ și cel de-al doilea argument al său (primul argument este subiectul). Cel de-al doilea argument poate fi un substantiv, un pronume, un numeral, un adjectiv, un adverb, o interjecție sau o propoziție subordonată (cu verb la un mod predicativ sau nepredicativ).
- **complementul posesiv** *poss* – acesta este realizat printr-un substantiv sau pronume care desemnează posesorul obiectului indicat de obiectul direct al

aceluiași verb. Când este exprimat prin substantiv, complexul verbal include și cliticul de dublare.

- **elementul predicativ suplimentar** *spe* – intră într-o relație ternară, cu verbul și cu un nominal din cadrul acestuia de subcategorizare; poate fi exprimat prin substantiv, pronume, adjectiv, numeral, adverb sau o propoziție subordonată (cu verb la un mod predicativ sau ne-predicativ).

  Exemplu:

  Femeia se arată simpatică.

  (simpatică, arată, *spe*)

2.2.4.5. Dependenții opționali ai verbului

Aceştia pot fi:

- **modificatorul adverbial** *advmod* – realizat prin adverbe
- **modificatorul nominal** *nmod* – realizat prin substantive sau pronume
- **modificatorul prepozițional** *pmod* – realizat printr-un grup prepozițional
- **adjectiv, centru al unei subordonate circumstanţiale** *xcomp* – realizat printr-un adjectiv care este centrul unui grup sintactic ce funcționează ca un dependent opțional al verbului (de obicei un circumstanţial de cauză)

  Exemplu:

  Bucuros că a câștigat, și-a invitat prietenii la o cină.

  (Bucuros, invitat, *xcomp*)
- **subordonata circumstanţială** *advcl* – este realizată printr-o propoziție subordonată care exprimă un loc, timpul, o condiție, etc.

2.2.4.6. Tratamentul grupului nominal

Grupul nominal poate avea ca centru un substantiv, un pronume sau un numeral care funcționează ca un substantiv. Relațiile pe care modificatorii le pot stabili cu aceste tipuri de centru sunt:

- **determinatori** *det* – realizat prin articole hotărâte, nehotărâte, demonstrative, posesive/genitivale.
- **modificatorul adjectival** *amod* – realizat prin adjectiv.
- **modificatorul nominal** *nmod* – realizat printr-un substantiv sau pronume.
- **subordonată adjectivală** *acl* – realizată printr-un gerunziu, un participiu, o subordonată relativă (obligatorie) sau o subordonată introdusă de o conjuncţie subordonatoare.
- **modificator adverbial** *advmod* – realizat printr-un adverb.
- **modificator prepozițional** *pmod* – realizat printr-un grup prepozițional.

- **negaţie de constituent** *neg*.

  Exemplu :

  Maria a cumpărat nu trandafiri, ci lalele.

  (nu, trandafiri, *neg*)

2.2.4.7. Tratamentul grupului adjectival

Grupul adjectival poate avea ca centru un adjectiv sau un numeral care funcționează ca adjectiv. Relațiile pe care modificatorii le pot stabili cu aceste tipuri de centru sunt:

- **modificator adverbial** *advmod* – realizat printr-un adverb.
- **subordonată circumstanţială** *advcl* – realizat printr-o subordonată care exprimă cauza, elementul de comparație etc.
- **modificatorul prepozițional** *pmod* – realizat printr-un grup prepozițional.
- **obiectul indirect** *iobj* – realizat printr-un substantiv sau pronume; exprimă beneficiarul sau cel care experimentează starea exprimată de adjectiv).
- **obiectul prepozițional** *pobj* – realizat printr-un grup prepozițional obligatoriu, i.e. absența sa din enunț generează o structură negramaticală.

  Exemplu:

  persoane dependente de droguri

  (de, dependente, *pobj*), (droguri, de, *prep*)
- **complement de agent** *agc* – realizat ca grup prepozițional.

  Exemplu:

  lucru inacceptabil de nimeni

  (de, inacceptabil, *agc*),

  (nimeni, de, *prep*)
- **negaţie** de constituent *neg*.

2.2.4.8. Numeralele

Numeralele se comportă fie ca substantive, fie ca adjective și pot intra în relațiile enumerate pentru grupul nominal și, respectiv, grupul adjectival.

2.2.4.9. Adverbele

Pot fi rădăcini sau modificatori. Astfel, ele intră în relații cu

- un **subiect** *subj* – exprimat printr-un nominal.
- un **obiect indirect** *iobj* – realizat printr-un nominal în cazul dativ; aceasta este o restricție de selecție a anumitor adverbe.

Exemplu:

A procedat adecvat situaţiei.

(situaţiei, adecvat, *iobj*)

- un **obiect prepoziţional** *pobj* – acesta este realizat ca grup prepoziţional cu prepoziția selectată de adverb.
- o **subordonată circumstanţială** *advcl* – aceasta poate exprima locul, timpul, condiția, etc.
- **un modificator prepoziţional** *pmod* – exprimat printr-un grup prepoziţional.
- **un modificator adverbial** *advmod* – realizat printr-un adverb.

2.2.4.10. Prepoziţiile

Prepoziţiile sunt centrele grupurilor prepoziţionale

- Când grupurile prepoziționale sunt dependenți opționali ai adjectivelor, adverbelor, substantivele sau verbelor (predicative sau ne-predicative), prepoziția stabilește relația *pmod* cu aceste centre.
- Când centrul (care poate fi un adjectiv, adverb, interjecție sau verb) selectează un grup prepozițional ca argument, prepoziția intră în relația *pobj* cu acel centru.
- Relația dintre prepoziție și complementul acesteia este întotdeauna *prep*, indiferent de ce tip de modificator (opțional sau obligatoriu) este grupul prepozițional.

2.2.4.11. Interjecţiile

Interjecţiile pot fi rădăcini, moştenind astfel toate posibilităţile de combinare ale verbului:

- **subiect** *subj* – realizat printr-un grup nominal.
- **obiect direct** *dobj* – realizat printr-un grup nominal.
- **obiect** *iobj* – realizat printr-un grup nominal.
- **complement posesiv** *poss* **–** acesta este un substantiv sau un pronume care desemnează posesorul obiectului indicat de obiectul direct al interjecţiei.
- **obiectul prepoziţional** *pobj* – anumite interjecţii selecteză argumente prepoziţionale.
- **modificator adverbial** *advmod* **–** realizat printr-un adverb.
- **modificator prepoziţional** *pmod* – realizat printr-un grup prepoziţional.
- **subordonată adverbială** *advcl*.

2.2.4.12. Apoziţiile

Toate părţile de vorbire conţinut pot avea apoziţii. Relaţia *appos* se stabileşte între un cuvânt din propoziţia principală şi centrul grupului apozitiv. Apozemele, i.e. cuvintele care

introduc apoziţii (*adică, anume, cu_alte_cuvinte, altfel_spus, mai_bine_zis, mai_exact, respectiv, şi_anume, alias, sau*), sunt marcatori de apoziţii şi se leagă de centrul apoziţiei prin relaţia *mark*.

Exemplu:

Munceşte delocalizat, adică unde i se cere.

(cere, delocalizat, *appos*)

(adică, cere, *mark*)

Relația *appos* este folosită şi pentru a lega perechi atribut-valoare în adrese, semnături, etc:

Exemplu:

Ana Ionescu, Str Rozelor, tel: 0245.756.547, email: ana@yahoo.com

(Str, Ana, *list*)

(Rozelor, Str., *appos*)

(tel, Ana, list)

(0245.756.547, tel, *appos*)

(email, Ana, list)

(ana@yahoo.com, email, *appos*)

2.2.4.13 Structurile eliptice

Când o elipsă apare într-o secvenţă de grupuri sintactice similare, folosim relaţia *remnant*. Pentru celelalte tipuri de elipse, centrul grupului sintactic lexicalizat devine rădăcina propoziţiei.

Exemplu:

Maria a mers la$_1$ Berlin, Elena la$_2$ Barcelona.

(Maria, mers, *subj*)

(la$_1$, mers, *pmod*)

(Berlin, la$_1$, *prep*)

(Elena, Maria, *remnant*)

(la$_2$, la$_1$, *remnant*)

(Barcelona, la$_2$, *prep*)

2.2.4.14. Alte tipuri de relaţii

- *foreign*: aceasta este o relaţie folosită pentru legarea între ele a cuvintelor din alte limbi. Pentru un şir de mai multe cuvinte dintr-o limbă străină, analiza se face astfel: primul

cuvânt din şir primeşte relaţia cerută de gramatica limbii române pentru poziţia ocupată de şir în propoziţie, iar celelalte cuvinte sunt legate prin relaţia *foreign* de primul cuvânt.

Exemplu:

A spus good bye și a plecat.

(good, spus, *dobj*)

(bye, good, *foreign*)

- *name***:** în cazul numelor compuse, toate elementele se leagă de primul, în ordine liniară, prin relaţia *name*.

  Exemplu:

  Numele meu este Ana Maria Ionescu.

  (Ana, este, *pred*)

  (Maria, Ana, *name*)

  (Ionescu, Ana, *name*)

- *mwe***:** am preluat această relaţie, folosită pentru legarea expresiilor multi-cuvânt, din UD, dar am adaptat-o pentru că am dorit să păstrăm informații despre structura sintactică a expresiilor. De asemenea, am folosit relația și pentru adnotarea entităților denumite (nume de instituții, de comisii, etc.). Fiecare element dintr-o expresie adnotată folosind mwe păstrează informația sintactică prin concatenarea relației de dependență din interiorul expresiei la relația mwe.

  Exemplu: Curtea Europeană a Drepturilor Omului

  (Europeană, Curtea, *mwe-amod*)

  (Drepturilor, Curtea, *mwe-nmod*)

  (a, Drepturilor, *mwe-det*)

  (Omului, Drepturilor, *mwe-nmod*)

- *list***:** această relaţie este folosită în liste precum adresele sau datele.

  Exemplu:

  miercuri, 29 decembrie

  (decembrie, miercuri, *list*)

  (29, decembrie, amod)

  Str. Popa Șapcă, nr. 35

  (nr., Str., *list*)

  (Popa, Str., *appos*)

  (Șapcă, Popa, *name*)

  (35, nr, *appos*)

- *parataxis***:** relaţia este utilizată pentru a lega centrul unui element de vorbire directă de centrul unui element de vorbire indirectă; am folosit această relație și pentru secvențe de cuvinte care funcționează ca etichete pentru întreaga propoziție: de exemplu, în secțiunea jurnalistică, multe dintre propoziții încep cu secvențe de cuvinte care reprezintă identificatori pentru secțiunile de lege conținute în propoziții.

  Exemplu:

  Ce faci? a întrebat el.

  (întrebat, faci, *parataxis*)

  (3) Prezentul acord poate fi modificat pe baza acordului în scris al părților.

  (3, poate, *parataxis*)

  (poate, *, *ROOT*)

- *goeswith***:** această relaţie leagă două părţi ale unui cuvânt care sunt separate în text datorită unei greşeli de editare sau segmentare. Centrul este primul dintre cele două elemente. Dacă fragmentarea prespune mai multe elemente separate, legarea se face în lanţ.

  Exemplu:

  Sunaţi la 0245 323 313.

  (0245, la, *prep*)

  (323, 0245, *goeswith*)

  (313, 323, *goeswith*)

- *discourse***:** această relaţie este folosită pentru interjecţii sau alte elemente şi particule de discurs, care nu sunt legate în mod direct de structura propoziţiei, ci aduc expresivitate enunţului (o!, aha, um, a, păi, de fapt, dar ştii, etc.). Centrul acestui tip de dependență va fi rădăcina propoziţiei.

  Exemplu:

  A, am uitat să cumpăr mere.

  (A, uitat, *discourse*)

  (“,”, A, *punct*)

- *dislocated***:** folosită pentru elemente ante-poziţionate sau post-poziţionate care dublează un argument al centrului propoziţiei. Elementul dislocat se ataşează aceluiaşi centru ca şi dependentul pe care îl dublează.

  Exemplu:

  Am băut-o, cafeaua.

  (-o, băut, *dobj*)

(cafeaua, băut, *dislocated*)

- *reparandum***:** folosim relaţia reparandum pentru tratamentul disfluenţelor în repararea vorbirii. Disfluenţa este dependentă de elementul de reparare.

  Exemplu:

  Mergi la_1 stân... la_2 dreapta.

  (la_2, mergi, *pmod*)

  (dreapta, la_2, *prep*)

  (la_1, la_2, *reparandum*)

  (stân, la_1, *prep*)

  (…, stân, *punct*)

CAPITOLUL 3

RESURSE ŞI INSTRUMENTE UTILIZATE

3.1. ROMBAC

Am selectat propoziţiile de adnotat din ROMBAC, un corpus balansat, care oferă avantajul că acoperă domenii şi stiluri literare diverse, permiţându-ne să imaginăm un treebank care să fie, de asemenea, balansat.

ROMBAC este distribuit prin platforma META-SHARE (dezvoltată de META-NET[12]) și este adnotat, conform recomandărilor acesteia, cu metainformații referitoare la numele resursei, numele autorilor resursei, detalii despre persoana de contact, despre condițiile de distribuție, despre dimensiunile resursei, codificarea datelor, tipurile de adnotare disponibile în corpus etc. Corpusul este disponibil pentru descărcare, cu îndeplinirea condițiilor de distribuție, la http://ws.racai.ro:9191/browse/, care este un punct de instanțiere locală al platformei MetaShare V1.1.

Corpusul acoperă patru stiluri funcţionale ale limbii (beletristic, oficial, publicistic şi ştiinţific) şi cuprinde cinci secţiuni, fiecare corespunzând unui domeniu distinct:

- sub-corpusul jurnalistic: provine din ziarul Agenda (http://www.agenda.ro/) şi conţine ştiri publicate între anii 2003 şi 2005, însumând 8.500.000 de cuvinte;
- sub-corpusul de ficţiune (literar): este o colecţie de romane şi poeme semnate de 28 de autori clasici români, care au publicat între sfârşitul secolului 19 şi începutul secolului 20; porţiuni din acest sub-corpus – care numără în total 6.800.000 de cuvinte – au fost iniţial redactate cu o ortografie românească veche, dar autorii ROMBAC au armonizat ortografia conform normelor actuale şi au unificat codificarea diacriticelor în text;
- sub-corpusul academic: numără 4.300.000 de cuvinte şi este bazat pe Dicţionarul General al Literaturii Române (Academia Română, 2009), o antologie critică ce cuprinde atât biografii ale unor scriitori, poeţi şi eseişti, cât şi comentarii critice despre operele acestora, definiţii ale unor concepte şi curente literare, etc;
- sub-corpusul medical: extras din corpusul multilingv paralel EMEA, compilat de Tiedemann (2009) din documente ce provin de la Agenţia Medicală Europeană. Din

[12] http://www.meta-net.eu/

componenta tradusă în română, descărcată de la http://opus.lingfil.uu.se/EMEA.php, au fost selectate aleatoriu 800 de documente numărând 9.100.000 de cuvinte;

- sub-corpusul juridic: extras din corpusul JRC-Acquis (corpus paralel disponibil în 22 limbi, (Steinberger et al., 2006)), bazat pe Acquis-ul Comunitar, o colecţie de texte legislative ale Uniunii Europene aplicabile în toate statele membre.

Corpusul a trecut iniţial printr-o etapă de pre-procesare, care presupune curăţarea datelor, uniformizarea diacriticelor şi codificarea UTF-8 a documentelor. Ulterior, a fost segmentat la nivel de propoziţie şi la nivel de cuvânt, adnotat morfo-lexical (eng. POS tagging) şi lematizat cu platforma de procesare de text TTL, dezvoltată la ICIA (Ion, 2007; Tufiș et al., 2008) şi disponibilă ca serviciu web la http://ws.racai.ro/ttlws.wsdl.

Componenta de adnotare morfo-lexicală a TTL este bazată pe Modele Markov Ascunse şi are o acurateţe de peste 98%. Setul de etichete utilizat cuprinde 614 descriptori morfo-lexicali (MSD-uri, din engl. Morpho-Syntactic Descriptor) şi este compatibil cu specificaţiile MULTEXT-East[13].

Componenta de lematizare foloseşte informaţia morfo-lexicală produsă la pasul anterior şi are trei scenarii posibile: 1) forma cuvântului plus eticheta pot identifica complet lema printr-o procedură de căutare într-un lexicon de mari dimensiuni (1.200.000 de intrări), validat manual; 2) forma cuvântului plus eticheta nu identifică lema în mod unic în lexicon, caz în care se optează pentru lema cea mai frecventă dintre cele posibile; 3) forma cuvântului plus eticheta nu produc nici un rezultat la căutarea în lexicon, caz în care se foloseşte un algoritm de predicţie bazat pe un Model Markov de sufixe de 5 litere, antrenat pe leme corecte din lexicon care au acelaşi MSD cu cuvântul ce trebuie lematizat; acurateţea algoritmului de predicţie este de 83%.

Informaţia returnată de lanţul de procesare TTL este codificată într-un format XML ne-standard, dar ulterior este convertită într-un format XCES (revizia 1.0.4, schema disponibilă la http://www.xces.org/schema/2003/) compatibil platformei METANET.

În Figurile 3.1 şi 3.2 puteți vedea un exemplu de propoziție din ROMBAC adnotată cu TTL. Figura 3.1 prezintă secţiunea de meta-informaţii a documentului *A01-05-Actualitate*, grupată sub eticheta *xces:cesHeader*. Figura 3.2 prezintă o propoziţie din același document. Fiecare cuvânt este codificat de un element *xces:tok*, ale cărui atribute *base* şi *msd* au ca valoare lema, respectiv MSD-ul cuvântului.

[13] http://nl.ijs.si/ME/V3/msd/html/msd.html

```
<?xml version="1.0" encoding="UTF-8"?>
<xces:cesCorpus xmlns:xces="http://www.xces.org/schema/2003" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://www.xces.org/schem
    <xces:cesHeader version="0.1" date.created="Wed Nov 9 16:24:00 2011" creator="TTL" type="automatic preprocessing">
        <xces:fileDesc>

            <xces:titleStmt />
            <xces:extent>
                <xces:wordCount>1649</xces:wordCount>
                <xces:byteCount units="bytes">31744</xces:byteCount>
            </xces:extent>
            <xces:sourceDesc>
                <xces:biblFull>

            <xces:titleStmt />
                <xces:publicationStmt>
                    <xces:distributor>Research Institute for Artificial Intelligence, Romanian Academy</xces:distributor>
                    <xces:pubAddress>Casa Academiei, Calea 13 Septembrie nr. 13, etaj 3, Bucureşti Sector 5, România, 050711</xces:pubAddress>
                    <xces:telephone>0040 21 3188103</xces:telephone>
                    <xces:fax>0040 21 3188142</xces:fax>
                    <xces:eAddress>tufis@racai.ro</xces:eAddress>
                </xces:publicationStmt>
                </xces:biblFull>
            </xces:sourceDesc>
        </xces:fileDesc>
        <xces:profileDesc>
            <xces:langUsage>
                <xces:language iso639="ro">Română</xces:language>
            </xces:langUsage>
        </xces:profileDesc>
    </xces:cesHeader>
```

Figura 3.1*:* Reprezentarea metainformaţiilor asociate unui document din ROMBAC în format XCES. În imagine se pot regăsi informaţii despre instrumentul care a produs documentul (TTL, prin preprocesare automată), despre dimensiunile documentului în număr de cuvinte şi număr de octeţi, despre limba în care este redactat documentul, precum şi detalii despre distribuitorul documentului.

```
<?xml version="1.0" encoding="UTF-8"?>
<xces:cesCorpus xmlns:xces="http://www.xces.org/schema/2003" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://www.xces.org/schema/
    <xces:cesDoc version="0.1">
        <xces:text id="A01-05-actualitate" complete="y">
            <xces:body>
                <xces:p>
                    <xces:s>
                        <xces:tok base="aşa" msd="Rgp" type="word">Aşa</xces:tok>
                        <xces:tok base="după_cum" msd="Cscsp" type="word">după_cum</xces:tok>
                        <xces:tok base="avea" msd="Va--3s" type="word">a</xces:tok>
                        <xces:tok base="promite" msd="Vmp--sm" type="word">promis</xces:tok>
                        <xces:tok base="," msd="COMMA" type="word">,</xces:tok>
                        <xces:tok base="premier" msd="Ncmsry" type="word">premierul</xces:tok>
                        <xces:tok base="desemnat" msd="Afpms-n" type="word">desemnat</xces:tok>
                        <xces:tok base="," msd="COMMA" type="word">,</xces:tok>
                        <xces:tok base="domnul_Călin_Popescu_Tăriceanu" msd="Edpm" type="word">domnul_Călin_Popescu_Tăriceanu</xces:tok>
                        <xces:tok base="," msd="COMMA" type="word">,</xces:tok>
                        <xces:tok base="avea" msd="Va--3s" type="word">a</xces:tok>
                        <xces:tok base="anunţa" msd="Vmp--sm" type="word">anunţat</xces:tok>
                        <xces:tok base="duminică" msd="Ncfsrn" type="word">duminică</xces:tok>
                        <xces:tok base="," msd="COMMA" type="word">,</xces:tok>
                        <xces:tok base="26" msd="Eni" type="word">26</xces:tok>
                        <xces:tok base="decembrie" msd="Ncms-n" type="word">decembrie</xces:tok>
                        <xces:tok base="," msd="COMMA" type="word">,</xces:tok>
                        <xces:tok base="structură" msd="Ncfsry" type="word">structura</xces:tok>
                        <xces:tok base="nou" msd="Afpmsoy" type="word">noului</xces:tok>
                        <xces:tok base="cabinet" msd="Ncms-n" type="word">cabinet</xces:tok>
                        <xces:tok base=":" msd="COLON" type="word">:</xces:tok>
                    </xces:s>
                </xces:p>
            </xces:body>
        </xces:text>
    </xces:cesDoc>
</xces:cesCorpus>
```

Figura 3.2. Reprezentarea unei propoziţii adnotate cu TTL din documentul ale cărui metadate sunt reproduse în Figura 3.1. De exemplu, pentru cuvântul “Aşa”, TTL a returnat lema “aşa” şi MSD-ul Rgp, corespunzător unui adverb de tip general şi grad pozitiv.

3.2. IULA LSP

Treebank-ul de dependențe IULA Spanish LSP, pe care a fost antrenat modelul statistic de limbă spaniolă utilizat de noi, este un corpus tehnic, care numără 40.000 de propoziții (550.000 de cuvinte) și este disponibil gratuit, ca şi ROMBAC, prin platforma META-SHARE[14] printr-o licență Creative Commons. Corpusul original pe care se bazează acest treebank, *Corpus Técnic de l'IULA*, cuprinde texte scrise din domeniile: juridic, economie, știința calculatoarelor, mediu și medicină, provenind din publicaţii specializate, teze de doctorat, etc. Propoziţiile selectate pentru acest treebank sunt reprezentative pentru corpusul original, atât ca număr de propoziţii pe domeniu cât şi în ceea ce priveşte lungimea propoziţiilor, rezultând o resursă balansată. Corpusul, codificat UTF-8, a fost adnotat morfo-lexical cu adnotatorul Freeling (Padró et al., 2010), folosind un set[15] de etichete bazat, ca şi setul MSD folosit în ROMBAC, pe specificaţiile EAGLES[16]. Accurateţea adnotării morfo-lexicale depăşeşte 98%.

Adnotarea cu relaţii sintactice de dependenţe se face în doi paşi: iniţial, s-a folosit mediul de procesare DELPH-IN (eng. Deep Linguistic Processing with HPSG Initiative) şi gramatica de tip HPSG Spanish Resource Grammar (Marimon, 2013) pentru analiza propoziţiilor. S-a folosit un algoritm stocastic de tip MaxEnt pentru ordonarea arborilor produşi de gramatică şi reducerea la un număr de 500 cei mai buni arbori, dintre care s-a selectat manual analiza corectă. Rezultatele, reprezentate ca arbori de derivare, au fost convertite automat în arbori de dependenţe.

Treebank-ul este distribuit în formatul standardizat CONLL, lansat de competiţiile de analiză sintactică cu dependenţe menţionate în Secţiunea 2.1.8. Fiecare fişier în format CONLL conţine propoziţiile separate printr-un rând liber, în timp ce cuvintele din propoziţie se găsesc fiecare pe un rând nou. Fiecare cuvânt este descris prin 10 câmpuri (a căror semnificaţie este detaliată în Tabelul 3.1) separate printr-un caracter tab.

Numărul câmpului:	**Numele câmpului:**	**Descrierea:**
1	ID	Un contor de cuvânt, care începe de la 1 pentru fiecare propoziţie nouă
2	FORM	Forma cuvântului sau un simbol de punctuaţie
3	LEMMA	Lema

[14] http://metashare.upf.edu și http://hdl.handle.net/10230/20408.

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

[16] http://www.ilc.cnr.it/EAGLES96/browse.html

4	CPOSTAG	Etichetă morfo-lexicală nerafinată, indicând tipul de parte de vorbire reprezentat de cuvânt
5	POSTAG	Etichetă morfo-lexicală rafinată
6	FEATS	Un set neordonat de trăsături sintactice şi/sau morfologice, separate printr-o bară verticală, sau o liniuţă de subliniere (eng. underscore) dacă informaţia nu este disponibilă
7	HEAD	Centrul cuvântului curent, care este fie o valoare a câmpului ID fie zero
8	DEPREL	Tipul relaţiei care leagă cuvântul curent de centru
9	PHEAD	Centrul proiectiv al cuvântului curent, care este fie o valoare a câmpului ID, fie zero, fie o liniuţă de subliniere dacă informaţia nu este disponibilă
10	PDEPREL	Tipul relaţiei care leagă cuvântul curent de centrul proiectiv sau o liniuţă de subliniere dacă informaţia nu este disponibilă

Tabelul 3.1. Descrierea semnificaţiilor fiecărui câmp din formatul CONLL.

Tabelul 3.2 prezintă adnotarea sintactică şi morfo-lexicală pentru propoziţia “Transforma el clima exterior en un ambiente controlable y confortable.”/ ro. ”Transformă clima exterioară într-un mediu controlabil şi confortabil.” în format CONLL:

1	Transforma	transformar	v	VMM02S0	_	0	ROOT	0	_	_
2	el	el	d	DA0MS0	_	3	SPEC	3	_	_
3	clima	clima	n	NCMS000	_	1	DO	1	_	_
4	exterior	exterior	a	AQ0CS0	_	3	MOD	3	_	_
5	en	en	s	SPS00	_	1	OBLC	1	_	_
6	un	un	z	Z	_	7	SPEC	7	_	_
7	ambiente	ambiente	n	NCMS000	_	5	COMP	5	_	_
8	controlable	controlable	a	AQ0CS0	_	7	MOD	7	_	_
9	y	y	c	CC	_	8	COORD	8	_	_
10	confortable	confortable	a	AQ0CS0	_	9	CONJ	9	_	_
11	.	.	f	Fp	_	10	punct	10	_	_

Tabelul 3.2 O propoziţie din corpusul IULA LSP adnotată sintactic cu dependenţe.

În proiectul nostru, am utilizat un model statistic de analiză sintactică antrenat cu MaltParser (secțiunea 3.3) în mod de-lexicalizat (bazat doar pe trăsături de tip etichetă morfo-lexicală sau etichetă sintactică, detaltii în secțiunea 3.3.2) pe corpusul IULA LSP. Modelul este același cu cel utilizat de Arias et al. (2014) și a fost obținut de la echipa IULA cu prilejul stagiului de mobilitate.

Pentru a putea adnota corpusul românesc de 5.000 de propoziţii dezvoltat pe baza ROMBAC (ce va fi descris în Secţiunea 4.1.) cu analizorul statistic MaltParser antrenat pe IULA LSP, este nevoie ca cele două seturi de etichete morfologice din cele două corpusuri să fie armonizate: mai precis, setul de etichete MSD din corpusul românesc de adnotat trebuie să fie transformat în setul de etichete specific IULA. Cum ambele seturi sunt derivate din specificaţiile EAGLES, nu a fost dificil de realizat o corespondenţă între cele două seturi, pe care am folosit-o pentru a înlocui automat MSD-urile cu etichetele IULA LSP corespunzătoare. Pentru lista detaliată de corespondenţe, vedeţi Anexa 2.

3.3. MaltParser

MaltParser este de fapt un generator de parsere: pornind de la un treebank adnotat cu depedenţe sintactice într-o anumită limbă, poate fi folosit pentru inducerea unui parser pentru acea limbă. Este disponibil liber pentru cercetare şi scopuri educaţionale şi a fost evaluat empiric pe mai multe limbi, printre care şi engleza şi spaniola. MaltParser implementează analiza cu dependenţe inductivă (Nivre, 2005), care apelează la tehnici de învăţare automată inductivă pentru ghidarea parserului în alegeri ne-deterministe. Cele trei componente principale ale acestei metodologii sunt:

3.3.1. Algoritmi determinişti pentru construirea grafurilor de dependenţe

Nivre et al. (2006) definesc structurile de date pe care se bazează un algoritm determinist compatibil arhitecturii MaltParser după cum urmează:

- STACK: este o stivă de elemente (cuvinte) parţial procesate; STACK[i] este elementul aflat în poziţia i+1 faţă de vârful stivei, care este STACK[0];
- INPUT: este o listă de elemente ne-procesate (cuvinte din propoziţie); INPUT[i] este elementul i+1 din listă, al cărei prim element este INPUT[0];
- CONTEXT: este o stivă de elemente ne-ataşate care apar în propoziţie între vârful lui STACK şi următorul element din INPUT; vârful acestei stive, CONTEXT[0] este cel mai apropiat element de STACK[0] şi cel mai depărtat de INPUT[0]:
- HEAD: este o funcţie care defineşte structura de dependenţe parţială deja construită, unde HEAD[i] este centrul sintactic al elementului i; HEAD[i]=0 dacă centrul lui i nu a fost încă identificat;
- DEP: este o funcţie care etichetează structura parţială de dependendenţe; DEP[i] este tipul dependenţei care leagă elementul i de centrul său sintactic, HEAD[i]; DEP[i] = ROOT dacă lui i nu i s-a ataşat încă un centru.

- LC: este o funcţie care defineşte „copilul cel mai la stânga” (eng. leftmost child) al unui element în structura de dependenţe parţială; LC[i] = 0 dacă i nu are copii la stânga.
- RC: este o funcţie care defineşte „copilul cel mai la dreapta” (eng. rightmost child) al unui element în structura de dependenţe parţială; RC[i] = 0 dacă i nu are copii la dreapta.
- LS: este o funcţie care defineşte următorul frate la stânga în structura de dependenţe parţială; LS[i] =0 dacă i nu are fraţi la stânga;
- RS: este o funcţie care defineşte următorul frate la dreapta în structura de dependenţe parţială; RS[i] =0 dacă i nu are fraţi la dreapta;

Structurile de bază sunt STACK, INPUT, HEAD și DEP, ele definind o configurație pentru un graf de dependențe asociat unei anumite propoziții. Structura CONTEXT este folosită doar de algoritmi care funcționează în mod non-proiectiv, pentru stocarea elementelor ne-atașate care apar între STACK[0] și INPUT[0] (de la dreapta la stânga). Parserul este inițializat cu o stivă vidă, cu toate elementele propoziției în lista INPUT, și cu un graf de dependențe în care toate nodurile sunt dependente de ROOT și toate arcurile sunt etichetate cu eticheta ROOT (pentru oricare i, HEAD[i]=0 și DEP[i] = ROOT). La sfârșitul analizei, lista INPUT trebuie să fie vidă și a avut loc o trecere de la stânga la dreapta prin toate elementele propoziției.

Algoritmul folosește patru tipuri de tranziții pentru a construi graful final de dependențe, două dintre ele bazate pe tipuri de relații de dependențe posibile ($r \in R$, unde R este setul de relații de dependențe dintr-o gramatică):

- LEFT-ARC(r): face ca elementul i din vârful stivei să fie dependent (la stânga) de următorul element j din INPUT, cu tipul dependenței r; adică h[i]=j și d[i]=r; în algoritmul proiectiv, i este eliminat din stivă, deoarece în acest moment el trebuie să aibă toți dependenții asociați; această tranziție are loc doar dacă h[i]=0;
- RIGHT-ARC(r): face ca următorul element j din lista INPUT să devină dependent (la dreapta) de elementul i din vârful stivei cu tipul dependenței r și împinge pe j în stivă; adică h[j]=i și d[j]=r; această tranziție are loc doar dacă h[j]=0; în acest moment, j ar trebui să aibă toți dependenții la stânga asociați, dar mai poate primi dependenți la dreapta;
- REDUCE: elimină vârful stivei; se poate aplica doar dacă vârful stivei are deja centrul asociat; tranziția este necesară pentru a elemina un nod care a fost împins în stivă printr-o tranziție RIGHT-ARC și care și-a găsit între timp toți dependenții la dreapta;

- SHIFT: împinge în STACK următorul element din INPUT; se aplică atâta vreme cât există elemente în INPUT; este necesară pentru procesarea nodurilor care au centrul la dreapta lor și pentru atașarea nodului ce are ca centru nodul artificial rădăcină.

Sistemul de tranziții definit mai sus este în sine ne-deterministic, unei anumite configurații putându-i fi aplicate mai multe tranziții. Pentru prezicerea următoarei tranziții, MaltParser folosește modele de trăsături bazate pe istoric. Algoritmii determiniști pe care îi implementează MaltParser pentru construirea grafului de dependențe sunt algoritmul lui Nivre (2003) pentru structuri de dependenţe proiective şi algoritmul lui Convington (2001) care poate fi rulat atât în mod proiectiv cât şi non-proiectiv.

3.3.2. Modele de trăsături bazate pe istoric

Pentru ca algoritmul de analiză să fie determinist, sistemul de tranziții este suplementat cu un mecanism care prezice care este următoarea tranziție și alege tipul dependenței pentru tranzițiile LEFT-ARC(r) și RIGHT-ARC(r). Acest mecanism este un model de trăsături bazat pe istoricul analizei. Modelele de trăsături bazate pe istoric, introduse de Black et al. (1992), își propuneau să extindă contextul din care modelele probabilistice își adună informația, incorporând trăsături diverse din istoria discursului, trecând chiar peste limitele propoziției.

Modelul utilizat de MaltParser este definit pe cuvinte (prin trăsături de tip LEX), părți de vorbire (prin trăsături de tip POS) și tipuri de dependențe (prin trăsături de tip DEP), relativ la una dintre structurile de date STACK, INPUT sau CONTEXT, folosind funcțiile HEAD, LC, RC, LS și RS.

Modelul de trăsături este specificat extern parserului, urmând sintaxa de mai jos:

<fspec> ::= <feat>+
<feat> ::= <lfeat>|<nlfeat>
<lfeat> ::= LEX\t<dstruc>\t<off>\t<suff>\n
<nlfeat> ::= (POS|DEP)\t<dstruc>\t<off>\n
<dstruc> ::= (STACK|INPUT|CONTEXT)
<off> ::= <nnint>\t<int>\t<nnint>\t<int>\t<int>
<suff> ::= <nnint>
<int> ::= (...|-2|-1|0|1|2|...)
<nnint> ::= (0|1|2|...)

Pentru exemplificare, în Tabelul 3.3 redăm un model pe care Nivre și Hall (2005) îl consideră util pentru orice limbă adnotată. Modelul poate fi de asemenea optimizat individual

pentru fiecare limbă și set de date de antrenare. Capul tabelului descrie semnificația fiecărei informații în fiecare coloană a fișierului de trăsături. În fișierul original, informațiile redate în tabel sunt separate prin tab-uri ("\t"). Fiecare linie din tabelul de mai jos descrie o singură trăsătură din fișierul de trăsături.

Tipul trăsăturii	**Locația elementului la care se referă trăsătura**	**Indice**				
		Elementul i+1 din lista menționată în coloana 2	**Valoare pozitivă: deplasare înainte în șirul original INPUT Valoare negativă: deplasare înapoi în șirul original INPUT**	**i aplicări ale funcției HEAD**	**Valoare negativă: i aplicări ale funcției LC Valoare pozitivă: i aplicări ale funcției RC 0: nici o aplicare**	**Valoare negativă: i aplicări ale funcției LS Valoare pozitivă: i aplicări ale funcției RS 0: nici o aplicare**
POS	STACK	0	0	0	0	0
POS	INPUT	0	0	0	0	0
POS	INPUT	1	0	0	0	0
POS	INPUT	2	0	0	0	0
POS	INPUT	3	0	0	0	0
DEP	STACK	1	0	0	0	0
DEP	STACK	0	0	0	0	0
DEP	STACK	0	0	0	0	-1
DEP	STACK	0	0	0	0	1
DEP	INPUT	0	0	0	0	-1
LEX	STACK	0	0	0	0	0
LEX	INPUT	0	0	0	0	0
LEX	INPUT	1	0	0	0	0
LEX	STACK	0	0	1	0	0

Tabelul 3.3. Configurația standard a fișierului de trăsături, cunoscută și ca modelul 7, recomandată pentru orice limbă de analiză și set de date

În continuare, vom explica ce trăsătură introduce fiecare linie din Tabelul 3.3:

- Linia 1: partea de vorbire a primului element (0+1 în coloana 3) din stiva STACK, adică trăsătura POS a vârfului (TOP) stivei;
- Linia 2: partea de vorbire a următorului element (0+1 în coloana 3), adică NEXT, din lista INPUT;

- Linia 3: partea de vorbire a primului element de după următorul element (1+1 în coloana 3), adică primul element după NEXT, din lista INPUT;
- Linia 4: partea de vorbire a celui de-al doilea element de după următorul element (2+1 în coloana 3), adică al doilea element după NEXT, din lista INPUT;
- Linia 5: partea de vorbire a celui de-al treilea element de după următorul element (3+1 în coloana 3), adică al treilea element după NEXT, din lista INPUT;
- Linia 6: tipul dependenței elementului TOP din stiva STACK;
- Linia 7: tipul dependenței primului dependent la stânga (un pas în jos către dependentul la stânga: coloana 6 are valoarea -1) al lui TOP (STACK în coloana 2, 0 în coloana 3);
- Linia 8: tipul dependenței primului dependent la dreapta (un pas în jos către dependentul la dreapta: coloana 6 are valoarea 1) al lui TOP (STACK în coloana 2, 0 în coloana 3);
- Linia 9: tipul dependenței primului dependent la stânga (un pas în jos către dependentul la stânga: coloana 6 are valoarea -1) al lui NEXT (INPUT în coloana 2, 0 în coloana 3);
- Linia 10: forma cuvântului din poziția TOP în stivă;
- Linia 10: forma cuvântului din poziția NEXT în INPUT;
- Linia 11: forma cuvântului imediat vecin lui NEXT în INPUT;
- Linia 12: definește forma cuvântului care este centrul elementului TOP din stivă (se aplică funcția HEAD o dată, vezi coloana 5).

3.3.3. Învățare automată discriminativă pentru stabilirea corespondenței între istoric și acțiunile parserului

MaltParser implementează doi algoritmi de învățare, care induc o funcție de la istoricul parserului (relativ la modelul de trăsături dat) la acțiunile parserului (relativ la un algoritm de analiză dat):

- învățare și clasificare bazată pe memorie (eng. memory-based learning, MBL) (Daelemans și Van den Bosch, 2005): stochează toate instanțele de antrenare în momentul învățării și folosește o variantă a metodei de clasificare k-NN (eng. k-nearest neighbors, k cei mai apropiați vecini) pentru a prezice următoarea acțiune în momentul analizei; algoritmul este implementat cu pachetul software TiMBL; acesta este algoritmul de învățare folosit de cele mai multe experimente cu MaltParser până în prezent;

- mașinile vectori pentru suport (eng. support vector machines, SVM): implementat folosind librăria LIBSVM (Chang și Lin, 2001).

3.3.4. Rularea MaltParser

MaltParser are două moduri de rulare:

- modul ”learn” (“ro. învățare”), care are ca date de intrare un treebank cu dependențe și induce un clasificator pentru a prezice acțiunile parserului, dat fiind un anumit algoritm de analiză, un model de trăsături și un algoritm de învățare;
- modul “parse” (ro. “analiză”), care are ca date de intrare un set de propoziții și construiește un graf de dependențe pentru fiecare propoziție, folosind clasificatorul indus în modul “learn” și același algoritm de analiză și model de trăsături.

Formatul de intrare/ieșire este CONLL, descris în secțiunea 3.2. În modul “parse”, informația din coloanele corespunzătoare centrului și dependenței în formatul CONLL este ignorată, dar coloanele nu trebuie să lipsească: valoarea recomandată este “_”, folosită în formatul CONLL în cazul în care informația este absentă.

3.4. yEd

Ca mediu de corectare a erorilor de adnotare inerente procesului de adnotare automată cu MaltParser, am folosit instrumentul yEd[17], o aplicație foarte prietenoasă care facilitează crearea de diagrame, fie manual, element cu element, fie prin import din date externe în format xcel sau xml. Editarea diagramelor în yEd este intuitivă și confortabilă datorită diverselor sale funcționalități: instrumente de căutare și de selecție în diagramă, funcția de mărire și micșorare (zoom-in/zoom-out) asociată rotiței mouse-ului, facilități clipboard, revenire (undo) extensivă, posibilitatea de a lucra cu mai multe diagrame simultan, comenzi rapide de la tastatură, etc.

O facilitate importantă este cea legată de modul în care sunt prezentate vizual datele. yEd folosește algoritmi sofisticați pentru a aranja în mod automat diagramele în imagine sau pentru a asista utilizatorul să-și aranjeze singur diagrama. Utilizatorul poate alege între scheme de aranjare ierarhice, radiale, de tip arbore, circulare, organice, ortogonale etc. Am considerat că cea mai potrivită și ușor de utilizat schemă pentru reprezentarea arborilor de dependențe este cea ierarhică, cu orientare de jos în sus. În Figura 3.3. se poate vedea meniul de ajustare a schemei de organizare ierarhice, accesat prin selectarea în meniul principal yEd a căii Layout/Hierarchical.

[17] https://www.yworks.com/en/products/yfiles/yed/

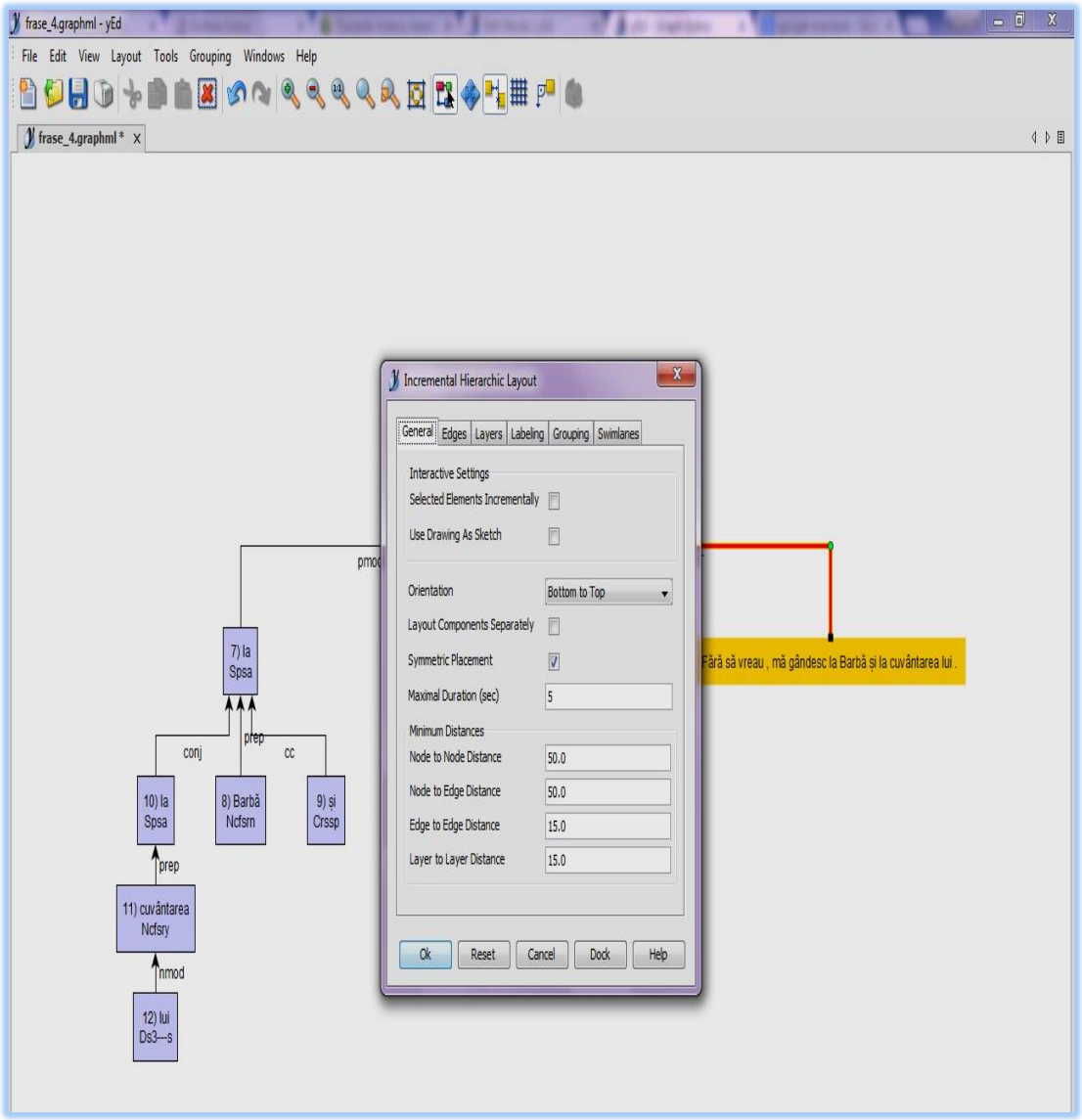


Figura 3.3. Meniul de ajustare a schemei ierarhice (Incremental Hierarhic Layout) în yEd.

Dintre cele două modalități de lucru cu yEd, Edit Mode și Navigation Mode, ne interesează primul, care ne permite să intervenim pe structura diagramei; vom menționa doar funcțiile pe care le-am utilizat cel mai frecvent în procesul de corectare a arborilor de dependențe:

- selectarea nodurilor și muchiilor: pentru selectarea unui nod sau a unei muchii, un simplu click pe element este suficient; pentru a selecta mai multe noduri sau muchii, trebuie menținută apăsată tasta SHIFT;
- crearea unei noi muchii: menținând apăsat butonul din stînga al mouse-ului, se trage mouse-ul de la nodul dependent către nodul centru; dacă mouse-ul nu este eliberat pe un nod, se crează un punct de control, ceea ce permite crearea unei muchii care să ocolească un anumit nod și facilitează vizualizarea arborelui până la o nouă executare a algoritmului de reprezentare ierarhică;

- crearea etichetelor pentru muchii: se activează meniul disponibil pe butonul de click din dreapta (meniu senzitiv la context, după cum este denumit în manualul yEd) pentru o anumită muchie şi se selectează Add Label; apare o locaţie de etichetă lângă muchie, cursorul fiind activ şi permiţând introducerea unei valori pentru etichetă;
- mutarea nodurilor: se selectează nodul și, apăsând butonul din stânga al mouse-ului, se trage cursorul mouse-ului către poziția dorită;
- ştergerea elementelor: se selectează elementul și se apasă tasta Delete sau se deschide meniul senzitiv la context și se selectează DELETE;
- selectarea şi editarea etichetelor: o etichetă de element se poate selecta printr-un click; editorul acestei etichete se selectează printr-un dublu-click sau prin selectarea elementului, activarea meniului senzitiv la context şi selectarea opţiunii Edit Label;
- funcţia Fit Node to Labels (ro. Potriveşte nodurile după etichete), disponibilă din meniul Tools, este utilă pentru a ajusta dimensiunea nodurilor la dimensiunea etichetelor pentru noduri, care pot conţine multă informaţie: de exemplu, nodul rădăcină are ca etichetă conținutul întregii propoziții (vedeți nodul galben din Figura 3.3.), iar un nod obișnuit poate conține ID-ul în propoziție al cuvântului pe care-l reprezintă, forma și msd-ul cuvântului respectiv.
- funcția Reverse Selected Edges (ro. Inversează muchiile selectate), inversează direcția relației de dependență pentru una sau mai multe muchii selectate, inversând raportul centru-dependent între cuvintele pe care muchia sau muchiile le unesc.

Dintre formatele pe care le poate interpreta yEd, am ales să utilizăm formatul GRAPHML, de tip XML, care permite definirea de către utilizator a proprietăților elementelor. Pentru conversia din formatul CONLL în formatul GRAPHML am folosit un script perl furnizat de colaboratorii de la IULA (care au recomandat, de altfel, utilizarea instrumentului yEd), iar pentru conversia din formatul GRAPHML în formatul CONLL (după efectuarea corecturilor) am implementat un script C#.

Anexa 3 reproduce în formatul CONLL şi în formatul corespunzător GRAPHML propoziția "Are 52 de ani, este căsătorit și are o fiică.", adnotată sintactic automat și corectată manual în interfața yEd. Arborele așa cum poate fi vizualizat cu yEd, exportat în format png, este reprodus în continuare Figura 3.4. yEd oferă și posibilitatea modificării culorii nodurilor și muchiilor, pe care am folosit-o în procesul de corectură pentru a marca cu roșu noduri sau muchii asupra cărora nu am putut decide pe moment o adnotare potrivită. În figura 3.4, nodul rădăcină se distinge de celelalte, fiind colorat în galben, o procedură standard pentru toți arborii de dependențe.

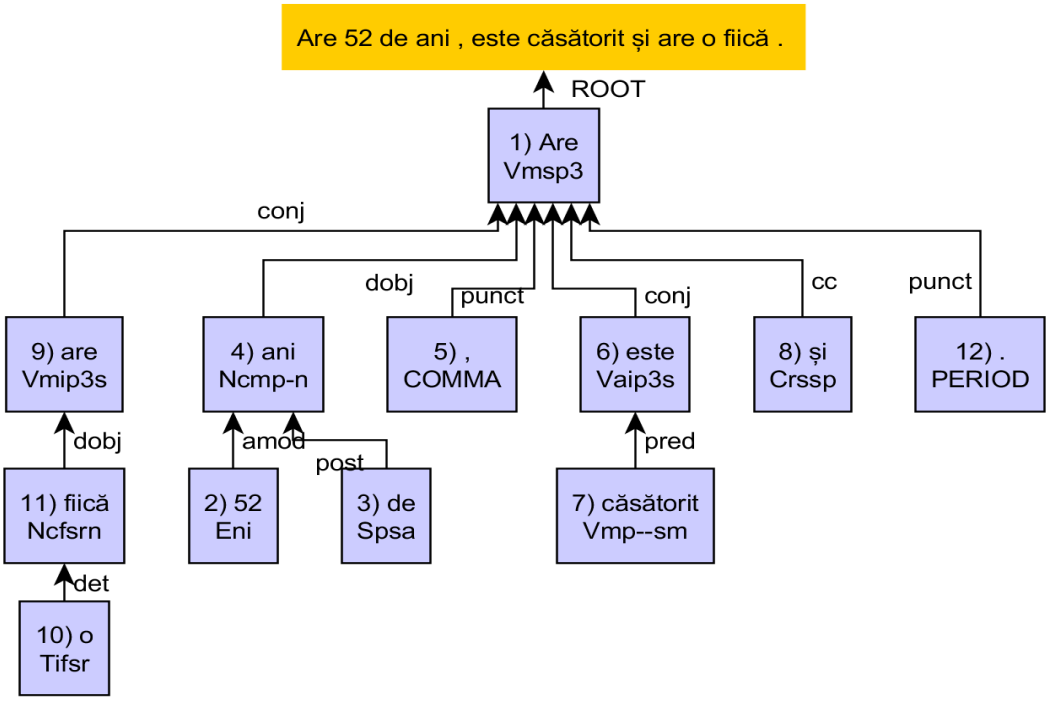


Figura 3.4. Arborele sintactic cu dependenţe pentru propoziţia “Are 52 de ani, este căsătorit și are o fiică.” aşa cum este vizualizat cu yEd.

3.5. MaltEval

Spuneam în secțiunea 1.3. că am folosit pentru evaluarea rezultatelor instrumentul MaltEval**,** implementat în Java prin adaptarea scripturilor Perl de evaluare lansate de competițiile CONLL 2006 și 2007 (eval.pl și eval07.pl). Aceste competiții au consacrat Labeled Attachment Score (LAS), introdus de (Nivre et al. 2004) – care reprezintă raportul dintre numărul de cuvinte cu centre și etichete corect identificate și numărul total de cuvinte din propoziție – drept măsură a performanței analizei sintactice, dar instrumentul MaltEval oferă și posibilitatea calculării altor măsuri: LA(numărul de cuvinte cu etichete corecte raportat la numărul total de cuvinte din propoziție), UAS (numărul de cuvinte cu centru corect identificat raportat la numărul total de cuvinte din propoziție), AnyRight (numărul de cuvinte care au centrul, eticheta sau pe amândouă corect identificate raportat la numărul total de cuvinte din propoziție) etc.

De asemenea, prin intermediul unor fișiere externe de evaluare, editabile de către utilizator, și al unui flag special –e care rulează toate aceste fișiere de evaluare, MaltEval poate reproduce rezultatele statistice oferite de eval7.pl referitoare la 1. acurateţe și distribuția acurateții de adnotare relativ la setul de etichete morfo-lexicale, 2. rata erorii și distribuția acesteia relativ la setul de etichete morfo-lexicale; 3. precizia și recall-ul pentru identificarea tipului relației de dependență; 4. precizia și recall-ul pentru identificarea tipului și centrului

relației de dependență; 5. cuvinte care sunt cel mai des adnotate greșit, contexte în care apar cel mai des greșeli, perechi de relații care sunt confundate cel mai des, etc.

Cu acest sistem de flag-uri și fișiere externe de evaluare, MaltEval permite specificarea de către utilizator a mai mult de 25 de parametri de evaluare. De exemplu, cu fișierul evaluare.xml de mai jos, executat prin flag-ul –e, am putut specifica lui MaltEval că ne interesează evaluarea pentru scorul AnyRight (similar pentru LA și UAS):

```
<evaluation>
        <parameter name="Metric">
                <value>AnyRight</value>
        </parameter>
        <parameter name="GroupBy">
                <value>Token</value>
        </parameter>
</evaluation>
```

Cele două flag-uri obligatorii pentru utilizarea MaltEval, -g și -s, specifică fișierul gold-standard (cel a cărui adnotare sintactică a fost validată manual) și fișierul sursă, a cărui adnotare automată dorim să o evaluăm în raport cu adnotarea gold-standard.

java -jar MaltEval.jar -g goldfile -s sourcefile

Am apreciat ca foarte utilă posibilitatea de a evalua directoare întregi, în loc de fișiere, cu condiția ca numărul de fișiere din cele două directoare evaluate să fie egal, specificând calea către directoare după cum urmează:

java -jar MaltEval.jar -g golddir/ -s sourcedir/

Prin flag-ul -v urmat de opțiunea 1, se activează un modul MaltEval de vizualizare arborescentă a fișierelor gold-standard și sursă, arbore cu arbore, în paralel, în aceeași fereastră. Flag-ul -v dezactivează celelalte flag-uri cu excepția lui -s și -g și nu oferă nici o evaluare propriu-zisă, ci o serie de facilități de căutare în corpusuri, după criterii precum adâncimea arcului, direcția arcului, poziția în propoziție a relației, versiunea scurtă a etichetei morfo-lexicale, eticheta morfo-lexicală completă, lema, tipul relației de dependență, etc. Diferențele între fișierul gold-standar și cel de evaluat, adică erorile produse de adnotatorul automat, sunt marcate prin colorarea în roșu a relațiilor greșite.

Figura 3.5: Opțiunea de vizualizare a erorilor (marcate cu roșu în arborele din partea inferioară a imaginii), căutare (butoanele de Search din partea superioară a imaginii) și navigare între propoziții (Prev sent, Next sent) și erori (Prev error și Next error) în seturile de propoziții comparate.

În imaginea din Figura 3.5., în căsuţa *Search in* a fost selectat Gold-Standard, cealaltă opţiune posibilă fiind parse_1, adică fişierul de evaluat (sau sursă). Din căsuţa *Search by* a fost selectată opţiunea *Postag*, selecţie ce încarcă automat în căsuţa *Search for* o listă cu toate etichetele morfo-lexicale prezente în corpusul în care se face căutarea. De aici am selectat *Vag*, ceea ce înseamnă că dorim să căutăm verbe auxiliare la modul gerunziu. Căutarea returnează o listă cu indicii propoziţiilor în gold-standard (vezi căsuţa *Result*), din care pot fi selectate, pe rând, spre vizualizare, toate propoziţiile care conţin un verb cu eticheta morfo-lexicală *Vag*. În imagine, propoziţia cu indicele 29 în gold-standard este încărcată spre vizualizare, împreună cu echivalenta ei în corpusul de evaluat, iar verbul etichetat *Vag* (“fiind”) este marcat prin caractere îngroşate.

CAPITOLUL 4

CONSTRUIREA NUCLEULUI DE BANCĂ DE ARBORI PENTRU LIMBA ROMÂNĂ

4.1. Construirea corpusului de lucru

Reamintim că obiectivul cercetării noastre este construirea unui corpus care, deşi de dimensiuni modeste, să fie suficient de divers şi reprezentativ pentru limbă, astfel încât pe baza sa să poată fi construit un model statistic fiabil.

Ne-am propus să urmărim două criterii importante în selecţia corpusului de adnotat:

- **diversitate sintactică**: pe care considerăm că am asigurat-o alegând ROMBAC (vezi descrierea detaliată din Secţiunea 3.1.) drept corpus de selecţie; au fost astfel acoperite patru stiluri funcţionale ale limbii şi cinci domenii literare diferite;
- **reprezentativitate sintactică**: pe care am atins-o luând drept criteriu esenţial de selecţie a propoziţiilor de adnotat frecvenţa verbelor în ROMBAC.

Pe baza informației morfo-lexicale din ROMBAC, am putut identifica automat verbele predicative și calcula frecvențele acestora în corpus. Ne-am concentrat pe cele mai frecvente 500 de verbe din fiecare dintre cele 5 secțiuni ale corpusului și am extras din ROMBAC câte 1.000 de propoziții din fiecare secțiune, astfel încât fiecare dintre cele 500 de verbe frecvente să apară în cel puțin două propoziții din fiecare domeniu. Cele 5.000 de propoziţii extrase astfel din ROMBAC vor reprezenta corpusul de lucru în continuare (treebank-ul). Propoziţiile selectate trebuie să aibă o lungime cuprinsă între 10 şi 40 de cuvinte şi cel puţin un verb predicativ în structură.

În mod natural, unele verbe se vor întâlni în mai multe sau chiar toate secțiunile corpusului; în plus, cele mai multe dintre propoziţii conţin mai mult de un verb predicativ. De aceea, multe dintre verbe vor avea în resursa noastră o frecvenţă mai mare de 2, aceasta fiind doar frecvenţa minimă garantată fiecărui verb. De exemplu, pentru v_1, v_2, .., v_i, .. v_{500} lista celor mai frecvente 500 de verbe din secţiunea jurnalistică a corpusului ROMBAC, pentru orice i, v_i va apărea cel puţin de două ori în secţiunea jurnalistică a treebank-ului pe care îl vom construi.

În funcţie de particularităţile stilistice ale sub-domenilor, distribuţia verbelor în ROMBAC este diferită: de exemplu, secţiunea jurnalistică şi cea literară au o diversitate mai mare a verbelor, secţiunea medicală abundă în verbe dar diversitatea acestora este mică, în timp ce secţiunea academică are mult mai puţine verbe (şi, implicit, mai multe adjective şi substantive). Tabelul 4.1. arată frecvenţele minime şi maxime în ROMBAC ale celor 500 de

verbe selectate ca reprezentative din fiecare secţiune. De exemplu, în secţiunea jurnalistică, verbul cu cea mai mare frecvenţă din lista de 500 de verbe selectate apare de 25.444 ori doar în această secţiune, în timp ce verbul cu cea mai mică frecvenţă apare de 197 ori. După cum se poate vedea în tabel, verbul selectat care are cea mai mică frecvenţă în secţiunea sa (78) are totuşi un număr important de apariţii, ceea ce ne asigură că lucrăm cu verbe frecvent folosite în limbă. Pentru o listă completă a verbelor selectate împreună cu frecvenţele lor în fiecare secţiune din ROMBAC și frecvențele în treebank-ul asamblat de noi, vedeţi Anexa 4.

Secţiune ROMBAC	**Frecvenţa minimă**	**Frecvenţa maximă**
Jurnalistic	197	25.444
Academic	104	6.388
Literar	262	30.664
Medical	78	63.053
Juridic	91	29.291

Tabelul 4.1. Frecvenţele minime şi maxime în ROMBAC ale celor 500 de verbe selectate ca reprezentative din fiecare secţiune.

Tabelul 4.2 cuprinde informaţii statistice despre cele 5.000 de propoziţii selectate pentru adnotare: numărul de cuvinte pe care îl cuprinde fiecare secţiune, precum şi numărul de leme distincte şi numărul de cuvinte distincte. Se poate observa că secţiunea literară este cea mai săracă, atât ca număr de cuvinte (propoziţiile extrase au în medie în jur de 16 cuvinte) cât şi ca număr de leme şi forme distincte, ceea ce indică faptul că avem de-a face cu un vocabular modest. Dacă lungimea medie a propoziţiei nu este surprinzătoare pentru acest stil literar – sunt rari autorii care se aventurează să creeze conţinut beletristic folosind enunţuri lungi şi stufoase – dimensiunea vocabularului indică faptul că nu avem de a face cu o proză pretenţioasă şi livrescă, ci cu o literatură bazată pe experienţe practice, de viaţă. Prin contrast, după cum e şi de aşteptat, secţiunea academică are cel mai bogat vocabular dintre secţiunile corpusului, cu aproape de 8 ori mai multe leme distincte decât secţiunea juridică, deşi numărul total de cuvinte din secţiunea juridică este de 1,5 ori mai mare decât numărul total de cuvinte din secţiunea academică. Vocabularul redus din secţiunea juridică se datorează limbajului tehnic şi controlat utilizat în texte de acest tip. Raportul scăzut, de 1,5, între numărul de forme şi numărul de leme distincte (3.610 versus 2.326, respectiv 7.272 versus 4.816) din secţiunea medicală şi din cea academică se explică prin faptul că aici se întâlnesc multe nume proprii, reprezentând denumiri de medicamente, respectiv denumiri de autori, care nu flexionează morfologic.

Secțiune	Cuvinte în total	Leme distincte	Forme distincte
Jurnalistic	23.710	2.150	5.155
Literar	16.697	658	2.664
Academic	20.408	4.816	7.272
Medical	20.818	2.326	3.610
Juridic	30.188	632	3.185

Tabelul 4.2. Statistici pe treebank-ul dezvoltat: număr total de cuvinte, număr de leme distincte, număr de forme distincte pentru fiecare dintre secţiuni.

Tabelul 4.3. prezintă distribuţia numărului de cuvinte din fiecare dintre cele cinci secţiuni ale treebank-ului în funcţie de partea de vorbire asociată fiecărui cuvânt. Se observă că abundă substantivele, cu numărul cel mai mare de ocurenţe în secţiunea juridică şi în cea jurnalistică. Numărul mare de cuvinte asociat fiecărei părţi de propoziţie din secţiunea juridică este datorat unei lungimi medii de 30 de cuvinte pe propoziţie în această secţiune şi, implicit, unui număr total de cuvinte semnificativ mai mare în raport cu celelalte secţiuni. Totuşi, numărul foarte mare de numerale din secţiunea juridică este o particularitate a acestui stil şi se datorează prezenţei în corpus a numeroşi identificatori numerici care reprezintă denumiri pentru legi şi articole de legi.

Partea de vorbire	Jurnalistic	Literar	Academic	Medical	Juridic
Adjective	1.638	925	1.769	1.731	2.099
Conjuncții	584	870	693	855	1.105
Determinanţi	319	546	301	300	609
Numerale	97	103	116	102	1.303
Substantive	7.009	3.502	5.491	5.875	8.925
Pronume	992	1.458	866	860	1.253
Punctuație	2.888	2.380	3.561	2.347	4.108
Particule	358	639	266	530	679
Adverbe	687	1.014	777	783	686
Prepoziţii	3.426	1.860	2.498	2.849	4.319
Articole	869	572	881	588	955
Verbe	3.695	2.937	2.598	3.343	4.011
Cuvinte reziduale	1	1	2	15	17
Abrevieri	44	4	10	105	116

Tabelul 4.3. Distribuţia frecvenţei cuvintelor fiecărei secţiuni din treebank pe părţi de vorbire.

4.2. Adnotarea corpusului de lucru

Așa cum menționam în secțiunea 1.3, am ales ca în dezvoltarea treebankului să nu pornim de la corpus ne-adnotat, ci să exploatăm o metodologie deja testată, anume să adnotăm corpusul automat cu un parser statistic (MaltParser) și un model antrenat pe limba spaniolă (pe treebank-ul IULA LSP) și să corectăm manual adnotarea propusă pentru a corespunde standardelor gramaticii limbii române.

Metodologia a fost deja aplicată de către Arias et al. (2014) pentru a grăbi crearea unui treebank pentru limba catalană. Argumentele principale ale autorilor pentru soluția de adnotare aleasă sunt scorul LAS foarte bun (94%) obținut pentru modelul statistic spaniol atunci când se adnotează texte spaniole, precum și facilitatea oferită de MaltParser de a produce modele statistice de-lexicalizate, excluzând din modelul de trăsături (vezi secțiunea 3.3.2.) trăsăturile de tip LEX. În experimentul pentru catalană, modelul de limbă spaniolă aplicat pe propoziții în catalană a fost evaluat la 79% scor LAS; primul model lexicalizat de limbă catalană a fost antrenat după 1.000 de propoziții corectate și a dus la un scor de adnotare de 84% pentru următorul set de propoziții adnotate, cu o lungime medie de 20 de cuvinte. După antrenări succesive ale modelului statistic pe catalană până la 2.400 de propoziții, scorul LAS a crescut mult mai lent, ajungând la 86%.

Pentru a menține consistența adnotării, am decis să pornim, într-o primă etapă, cu prima jumătate a corpusului de adnotat în care am inclus propoziții de lungime cuprinsă între 10 și 20 de cuvinte, și să lăsăm propozițiile mai lungi, și implicit mai complexe sintactic, pentru adnotare și corectare într-o etapă secundară. Fiecare secțiune a corpusului a fost împărțită astfel în două tranșe a câte 500 de propoziții: în prima etapă se corectează prima tranșă, cu propoziții mai scurte, din fiecare secțiune, iar în cea de-a doua etapă se corectează propozițiile de lungime mai mare rămase. Ipoteza este că procedând în acest mod:

1) ne vom concentra în prima parte pe familiarizarea cu principiile de corectare, aplicându-le pe propoziții mai scurte, care să pună mai puține probleme de corectare;
2) corectura din etapa a doua va fi mai facilă, deoarece fiecare dintre seturile secundare de propoziții corectate corespunzătoare unei anumite secțiuni din text și unui anumit stil literar (jurnalistic, beletristic, academic, științific și juridic) va beneficia de un model statistic de adnotare antrenat pe datele similare din seturile corectate în prima etapă. De exemplu, pentru al doilea set de 500 de propoziții aparținând stilului jurnalistic, adnotarea automată va beneficia de un model

statistic antrenat deja pe date de tip jurnalistic din primul set de 500 propoziții corectat (vezi Figura 4.1.)

Am început adnotarea automată cu un set de 500 de propoziții din sub-corpusul jurnalistic folosind modelul statistic de-lexicalizat de limbă spaniolă. Am optat să începem cu stilul jurnalistic datorită intuiției că modelul statistic obținut va fi unul destul de divers atât sintactic cât și lexical (nu controlat și specific, cum ar fi fost un model antrenat pe sub-corpusul medical sau juridic, de exemplu); în același timp, datorită particularităților stilistice, ne-am așteptat ca procesul de corectură să fie mai facil decât cel al unui text beletristic, în care un limbaj figurativ poate pune probleme de interpretare sintactică și semantică chiar și unui adnotator experimentat.

Am decis antrenarea unui model lexicalizat pe limba română după doar 500 de propoziții corectate, intuind că modelul obținut va avea deja performanțe mai bune decât cel spaniol, lucru confirmat de evaluările efectuate (vezi secțiunea 5.1). Am repetat procedura de reantrenare după corectura a 500 de propoziții din fiecare sub-corpus, adăugând de fiecare dată la corpusul de antrenare ultimele propoziții corectate.

După cum se poate vedea în Figura 4.1, ciclul de lucru este: 1) adnotare cu modelul statistic cel mai performant la dispoziție (în imagine, săgeata albastră indică procesul de adnotare); 2) corectura setului de propoziții adnotat la pasul 1 (în imagine, săgeata verde indică procesul de corectură); 3) adăugarea setului corectat la corpusul de antrenare și re-antrenarea unui model extins, mai performant decât precedentul (săgeata mare roșie din imagine simbolizează antrenarea progresivă pe seturi de date tot mai mari).

Fiecare tranşă de propoziţii a fost corectată manual de către doi adnotatori umani, un specialist informatician şi un specialist lingvist. Adeseori, aceştia au comunicat între ei pentru a conveni asupra cazurilor de adnotare problematice. Într-o etapă ulterioară finalizării acestui proiect, intenţionăm să folosim tehnici automate de identificare a erorilor pentru a corecta eventuale scăpări ale celor doi adnotatori umani.

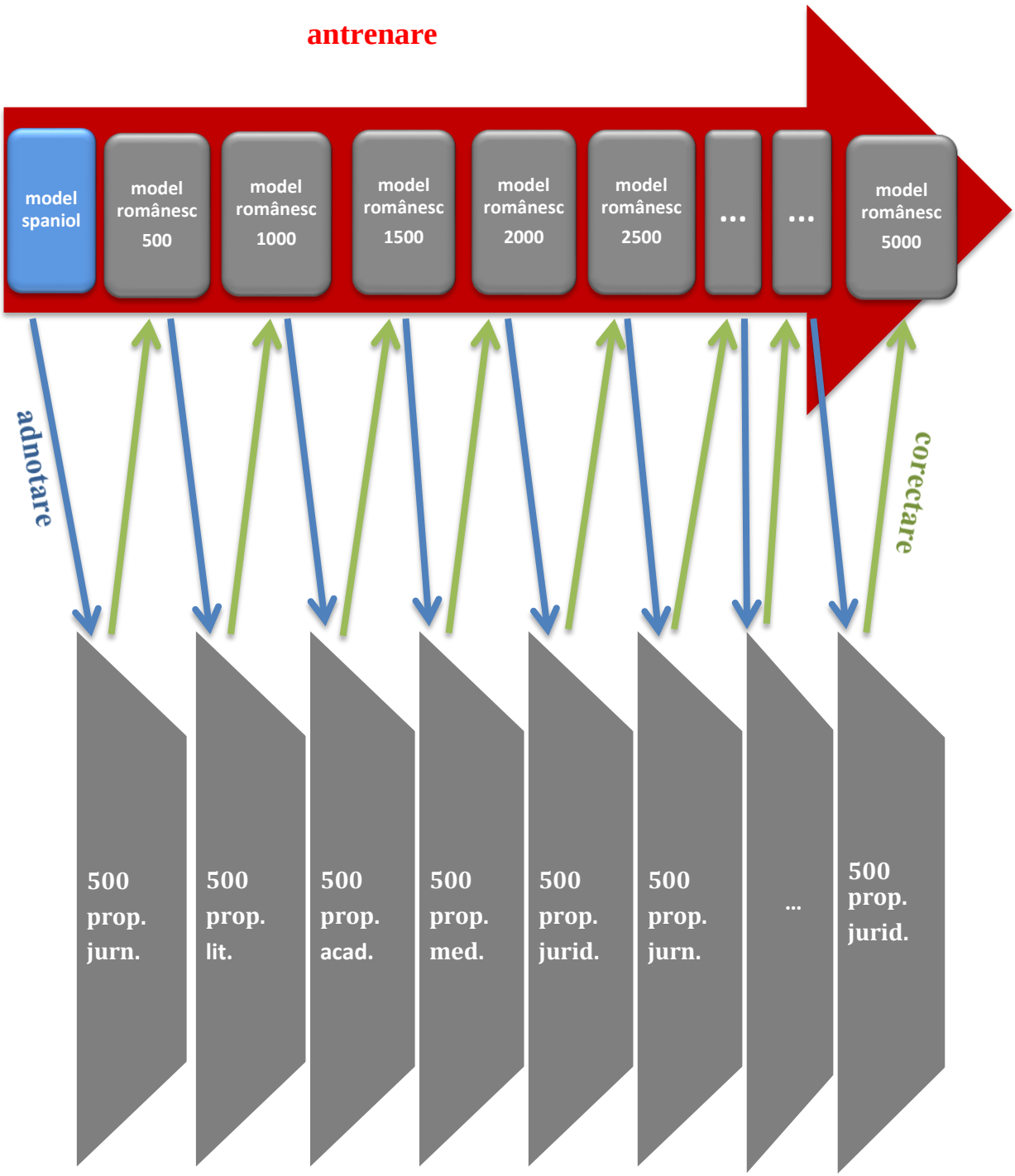


Figura 4.1. Ciclul de adnotare/corectare/re-antrenare iterat pe seturi de 500 de propoziţii.

CAPITOLUL 5

EVALUAREA REZULTATELOR

5.1. Evaluarea performanţelor modelelor statistice utilizate

Deoarece procesul de dezvoltare a treebank-ului s-a extins pe durata a mai mult de un an, am efectuat mai multe etape de evaluare, care ne-au ghidat în munca de adnotare şi corectare. Într-o etapă iniţială, după corectarea primelor 100 de propoziţii (din secţiunea jurnalistică), am calculat mai multe scoruri puse la dispoziţie de MaltEval, pentru a vedea cum arată, reprezentată numeric, cantitatea de muncă efectuată (pe care din punct de vedere al efortului uman o consideram dificilă, dar mai puţin împovărătoare decât dacă am fi început adnotarea de la zero). Fişierul gold-standard (transmis lui MaltEval prin parametrul -g) conţine cele 100 de propoziţii corectate manual, în timp ce fişierul de evaluat (transmis lui MaltEval prin parametrul -s) conţine aceleaşi propoziţii aşa cum au fost iniţial adnotate de MaltParser cu modelul statistic spaniol.

O rată a erorii mare (aproximativ 79% pentru condiţia ca atât centrul cât şi eticheta relaţiei de dependenţă să fie identificată corect; a se vedea Tabelul 5.1, linia scorului LAS) era de aşteptat, din moment ce am folosit un model de-lexicalizat antrenat pe o limbă diferită. De asemenea, după cum s-a văzut în secţiunea 2.2., am adoptat anumite principii de analiză diferite de cele din modelul statistic spaniol şi am rafinat anumite relaţii: aceste decizii măresc suplimentar distanţa dintre analiza propusă de modelul spaniol şi cea pe care noi o considerăm corectă. Valoarea scorului LAS semnalează că există un număr destul de mic de cuvinte a căror adnotare nu necesită corecturi manuale: aproximativ 21% din numărul total de cuvinte. Dar scorul AnyRight arată că un număr important de cuvinte (în jur de 71%) sunt deja adnotate cu informaţie corectă, fie la nivel de centru, fie la nivel de etichetă a relaţiei. Atât aceste scoruri, cât şi experienţa de lucru, ne-au încurajat să continuăm cu metodologia de adnotare aleasă, în perspectiva înlocuirii modelului spaniol cu cel românesc după 500 de propoziţii corectate.

Măsura	**Valoarea**
LAS	0,216
LA	0,417
UAS	0,514
AnyRight	0,715

Tabelul 5.1: Scorurile LAS, LA, UAS şi AnyRight pentru primele 100 de propoziţii corectate.

Cea de a doua evaluare a avut loc după corectarea primelor 500 de propoziţii (din secţiunea jurnalistică), când un prim model statistic de adnotare sintactică a fost antrenat pe propoziţii în limba română. Acest model este lexicalizat şi complet adaptat la setul de relaţii de dependenţă şi la principiile de adnotare alese de noi.

O practică comună în utilizarea instrumentelor statistice în domeniul PLN este optimizarea acestor instrumente pe anumite părţi ale corpusului de antrenare: parametrii modelului statistic sunt calculaţi şi fixaţi astfel încât modelul să producă cele mai bune rezultate posibile (în termenii unei măsuri statistice) pentru un anumit set sau pentru un anumit tip de date. Am folosit MaltOptimizer – un instrument disponibil liber, dezvoltat pentru MaltParser – pentru a antrena un model (lexicalizat) optimizat pe aceleaşi 500 de propoziţii corectate în prima etapă de corectare.

Cele două modele obţinute, cel ne-optimizat (Ro-non-opt-500) şi cel optimizat (Ro-opt-500), împreună cu cel spaniol, au fost utilizate pentru adnotarea următoarei tranşe de propoziţii de corectat, 500 de propoziţii din secţiunea literară (vezi Figura 4.1). Corectarea acestor propoziţii s-a făcut pe adnotarea cu modelul optimizat, iar evaluarea, după corectarea a doar o sută de propoziţii (vezi Tabelul 5.2), ne-a confirmat că modelul optimizat este mai bun decât cel ne-optimizat, chiar dacă este utilizat pe date aparţinând unui stil funcţional diferit. De asemenea, corectarea manuală a celei de-a doua tranşe de propoziţii a implicat mai puţine eforturi şi a necesitat semnificativ mai puţin timp decât corectarea primei tranşe, ceea ce justifică înlocuirea precipitată a modelului spaniol cu un model românesc lexicalizat.

După cum se poate vedea în Tabelul 5.2, creşterea scorului LAS este mult mai mare decât în cadrul experimentului pentru limba catalană: 0,345 (de la 0,202 la 0,547, pentru modelul optimizat) faţă de 0,074 (de la 0.790 la 0.864). Acest lucru poate fi explicat de valoarea deja importantă a scorului LAS pentru experimentul catalan utilizând modelul spaniol: în experimente statistice, valoarea unei măsuri este cu atât mai greu de îmbunătăţit cu cât este mai apropiată de valoarea spre care tinde (1 în cazul scorului LAS).

Modelul statistic	**LAS**
RO-non-opt-500	0.469
RO-opt-500	0.547
Spaniol	0.202

Tabelul 5.2. Evaluarea modelelor Ro-non-opt-500, Ro-opt-500 şi a modelului spaniol pe primele 100 de propoziţii din secţiunea literară.

Am continuat munca de corectare pe propoziţiile adnotate cu modelul optimizat până la finalizarea setului de 500 de propoziţii din secţiunea literară şi am evaluat din nou performanţa

modelului: scorul LAS obţinut a fost chiar mai bun, 0,580 (vezi Tabelul 5.3). Apoi, aşa cum am descris în secţiunea 4.2., am repetat procedura de re-antrenare a unui model optimizat pe setul de propoziţii acumulate şi corectare a următoarei tranşe de 500 de propoziţii, până la finalizarea corectării întregului treebank. După cum se poate observa în Tabelul 5.3, scorul LAS (calculat prin compararea fişierului corectat manual, drept gold-standard, cu cel adnotat automat, drept fişier de test) a crescut după fiecare pas, cu excepţia primei tranşe de propoziţii din secţiunea juridică, unde a avut loc o uşoară scădere. Cele mai multe propoziţii din acestă secţiune au o structură specifică: încep cu diferite secvenţe de cuvinte care funcţionează ca identificatori pentru articole sau secţiuni de articole de lege. Când se află la începutul propoziţiei, astfel de secvenţe nu fac parte de fapt din structura sintactică a propoziţiei, ci reprezintă nişte etichete pentru conţinutul propoziţiei: de aceea am decis să le legăm de centrul verbal prin relaţii de tip *parataxis*, la fel ca în cazul vorbirii indirecte. Modelul statistic, ne-antrenat până în acel moment pe un asemenea tip de enunţ, a asociat etichete greşite secvenţelor menţionate, acest lucru afectând de cele mai multe ori analiza propoziţiei, inclusiv prin adnotarea eronată a rădăcinii.

Secţiunea adnotată	**Corpus folosit pentru antrenarea modelului statistic**	**LAS**
Jurnalistic 1	Spaniol	0.243
Literar 1	Jurnalistic 1	0.580
Academic 1	Jurnalistic 1+ Literar 1	0.738
Medical 1	Jurnalistic 1+Literar 1 +Academic 1	0.773
Juridic 1	Jurnalistic 1+Literar 1 +Academic 1+Medical 1	0.710
Jurnalistic 2	Jurnalistic 1+Literar 1 +Academic 1+Medical 1 + Juridic 1	0.750
Literar 2	Jurnalistic 1+Literar 1 +Academic 1+Medical 1 + Juridic 1 + Jurnalistic 2	0.774
Academic 2	Jurnalistic 1+Literar 1 +Academic 1+Medical 1 + Juridic 1 + Jurnalistic 2 + Literar 2	0.813
Medical 2	Jurnalistic 1+Literar 1 +Academic 1+Medical 1 + Juridic 1 + Jurnalistic 2 + Literar 2 + Academic 2	0.817
Juridic 2	Jurnalistic 1+Literar 1 +Academic 1+Medical 1 + Juridic 1 + Jurnalistic 2 + Literar 2 + Academic 2	0.870

Tabelul 5.3. Evoluţia scorului LAS după fiecare etapă de corectare manuală.

5.2. Studiul erorilor de adnotare automată

Erorile ce apar în experimente statistice sunt fie *sistematice*, caz în care pot fi evitate atunci când modelul statistic este îmbunătăţit prin adăugarea de exemple adnotate sau prin căutarea şi corectarea erorilor în corpusul de antrenare, fie *ne-sistematice*, când avem de-a face cu erori ce vin din contexte ambigue, unde numai intervenţia umană poate determina soluţia corectă de analiză.

5.2.1. Erori în evaluarea distorsionată

O soluţie de identificare a erorilor *ne-sistematice* este o metodologie pe care am mai utilizat-o pentru identificarea erorilor de adnotare morfo-lexicală (Tufiș și Irimia, 2006) şi pe care o denumeam atunci „evaluare distorsionată” (eng. “biased evaluation”): utilizând un model statistic antrenat pe un anumit set de propoziţii (în cazul nostru cele 500 de propoziţii din prima tranşă a secţiunii jurnalistice, corectate manual), re-adnotăm acelaşi set de propoziţii. Se presupune că modelul învaţă să adnoteze toate cazurile sistematice şi nu vor fi cazuri ne-văzute în faza de re-antrenare, deci erorile rămase ar trebui să fie de tipul ne-sistematic, necesitând decizia umană.

Scorul LAS foarte mare obţinut pentru această evaluare distorsionată, 0.972, ne indică faptul că adnotarea umană este consistentă pe acest set de antrenare şi că o creştere constantă a setului de antrenare ne va permite să atingem la un moment dat valori similare pentru propoziţii neîntâlnite în faza de antrenare.

Prin utilizarea lui MaltEval cu flagul –e (vezi secţiunea 3.5.) am putut obţine rezultate statistice detaliate despre erorile apărute în cadrul acestui experiment, pe care le-am redirecţionat către un fişier .txt, pentru a putea sorta şi interpreta aceste informaţii. Fişierul rezultat conţine o secţiune care enumeră cele mai frecvente *n* erori (*n* este implicit 10, dar acest parametru poate fi modificat) şi care ne semnalează că, de exemplu, de 42 de ori, adnotatorul a calificat drept rădăcină a propoziţiei un element de punctuaţie (de cele mai multe ori (32) punctul final, de 10 ori o virgulă din propoziţie), atribuind eticheta ROOT în locul etichetei *punct*: aceasta este o eroare specifică MaltParser, care nu impune condiţia de rădăcină unică pentru un arbore de dependenţe, generând astfel multe situaţii (aprox. 8% din setul evaluat) în care propoziţia are între două şi şase rădăcini. Din experienţa noastră şi a echipei IULA cu care am colaborat, procentul de propoziţii pentru care MaltParser identifică mai mult de o rădăcină scade pe măsură ce dimensiunile şi, implicit, performanţa modelului statistic cresc: pentru evaluarea normală, ne-distorsionată, s-a observant scăderea treptată a acestui procent de la valori (aproximative) de 30% până la valori de 15%. Acest tip de eroare este deci una sistematică, dar pe care, în această etapă, modelul nu o poate elimina nici măcar prin evaluare distorsionată.

Figura 5.1. prezintă un exemplu de analiză automată în care adnotatorul, deşi identifică în mod corect rădăcina propoziţiei (verbul predicativ “cere”), pune eticheta ROOT pe alte două elemente ale prepoziţiei, locuţiunea prepoziţională “în legătură cu” şi punctul de la finalul propoziţiei.

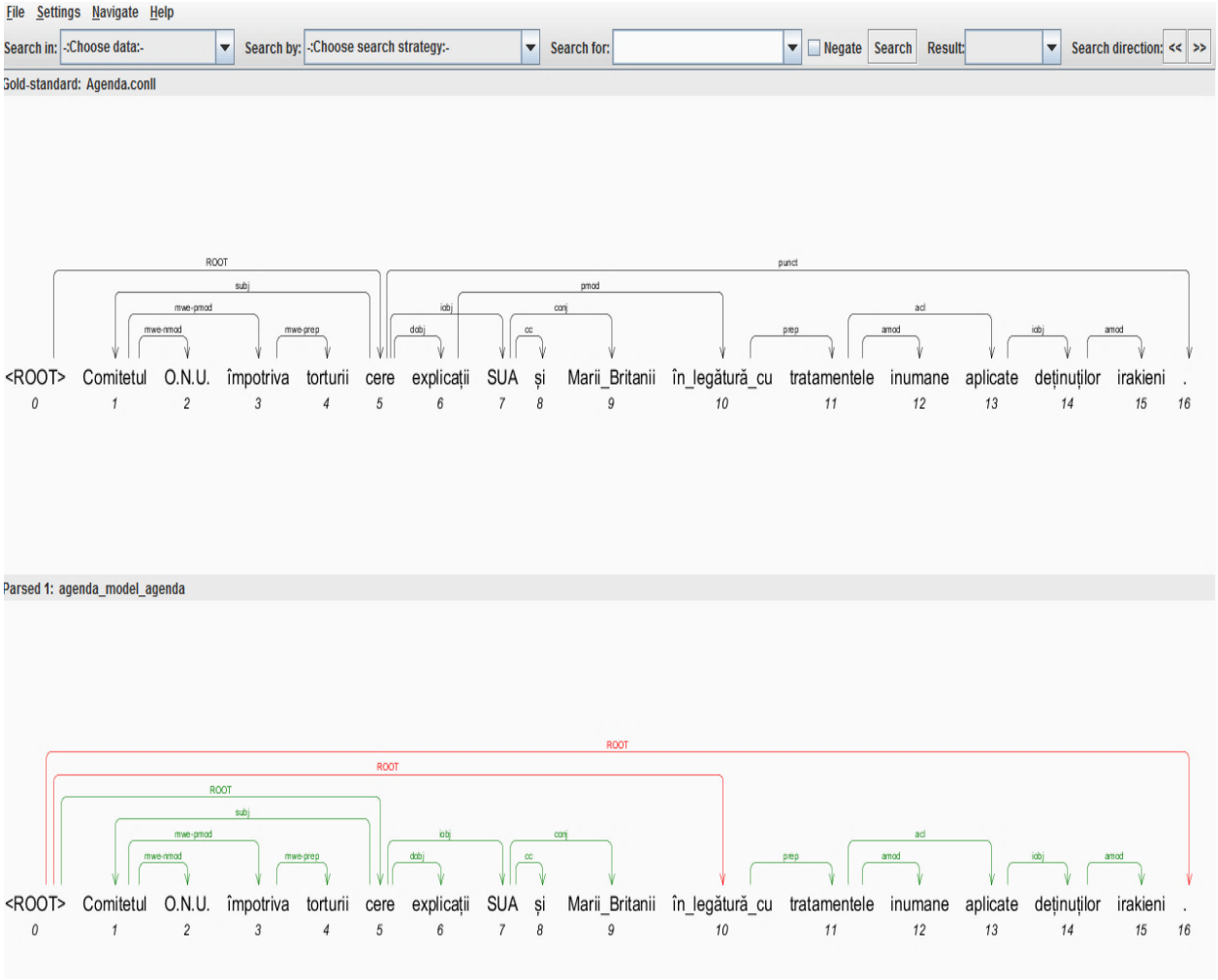


Figura 5.1: Adnotarea sintactică automată (în partea de jos a imaginii) şi adnotarea gold-standard (în partea de sus a imaginii) pentru propoziţia “Comitetul O.N.U. împotriva torturii cere explicaţii SUA şi Marii Britanii în legătură cu tratamentele inumane aplicate deţinuţilor iranieni.”

O altă secţiune a fişierului cu rezultate de evaluare prezintă distribuţia erorilor de adnotare (în termeni de recall şi precizie) pentru fiecare tip de relaţie de dependenţă din setul de date evaluate. Tabelul 5.4. prezintă cele mai frecvente relaţii din setul de evaluare (cu număr de ocurenţe în gold-standard de peste 500, vezi coloana 2 din tabel) împreună cu recall-ul şi precizia de identificare a etichetei corecte pentru o anumită relaţie. Tabelul 5.5 prezintă aceleaşi măsuri dar calculate pentru identificarea corectă atât a etichetei cât şi a centrului.

relaţia	**gold**	**identificate corect de sistem**	**identificate de sistem**	**recall (%)**	**precizie (%)**
prep	1338	1334	1342	99.7	99.4
punct	1325	1277	1287	96.38	99.22

pmod	1165	1151	1167	98.8	98.63
amod	857	853	856	99.53	99.65
subj	645	629	630	97.52	99.84
nmod	610	609	613	99.84	99.35
ROOT	553	548	693	99.1	79.08
dobj	513	507	512	98.83	99.02

Tabel 5.4. Recall-ul şi precizia de identificare a etichetei relaţiei de dependenţă pentru cele mai frecvente relaţii din setul de evaluare

relaţia	**gold**	**identificate corect de sistem**	**dentificate de sistem**	**recall (%)**	**precizie (%)**
prep	1338	1334	1342	99.7	99.4
punct	1325	1250	1287	94.34	97.13
pmod	1165	1128	1167	96.82	96.66
amod	857	851	856	99.3	99.42
subj	645	628	630	97.36	99.68
nmod	610	606	613	99.34	98.86
ROOT	553	548	693	99.1	79.08
dobj	513	507	512	98.83	99.02

Tabel 5.5. Recall-ul şi precizia de identificare a etichetei şi centrului relaţiei de dependenţă pentru cele mai frecvente relaţii din setul de evaluare

Eroarea semnalată iniţial, apariţia unui număr mai mare de rădăcini pentru o singură propoziţie, este surprinsă de valorile preciziei pentru relaţia ROOT: 79.08%, cea mai mică precizie din Tabelele 5.4. şi 5.5, cu 693 de ROOT-uri identificate de sistem, dintre care doar 548 sunt corecte.

Pentru relaţia *prep*, valorile egale pentru recall în ambele tabele ne indică faptul că nu există relaţii prep care să fie identificate ca atare, dar să fie ataşate unui centru greşit. În plus, recall-ul pentru această relaţie este foarte bun, 99,7%, cu doar 4 erori din 1338 de apariţii ale acestei relaţii în setul evaluat.

În schimb, despre relaţia *punct* se poate spune că în 27 (1277-1250, vezi coloana 3 din ambele tabele) din cazuri, punctuaţia nu a fost ataşată corect centrului său, chiar dacă a fost identificat corect tipul relaţiei. Acest lucru este reflectat şi în altă secţiune a fişierului cu rezultate

de evaluare, unde este sistematizată distribuţia erorilor conform părţii de vorbire a cuvântului adnotat: 31 de semne de punctuaţie din setul de analiză purtând eticheta morfo-lexicală COMMA (asociată virgulei) au centrul ataşat greşit; dintre acestea, 10 au eticheta dependenţei sintactice adnotată greşit (adică ROOT), iar celelate reprezintă 21 dintre cele 27 situaţii în care relaţia *punct* are eticheta corectă, dar nu şi centrul corect identificat. Celelalte 6 situaţii se datorează punctului (eticheta morfo-lexicală PERIOD): 38 de elemente etichetate PERIOD au centrul ataşat greşit, dintre care 32 sunt etichetate greşit drept ROOT, iar restul de şase au eticheta ataşată corect, conform distribuţiei erorilor pe părţi de vorbire.

În imaginea din Figura 5.2 sunt evidenţiate prin culoare roşie în partea de jos a ecranului două dintre relaţiile de dependenţă ce poartă eticheta punct: după cum se poate vedea în partea superioară, etichetele acestor relaţii sunt corect identificate, dar centrele relaţiilor sunt incorecte. MaltParser a asociat atât virgulele cât şi elementul interpus centrului căruia îi sunt asociate toate elementele precedente din listă; cum toate structurile precedente precum şi cea imediat următoare (“plata pe loc”), separate prin virgule, sunt modificatori pentru obiectul direct al propoziţiei, “apartament”, adnotatorul automat a învăţat să trateze similar modificatorul “limitrof” şi punctuaţia din jurul său. Dar adnotatorul uman, care dispune şi de cunoştinţe enciclopedice suplimentare, apreciază că “limitrof” este un modificator pentru “zona”, iar punctuaţia trebuie ataşată local.

Asemănător, relaţia *pmod* are erori de ataşare a centrului pentru 23 dintre relaţiile pentru care eticheta sa a fost corect identificată. Acest tip de eroare se datorează uneia dintre cele mai mari provocări pentru analizoarele sintactice automate: ambiguitatea de ataşare a grupului prepoziţional. De exemplu, în Figura 5.3, pentru sintagma “sejururile personalizate în străinătate”, este nevoie de intervenţia umană pentru a decide ataşamentul corect al modificatorului prepoziţional “în străinătate”; parserul alege drept centru modificatorul “personalizate” (etichetat morfo-lexical drept verb la modul participiu), care este mai apropiat şi care în gramatica de dependenţe utilizată de noi introduce de fapt o subordonată adjectivală (“acl”). Adnotatorul automat este astfel înclinat să considere că se află în cadrul limitelor unei noi propoziţii şi că “în străinătate” este modificator pentru centrul acestei propoziţii, aşa cum a învăţat din datele de antrenare că este cel mai probabil. În figura 5.4, adnotatorul automat foloseşte din nou criteriul vecinătăţii pentru a ataşa grupul prepoziţional “în 7 aprilie 1924”, iar decizia umană a identificat verbul “întâmpinaţi” drept centru corect pentru această relaţie.

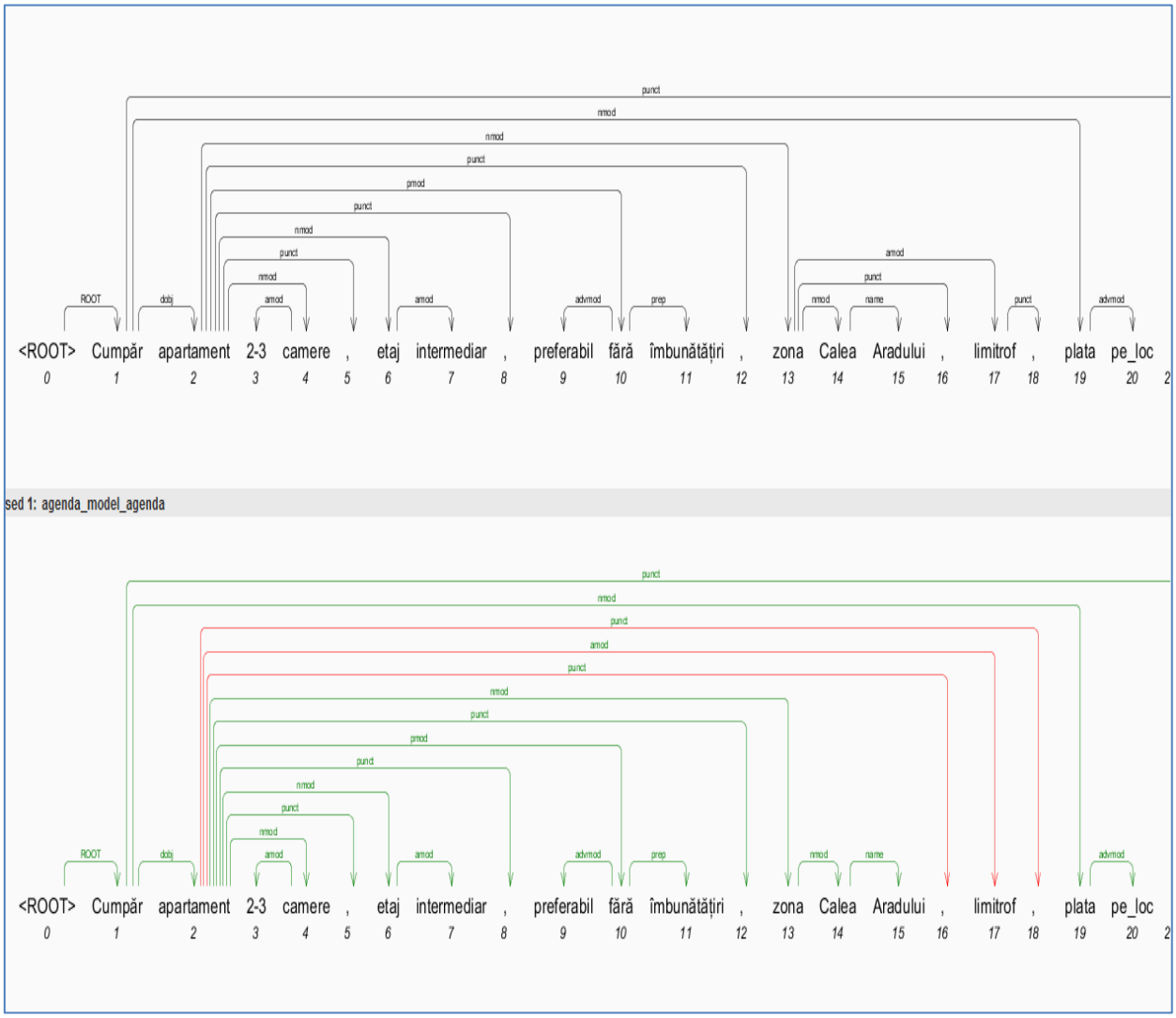


Figura 5.2. Exemplu de eroare pentru identificarea centrului relaţiei punct

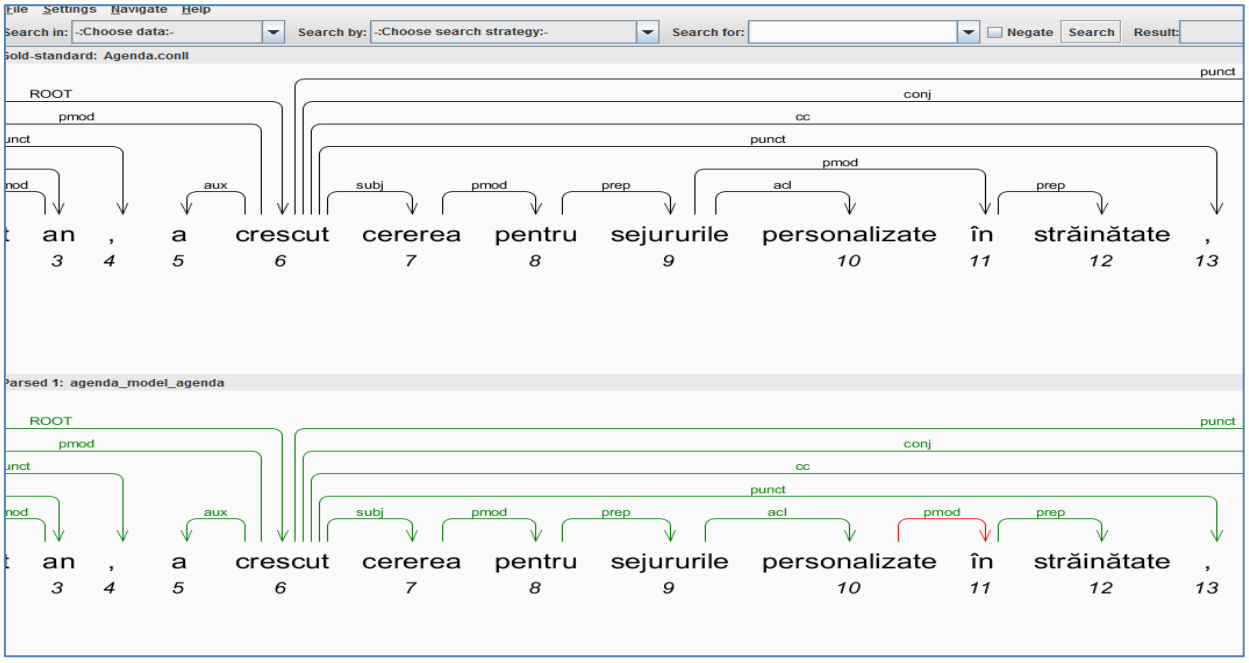


Figura 5.3. Exemplu de eroare pentru identificarea centrului relaţiei pmod.

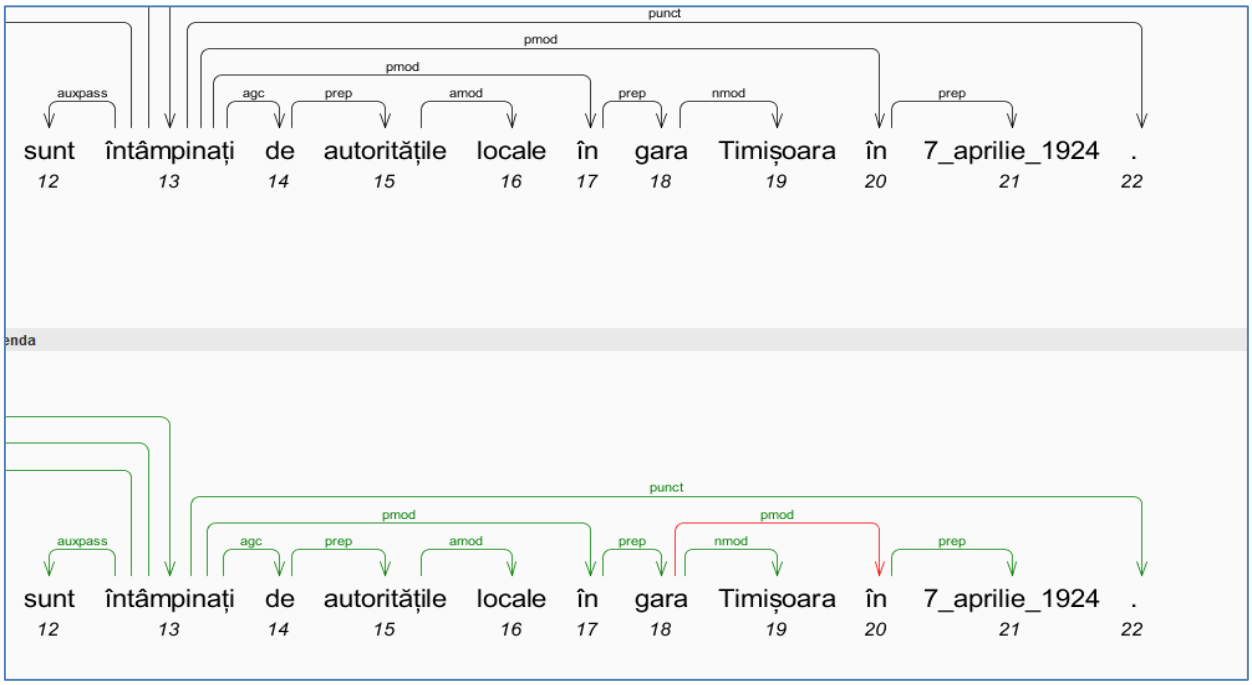


Figura 5.4. Exemplu de eroare pentru identificarea centrului relaţiei pmod.

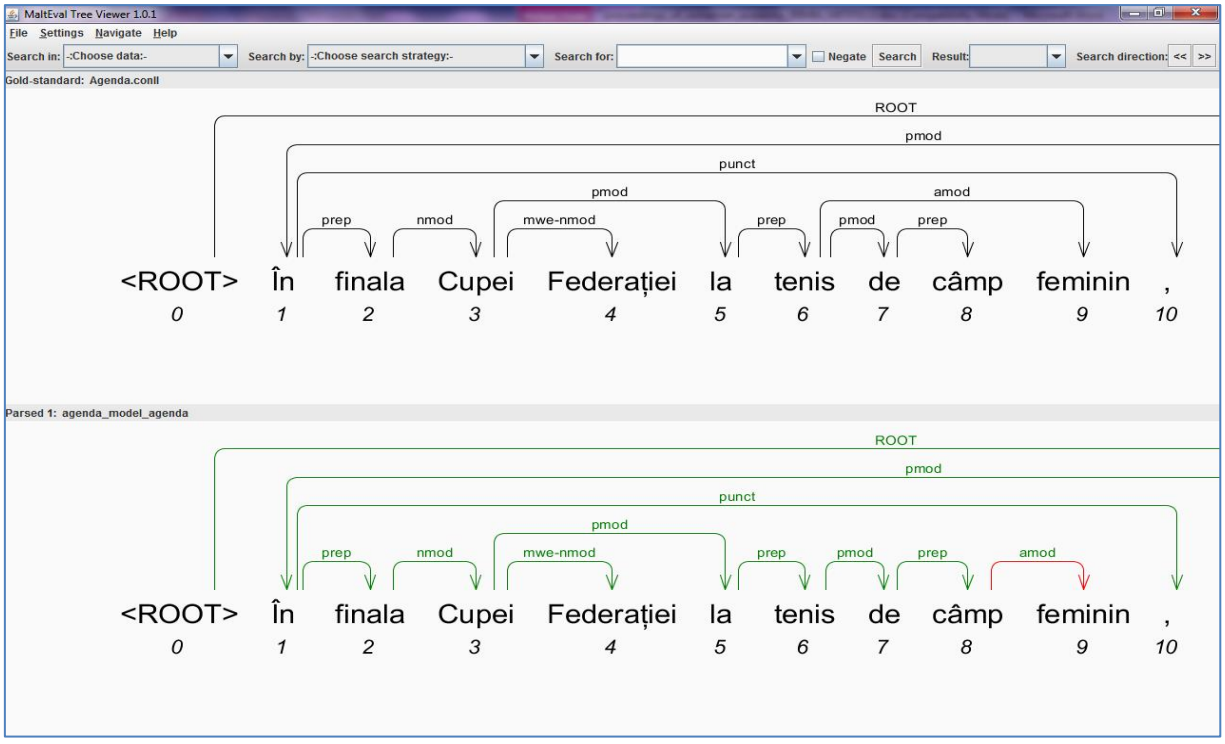


Figura 5.5. Exemplu de eroare pentru identificarea centrului relaţiei amod.

Relaţia *amod*, care identifică modificatorii adjectivali ai unui grup nominal, este corect adnotată în proporţie foarte mare (99.3% dintre relaţii au centrul şi eticheta corect identificată); totuşi cele două situaţii în care centrul nu este corect identificat (853-851, vezi coloana 3) sunt exemple ale unui alt tip de ambiguitate, mai rară în limbă, dar care necesită decizia adnotatorului uman: ambiguitatea de ataşare a modificatorului adjectival. În exemplul din Figura 5.5, observăm că avem de a face cu ataşarea modificatorului la un termen al unei expresii multi-cuvânt (tenis de câmp) şi că adnotatorul automat nu poate identifica constituentul din expresie la care trebuie ataşat modificatorul (“tenis”) şi îl ataşează celui mai apropiat substantiv (“câmp”).

Tabelele 5.4 şi 5.5 semnalează și un alt tip de eroare ne-sistematică: ataşarea grupului nominal (vezi relaţia *nmod*). În exemplul din Figura 5.6, doar adnotatorul uman poate identifica în mod corect centrul (substantivul “menţinerea”) pentru modificatorul nominal “preşedintelui”, în timp ce adnotatorul automat alege, conform criteriului vecinătăţii, un centru greşit (“funcţie”).

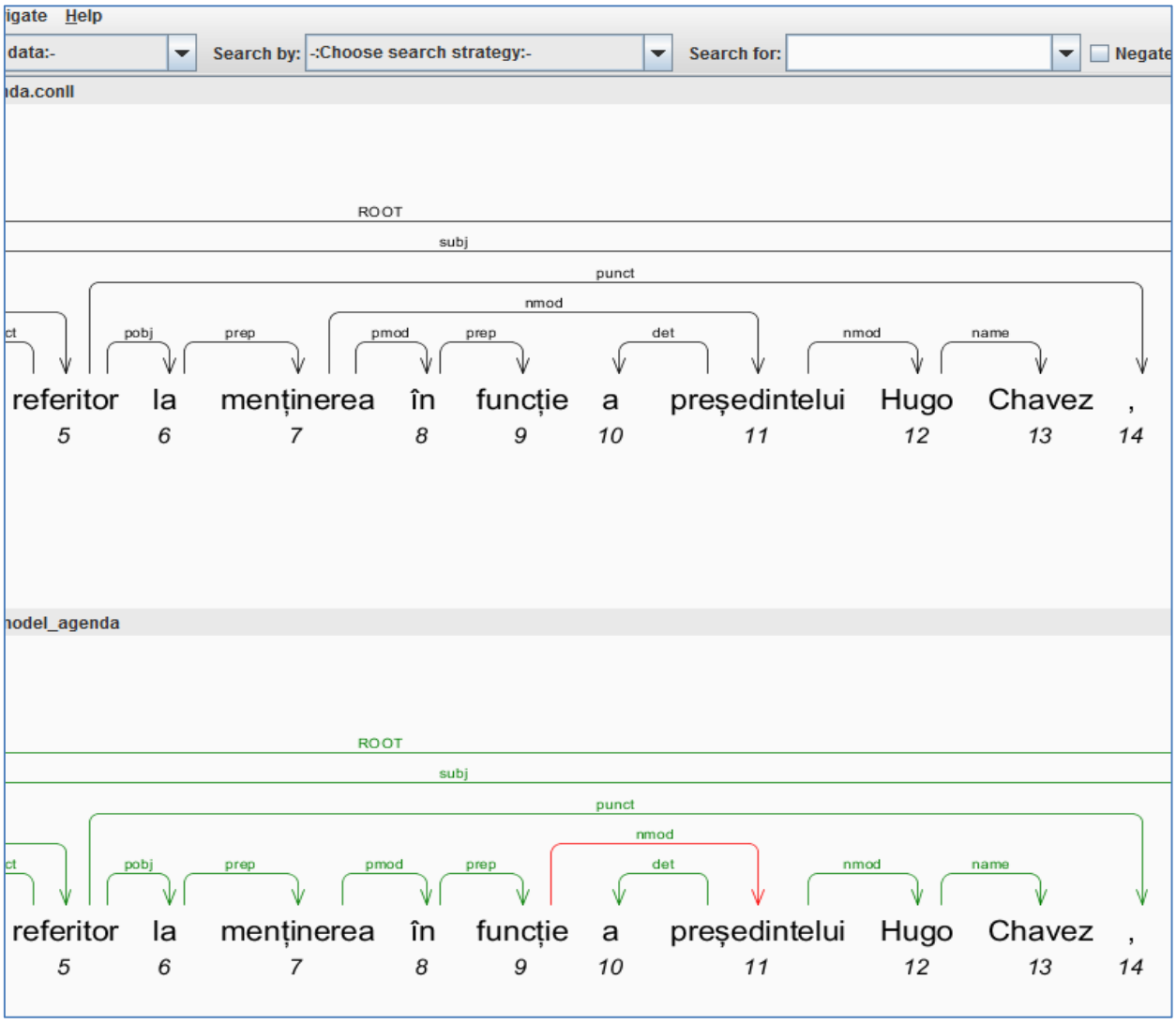


Figura 5.6. Exemplu de eroare pentru identificarea centrului relaţiei nmod

Prin capacitatea de a scoate în evidenţă erorile ne-sistematice, această metodologie ne ajută să identificăm şi erorile umane produse în timpul etapei de corectură manuală. De exemplu, în Figura 5.7. se poate vedea cum a fost descoperită o eroare în gold-standard: din neatenţie, adnotatorul uman a etichetat drept *pobj* o relaţie de tip *punct*.

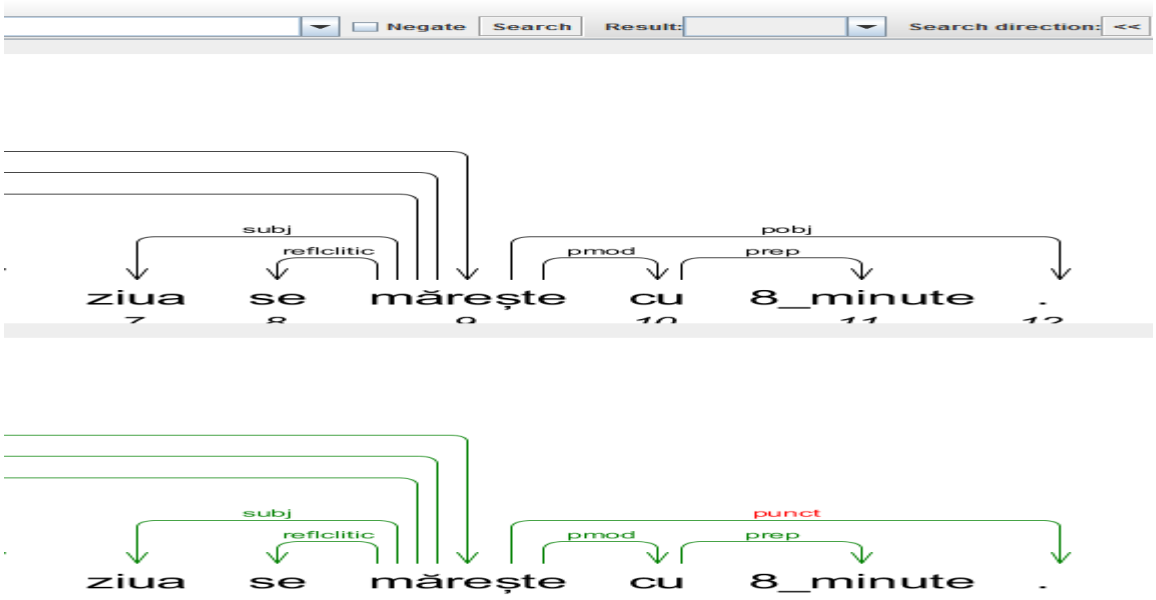


Figura 5.7. Exemplu de eroare în gold-standard

Acest experiment de evaluare distorsionată a avut loc într-o fază iniţială de corectare şi nu a fost repetat între timp, dar considerăm că rezultatele sunt similare şi pe alte tipuri de date, deoarece stilul publicistic este unul dintre cele mai versatile, folosind sintaxă şi vocabular mai puţin controlate decât stiluri literare oficiale. Totuşi, într-o etapă ulterioară finalizării acestui proiect, intenţionăm să folosim această metodologie pe întreg treebank-ul pentru a identifica eventualele erori de adnotare umană.

La momentul experimentului, am interpretat rezultatele acestuia ca fiind un bun indicator de a continua pe acest drum, din moment ce erorile ne-sistematice sunt într-un procent mic (mai puţin de 3%), iar numărul erorilor sistematice poate fi redus, ca în cazul oricărui experiment statistic, prin adăugarea de exemple diverse şi adnotate corect şi consistent.

5.2.2. Evoluţia erorilor sistematice în timpul ciclului de adnotare/corectare/re-antrenare

Pentru studiul erorilor sistematice, acele tipuri de erori care ar trebui să fie din ce în ce mai puțin întâlnite pe măsură ce modelul statistic devine mai cuprinzător, am urmărit evoluția erorilor de-a lungul procesului iterativ de adnotare/corectare/re-antrenare. Informațiile cuprinse în fișierele cu rezultate de evaluare generate după fiecare iterație a modelului statistic sunt redate

(parțial) în Anexa 5, separat pentru fiecare etapă în parte. În continuare vom sintetiza și interpreta aceste informații statistice pentru a da seamă de creșterea mai mică sau mai mare a performanțelor modelului statistic în funcție de eticheta morfo-lexicală a cuvintelor adnotate sau de tipul relației de dependențe identificat, pe parcursul procesului de construire a treebank-ului.

Tabelul 5.6. prezintă distribuţia acurateții de identificare a tipului relației de dependență și a centrului său pe părţi de vorbire: de exemplu, raportul dintre numărul de adjective cu tip de relație și centru corect identificate şi numărul total de adjective din setul de date evaluat. Pe axa verticală a tabelului se regăsesc părțile de vorbire evaluate (în ordine alfabetică, adjective, adverbe, conjuncții coordonatoare, conjucții subordonatoare, particule (de negație, de infinitiv, de conjunctiv), prepoziții, pronume (demonstrative, interogativ-relative, personale, reflexive, indefinite), substantive comune și proprii, verbe predicative), iar pe axa orizontală tipurile de date evaluate în iterațiile succesive ale procesului de adnotare (jurnalistic, literar, academic, medical, juridic). După cum se poate observa, lipsesc părți de vorbire precum articolele, determinanții și verbele auxiliare, care au o evoluție foarte bună încă de la al doilea set de date evaluat (Acad.1), cu acurateţe pornind de la valori medii de 80% și ajungând la valori de 100% la ultimul set evaluat (Jurid. 1). Aceste valori pot fi consultate în Anexa 5. Am prezentat în Tabelul 5.6 părți de vorbire care pornesc cu acurateţe de adnotare scăzută și au o evoluție lentă a acurateții pe parcursul procesului.

De asemenea, lipsesc din tabel valorile acurateții pentru primul set de propoziții adnotate (tranșa 1 din secțiunea jurnalistică, Jurnalistic 1): aceste date nu sunt relevante pentru evoluția modelului de limbă română, din moment ce corectarea manuală s-a făcut pe date adnotate cu modelul statistic spaniol. Pentru acest set, rezultatele detaliate nu sunt utile, mai ales pentru că se lucrează cu două seturi de etichete de dependențe diferite, IULAdep și ROdep, ceea ce conduce la un număr mare de erori în termenii oricărui scor de evaluare (LAS, acurateţe, recall, precizie). Rezultatele evaluării performanțelor după această primă etapă de corectare precum și concluziile pe care le-am tras în momentul respectiv sunt detaliate în secțiunea 5.2.1. În continuare, în nici unul dintre tabelele prezentate nu vom face referire la setul de propoziții Jurnalistic 1.

După cum se poate vedea atât în Tabelul 5.6 cât și în diagrama corespunzătoare din Figura 5.8, valorile acurateții cresc pentru fiecare secțiune de la o tranșă la alta (de exemplu, de la Literar 1 la Literar 2, de la Academic 1 la Academic 2 etc.) pentru toate părțile de vorbire examinate. În schimb, în unele situații, valorile scad când se trece de la o secțiune la alta (de la un stil literar la altul). Toate dreptele sunt ascendente la trecerea de la secțiunea literară la cea academică, în ambele tranșe: acest fenomen poate fi explicat prin faptul că stilul academic din corpusul nostru pare asemănător celui literar, dar este mai omogen și controlat, punând mai

puține probleme analizorului sintactic automat. În schimb, la trecerea de la stilul academic la cel medical, cele mai multe dintre părțile de vorbire au o creștere, chiar dacă modestă, cu excepția pronumelor (personal, interogativ-relativ și reflexiv) și al adverbelor.

După cum am menționat și explicat în secțiunea 5.1, la trecerea la secțiunea juridică, în tranșa 1, are loc o scădere a performanței generale, reflectată în pantele descendente ale acurateții majorității părților de vorbire în Figura 5.7. Cea mai eterogenă evoluție are loc la trecerea de la secțiunea Juridic 1 la secțiunea Jurnalistic 2, când adnotatorul automat s-a confruntat pentru prima dată cu propoziții lungi, cu lungimea cuprinsă între 20 și 40 de cuvinte, lungimea maximă până în acel moment fiind de 20 de cuvinte pe propoziție (după cum am explicat în secțiunea 4.2).

În continuare, la trecerile de la o secțiune la alta pentru tranșa a doua de propoziții, observăm fluctuații moderate, explicabile, ca și în prima etapă, prin schimbarea stilului funcțional și, implicit, a manierei de a structura sintactic propozițiile.

Pentru a face și mai clar faptul că acuratețea este în creștere atât de la o tranșă la alta în cadrul aceluiași set, cât și de la prima la ultima evaluare, Figura 5.9 transpune grafic aceleași cifre din Tabelul 5.6, dar de această dată nu în ordinea iterațiilor de adnotare/corectare/re-antrenare, ci ordonat pe secțiuni. Am exclus complet din Figura 5.9 secțiunea jurnalistică: este nerelevantă, deoarece nu avem informații decât despre tranșa 2, care nu are un termen de comparație, prin absența tranșei 1. Se observă creșterea pantelor pentru trecerea de la o tranșă la alta în fiecare secțiune, dar și faptul că punctul final al evoluției acurateței pentru fiecare parte de vorbire este superior punctului de plecare pe axa verticală (Juridic 2 versus Literar 1).

	Lit. 1	**Acad. 1**	**Med. 1**	**Jurid. 1**	**Jurn. 2**	**Lit. 2**	**Acad. 2**	**Med. 2**	**Jurid. 2**
Adjective	63	75	81	70	79	77	84	87	88
Adverbe	55	70	53	75	65	74	80	70	87
Conj. Coordonatoare	38	55	60	53	48	56	59	63	88
Conjuncții subordonatoare	37	52	56	50	53	58	67	60	70
Particule	54	82	84	78	78	75	82	92	87
Prepoziții	55	66	71	65	67	68	75	75	79
Pron. Demonstrative	38	54	75	73	69	61	75	88	75
Pron. Interogativ-	84	96	87	96	84	88	98	92	99

Relative									
Pron. Personale	37	71	50	50	68	65	95	78	66
Pron. Reflexive	57	71	63	71	71	84	88	72	77
Pron. Indefinite	44	75	75	88	63	68	81	79	91
Subst. Comune	69	79	82	80	80	84	82	87	89
Subst. Proprii	72	77	83	70	78	77	89	89	85
Verbe predicative	56	69	74	68	68	75	79	84	85

Tabelul 5.6. Distribuția pe partea de vorbire a acurateții de identificare a centrului și etichetei relației de dependență

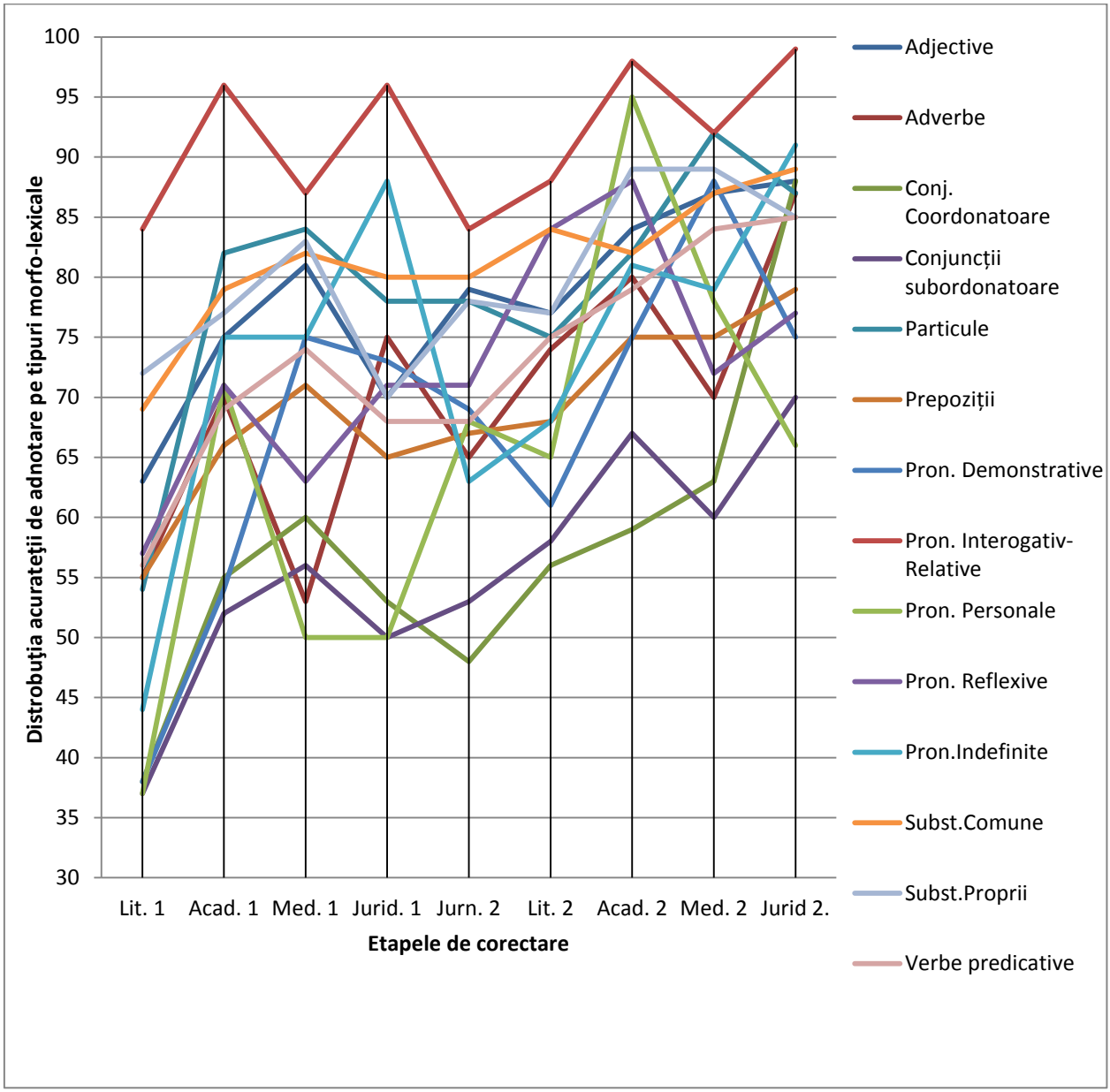


Figura 5.8. Diagrama corespunzătoare distribuției pe partea de vorbire a acurateții de identificare a centrului și etichetei relației de dependență

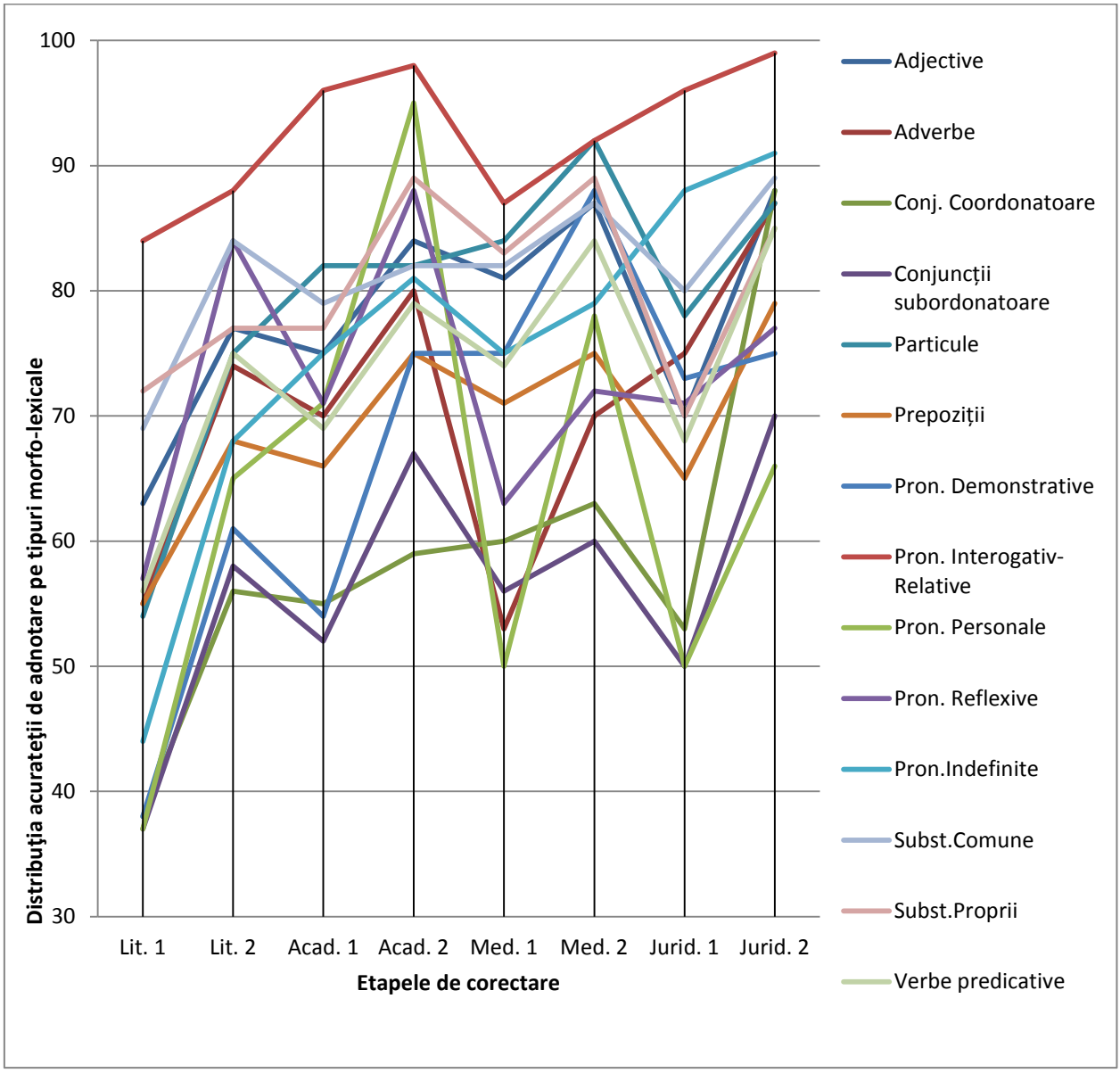


Figura 5.9. Diagrama corespunzătoare distribuției pe partea de vorbire a acurateții de identificare a centrului și etichetei relației de dependent; ordonarea etapelor de corectare pe secțiuni.

Informaţia legată de distribuţia recall-ului pe tipul de relaţie de dependenţă (pentru centru şi tip de relaţie corect identificate) a fost distribuită în mai multe tabele şi diagrame corespunzătoare, deoarece aglomerarea acestora într-un singur tabel şi o singură diagramă îngreunează foarte mult interpretarea datelor. Astfel, Tabelul 5.7 prezintă tipurile de relaţie de dependenţă care la prima tranşă evaluată (Literar 1) pornesc cu recall mic, cuprins (aproximativ) între 0% şi 50%, Tabelul 5.8 pe cele cu recall între 50 şi 70% iar Tabelul 5.9 pe cele cu recall între 70% şi 100%. Când un anumit tip de relație se încadrează, ca scor, într-un anumit tabel dar ca tip de comportament în alt tabel, am decis să clasificăm relația conform comportamentului, mai ales când scorul este foarte aproape de limita impusă pentru clasificare.

	Lit. 1	Acad. 1	Med. 1	Jurid. 1	Jurn. 2	Lit. 2	Acad. 2	Med. 2	Jurid 2.
appos	0	12.77	23.08	18.18	34.82	25.58	29.46	29.09	22.22
pobj	4.23	9.26	20.83	36.36	29.55	23.94	12.5	26.42	50
spe	11.11	20	20	20	14.29	36.36	37.02	42.8	50
advcl	12.73	48.89	33.04	39.62	45.29	44.38	63.8	63.1	66.67
post	18.75	16.67	100	30	80.22	50	80	71.43	75.33
pred	21.05	51.06	59.38	87.8	46.84	53.41	60.19	80.87	83.33
parataxis	21.43	2.5	7.59	7.41	26.09	42.86	28.57	17.8	82.76
iobj	22.12	42.86	57.89	51.16	36.78	49	52.78	70.89	56.52
conj	37.03	45.27	48.52	45.75	44.11	60.63	59.9	51.95	70.93
acl	37.58	49.28	61.59	53.75	44.29	60.54	51.3	69.38	65.12
cc	41.24	58.18	55.33	53.41	52.38	61.94	66.5	56.68	73.61
reflclitic	42.57	79.1	71.43	68.57	59.09	83.67	75.59	88.46	100

Tabelul 5.7. Distribuția recall-ului pentru centru și etichetă corect identificate pe tipuri de relație de dependenţă: relații care pornesc cu recall mic, 0-50%.

Printre relaţiile cele mai dificil de identificat, care pornesc cu un recall mic şi au o evoluţie modestă pe parcursul reantrenării, se numără:

- *appos*, o relaţie pe care adnotatorul o poate confunda uşor cu un conjunct, cu un tip de modificator (de multe ori *nmod*), iar în etapele finale de corectură cu *parataxis*; confuzia se datorează faptului că în apoziție se pot regăsi cele mai multe părți de vorbire (toate cuvintele conținut) și este foarte dificil pentru parser să sistematizeze această varietate de situații și posibilități;
- *pobj*: distincţia de *pmod* se face foarte dificil, pentru că exemplele de verbe care reclamă o anumită prepoziţie nu sunt suficiente în corpus; o altă soluţie (în afară de adăugarea unor exemple concludente) ar putea fi incorporarea de informaţie lexicală în parser, adică specificarea formală că un anumit verb cere o anumită prepoziţie, iar relaţia dintre ele este de tip *pobj*; în plus, chiar dacă am implementa una dintre cele două soluții, ne-am putea izbi de problema verbelor care pot avea două grupuri prepoziționale, introduse chiar de aceeași prepoziție (de ex.: „Mă gândesc la vacanță la masă”);
- *spe*: elementul predicativ suplimentar adjectival poate fi adesea confundat cu un complement circumstanțial exprimat prin adjectiv (etichetat eronat drept *amod*, ex. “situația se prezenta *gravă*” “să fie considerată *reușită*”) pe când cel substantival poate fi confundat cu un obiect direct (ex.: “S-a întors *bărbat*.”, “Fusese numit *subdirector*.”). Când urmează unui verb la participiu (ex.:”numită *Ioana*”) este tratat eronat drept *nmod*.

- subordonatele circumstanţiale (*advcl*) – introduse fie prin conjuncții (că, “dacă”, “deşi”), diverse locuțiuni conjuncționale (“indiferent dacă”, “pentru că”, “pentru ca”, “în timp ce”, “fără să”, “cu toate că” etc.), prepoziții (“spre”, “pentru” urmate de verbe la infinitiv), verbe la participiu sau gerunziu, sau verbe la moduri predicative când subordonata include un pronume relativ – au un comportament destul de eterogen, care poate fi învățat de analizor, dar într-un ritm mai lent: scorurile pentru ultimele seturi evaluate sunt în medie de 65%; neidentificarea consecventă de către analizorul morfo-lexical a locuțiunilor conjuncționale introduce şi ea o serie de erori;
- recall-ul relației *post* are fluctuații mari, dar per total creşterea este de aproximativ 50%; de cele mai multe ori, această relație este adnotată eronat drept *pmod* (cum adnotatorul învață că trebuie să trateze prepozițiile), dar, treptat, situațiile în care prepoziția este de fapt postpusă încep să se evidențieze (într-un procent covârşitor, este vorba de prepoziția “de”, atunci când apare în structuri de felul “30 *de* mii *de* lei”, sau de felul “astfel de”, “atât de” etc.)
- relația *pred* este şi ea fluctuantă, dar per total are o creştere de aproximativ 60%: acest comportament se datorează faptului că adnotatorul învață foarte repede că un nume predicativ urmează verbului copulativ “a fi”, dar dispune de puține exemple pentru alte verbe copulative; confuzia cea mai frecventă este cu obiectul direct (*dobj*);
- *parataxis*: e foarte dificil pentru adnotatorul automat să distingă între vorbirea directă şi vorbirea indirectă în propoziţie; de cele mai multe ori vorbirea indirectă este tratată ca propoziţie coordonată; pentru celelalte situaţii în care am folosit eticheta, precum structurile care identifică legile şi articolele acestora în secţiunea juridică, adnotatorul a învăţat foarte uşor să atribuie relaţia corectă: vezi creşterea de la tranşa 1 la tranşa 2 pentru secţiunea juridic corespunzătoare relaţiei *parataxis* în Tabelul 5.7 şi diagrama corespunzătoare;
- *iobj*: adesea confundat cu *dobj*, datorită omonimiei acuzativ/dativ pentru pronume; confundat şi cu *dblclitic*: parserul adnotează *dblclitic* pronumele în dativ, chiar dacă un alt dependent adnotat drept *iobj* nu există în propoziţie, caz în care ar trebui să-i acorde pronumelui această etichetă;
- relațiile *conj* și *cc* suferă o creştere lentă a recall-ului: este dificil pentru parser să adnoteze corect coordonările la distanță, în special când între conjuncți se interpun apoziții, propoziții subordonate, liste de tip adresă, modificatori ai conjuncților etc.
- subordonata atributivă (*acl*) are un comportament asemănător cu cea circumstanţială (*advcl*), cu care uneori o și confundă, selectând centrul subordonatei pe cel mai

apropiat verb; ea poate fi introdusă de verbe la participiu şi gerunziu, poate fi o subordonată relativă, se poate afla adeseori la distanță de centrul său; în plus, ca pentru orice altă subordonată, ieşirea din limitele propoziției presupune un grad superior de complexitate sintactică;

- creşterea spectaculoasă a recall-ului relaţiei *reflclitic* în ultimele două seturi de date analizate, asociate cu scăderea recall-ului relaţiei *passmark* în aceleaşi seturi (vezi Tabelul 5.8), se explică prin confuzia pe care o face parserul între diferitele valori ale pronumelui „se”. Valorile mici ale preciziei (vezi Anexa 5) pentru *reflclitic* în seturile respective sugerează că parserul a adnotat drept *reflclitic* multe dintre ocurenţele lui “se” cu valoare de *passmark*.

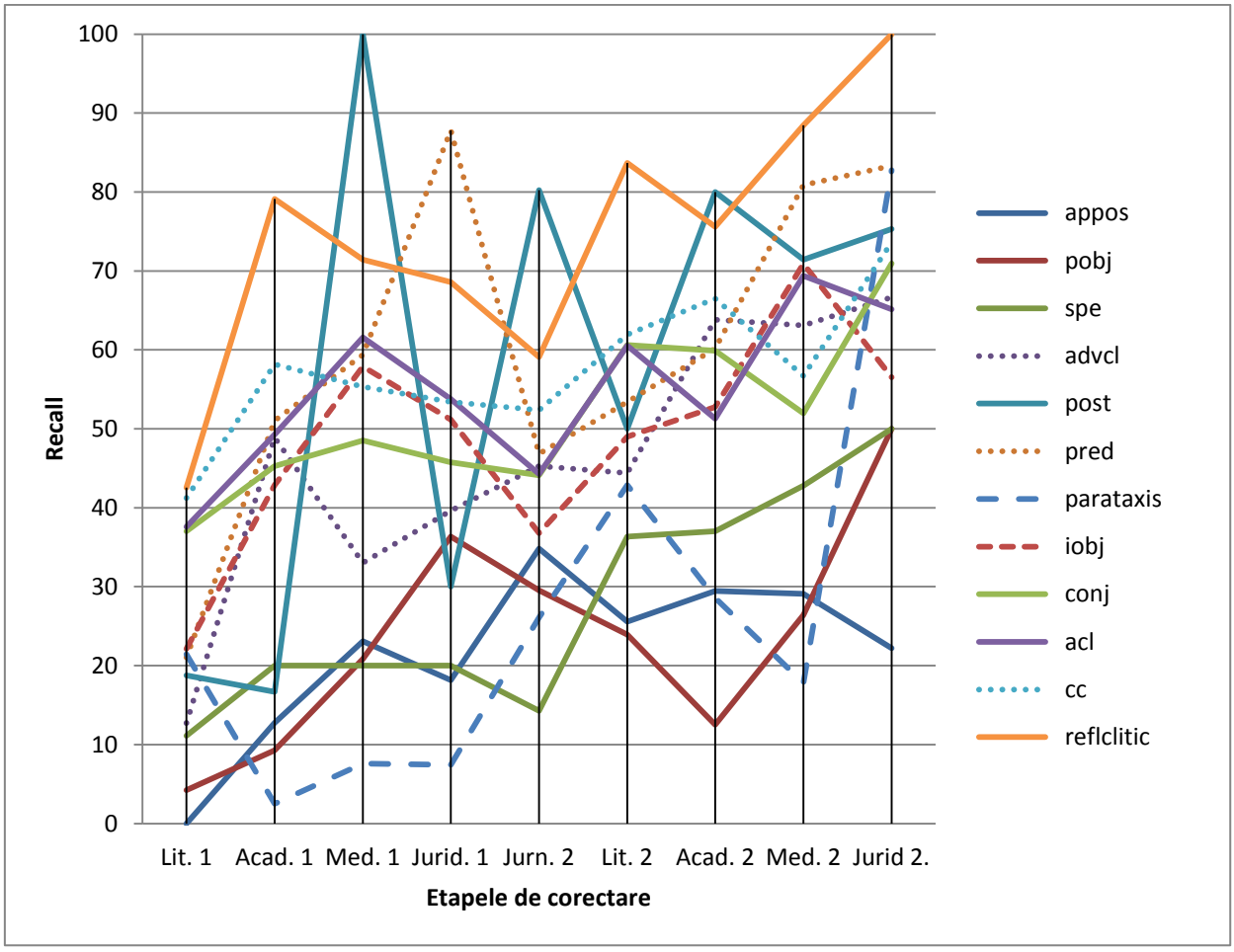


Figura 5.9. Diagrama corespunzătoare Tabelului 5.7. Distribuția recall-ului pentru centru și etichetă corect identificate pe tipuri de relație de dependență: relații care pornesc cu recall mic: 0-50%.

În Tabelul 5.8, şi cu atât mai clar în diagrama corespunzătoare (Figura 5.10) se poate observa că relaţiile care pornesc cu un recall mediu, cuprins între 50% şi 70% fluctuează în prima parte a procesului iterativ (cele mai mari scăderi la trecerea la primul set de date juridice,

din motivul deja explicat), dar au o creştere uşoară şi constantă în ultimele etape ale acestuia, depăşind valori ale recall-ului de 80%. Excepţie fac relaţiile *passmark* şi *dblclitic*, care sunt confundate adeseori cu *reflclitic*, respectiv cu *dobj* şi *iobj* atunci când acestea din urmă se realizează prin pronume în formă neaccentuată.

	Lit. 1	**Acad. 1**	**Med. 1**	**Jurid. 1**	**Jurn. 2**	**Lit. 2**	**Acad. 2**	**Med. 2**	**Jurid 2.**
advmod	49.34	64.47	51.2	76.53	59.09	68.36	74.81	75.51	77.33
subj	50.11	73.9	70.53	70.03	68.34	71.27	80.53	81.06	90.52
dblclitic	52.54	70.59	50	83.33	78.13	64.44	70.59	87.5	84.14
dobj	52.81	76.88	82.19	80.51	79.01	75.57	83.52	85.84	89.89
punct	53.57	69.36	79.44	70.87	68.77	79.44	82.59	85.96	90.17
poss	59.09	86.36	62.5	80	78.95	79.49	94.12	100	100
agc	60	55.88	76.47	46.15	57.5	68.97	71.13	85.71	90
sc	61.04	75.36	72.51	68.29	72.92	86.85	84.38	91.97	92.59
pmod	65.17	75.54	78.13	70.27	73.01	74.6	77.9	79.94	81.38
passmark	65.96	47.22	57.81	42.47	77.78	78.38	62.32	74.53	68.57
nmod	67.74	89.43	80.73	79.94	83.54	86.56	90.41	84.11	92.03
amod	68.02	85.93	91.3	75.86	91.34	89.28	94.66	92	93.86

Tabelul 5.8. Distribuția recall-ului pentru centru și etichetă corect identificate pe tipuri de relație de dependență: relații care pornesc cu recall mediu, 50-70%.

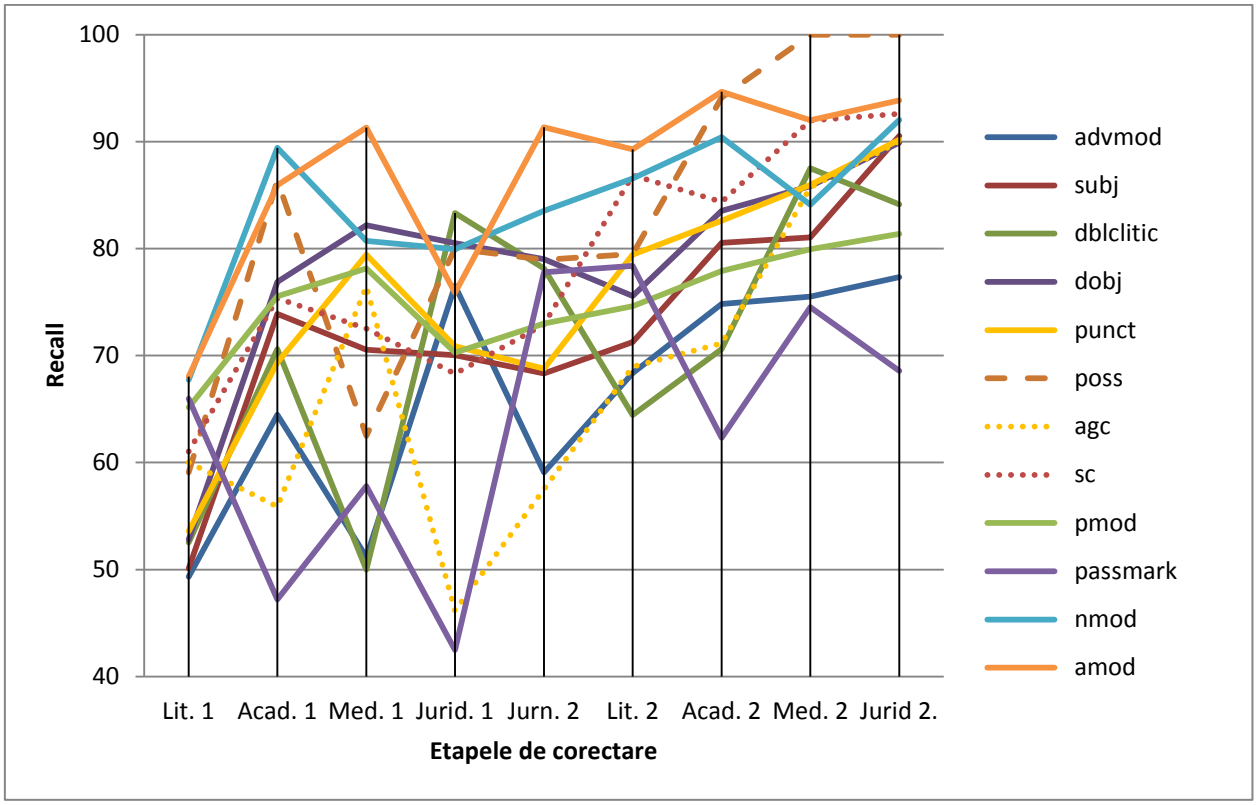


Figura 5.10. Diagrama corespunzătoare Tabelului 5.8. Distribuția recall-ului pentru centru și etichetă corect identificate pe tipuri de relație de dependență: relații care pornesc cu recall mediu: 50-70%.

Tabelul 5.9 şi diagrama corespunzătoare din Figura 5.11 cuprind tipuri de relaţii uşor de identificat de parser. Trebuie precizat că după evaluarea prezentată în secţiunea 5.2.1. şi identificarea comportamentului lui MaltParser de a genera mai multe rădăcini pentru o singură propoziţie analizată (pentru un procent din numărul propoziţiilor analizate care scade pe măsură ce modelul statistic devine mai performant), am decis corectarea acestui tip de erori manual, pe formatul CONLL, înainte de a trece la formatul GRAPHML, care nu permite mai mult de o rădăcină pentru un arbore. Evaluarea datelor s-a făcut luând drept fişiere de test aceste fişiere corectate: drept urmare, evoluţia recall-ului pentru eticheta ROOT, chiar dacă distorsionată (pentru că un procent dintre propoziţii au rădăcina corectată manual) surprinde totuşi capacitatea adnotatorului de a identifica rădăcina corectă pentru celelalte propoziţii (atunci când nu supragenerează noduri rădăcină).

Celelalte tipuri de relaţii din Tabelul 5.9 pornesc cu valori mari ale recall-ului şi cresc până la peste 85% pentru că implică proprietăţi morfologice sau topice ale cuvintelor analizate identificabile cu uşurinţă:

- cuvintele care primesc relaţia *name* sunt scrise cu majusculă, au etichetă morfo-lexicală Np şi apar întotdeauna după un alt cuvânt care are aceleaşi proprietăţi; datorită numărului foarte mic de nume proprii în secţiunea juridic, putem spune că scăderea recall-ului pentru această relaţie este nerelevantă statistic.
- cuvintele care primesc eticheta *mark* sunt particulele de conjunctiv şi infinitiv, identificabile prin etichete morfo-lexicale precise (Qn, Qs), particula de supin (“de” sau altă prepoziţie, înainte de un verb la participiu), elementele care compun gradele de comparaţie (“cea”, “mai”, “foarte” etc.) atunci când preced un adjectiv;
- relaţia *neg* este rezervată particulei de negaţie, identificată prin eticheta morfo-lexicală Qz;
- verbele auxiliare, care primesc eticheta de dependenţă *aux*, sunt identificate prin eticheta morfo-lexicală (variaţiuni pentru Va);
- *auxpass* este rezervată verbului „a fi” atunci când este însoţit de un verb la participiu (aici se întâlnesc situaţii de falşi pozitivi, pentru adjective care sunt confundate cu verbe la participiu);
- relaţia *prep* este atribuită cuvintelor care urmează unei prepoziţii; arareori se întâmplă ca centrul grupului prepoziţional să fie precedat de un determinant, adică un adjectiv (în cazul substantivelor) sau marcatori sau auxiliare (în cazul verbelor la infinitiv);
- relaţia *det* este specifică determinatorilor, identificaţi prin etichetele morfo-lexicale care încep cu litera D.

	Lit. 1	Acad. 1	Med. 1	Jurid. 1	Jurn. 2	Lit. 2	Acad. 2	Med. 2	Jurid 2.
neg	68.46	96.3	95.52	100	94.83	95.7	98.31	97.75	100
ROOT	70.11	80.29	82.92	77.86	87.01	84.88	86.09	86.07	90
name	70.83	90.1	80	54.55	87.12	94.29	90.84	92.86	83.33
mark	77.33	98.15	87.14	94.23	84.5	93.23	93.7	91.62	84.62
auxpass	81.25	80	88.79	92	96.43	92.68	95.83	93.72	95.83
aux	89.27	100	95	100	98.7	96.24	99.33	97.37	100
det	90.28	92.46	94.77	92.61	91.92	95.82	99.04	98.84	100
prep	92.3	97.33	97.45	97.69	96.92	96.8	97.96	97.06	100

Tabelul 5.9. Distribuția recall-ului pentru centru și etichetă corect identificate pe tipuri de relație de dependență: relații care pornesc cu recall mediu, 70-100%.

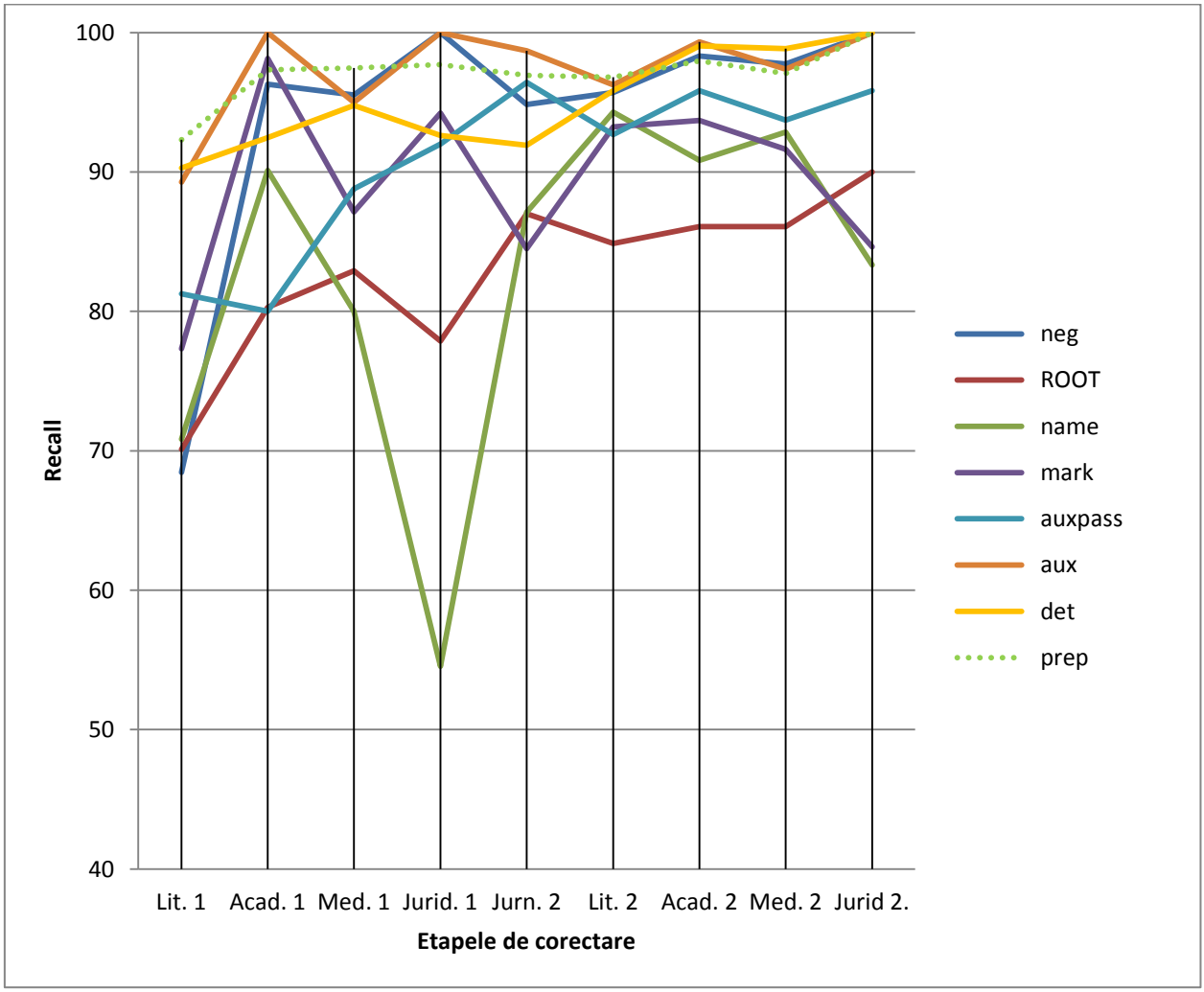


Figura 5.11. Diagrama corespunzătoare Tabelului 5.9. Distribuția recall-ului pentru centru și etichetă corect identificate pe tipuri de relație de dependență: relații care pornesc cu recall mic: 50-70%.

Am omis din tabelele și diagramele acestei secțiuni acele relații care apar sporadic în corpus: *goeswith*, *list*, *remnant*, *secobj*, *xcomp*, *voc*. Prezența acestora doar în anumite secțiuni ale corpusului face imposibil studiul evoluției recall-ului de-a lungul etapelor de corectare. În

plus, datorită numărului foarte mic de ocurențe în corpus, adnotatorul automat nu poate învăța cum să le analizeze corect: responsabilitatea pentru aceste tipuri de relații îi revine în întregime adnotatorului uman. Alte relații, deși incluse în setul pe care l-am proiectat (descris în secțiunea 2.2.4), nu se regăsesc în treebank: *discourse*, *reparandum*, *dislocated*.

CONCLUZII

Pe fondul unui decalaj important în dezvoltarea de tehnologii lingvistice între limba română și limbi europene avantajate tehnologic, precum engleza, ne-am propus să dezvoltăm o resursă digitală importantă: **un nucleu de bancă de arbori sintactici (treebank), adnotați în formalismul gramaticii de dependențe, numărând 5.000 de propoziții**. Selecția propozițiilor care să fie incluse în treebank s-a făcut riguros: s-a pornit de la un corpus de limbă română balansat (care acoperă patru stiluri funcționale și cinci domenii) și au fost alese propoziții care să conțină verbe frecvente în acest corpus. Am urmărit în acest fel să obținem un set de propoziții atât divers, cât și reprezentativ pentru limba română (vezi Secțiunea 4.1.).

Lucrarea de față descrie munca de dezvoltare a acestei resurse, extinsă pe o perioadă de 16 luni. Sunt reflectate atât activitatea de cercetare cât și cea practică.

Au fost parcurse diverse surse bibliografice atât pentru a da seamă de stadiul actual al cercetării și tehnologiei în domeniu (vezi secțiunea 1.2) cât și pentru a descrie evoluția formalismului gramaticii de dependențe în teoriile lingvistice și în aplicațiile practice (secțiunea 2.1.). Ne-am concentrat atenția pe controversa Gramatici de Dependențe (GD) versus Gramatici de Constituenți (GC), concluzionând că GD este un formalism mult mai potrivit pentru aplicații informatice, datorită minimalismului său. Un rezultat util al acestui studiu este crearea unui **inventar de relații de dependență** adaptat tradiției gramaticii limbii române (secțiunea 2.2.), însoțit de un **ghid de adnotare sintactică cu dependențe** (vezi Anexa 1).

Pentru că ne-am dorit încă de la început să creăm o **resursă compatibilă standardelor internaționale în domeniu**, inventarul de relații de dependențe este în mare parte bazat pe specificațiile inițiativei de standardizare Universal Dependencies. Devierile de la aceste caracteristici standard se datorează specificităților gramaticii limbii române, pe care ne-am dorit să le conservăm în această variantă a resursei, fără a face compromisuri cu scopul de a obține o resursă care să se conformeze complet standardelor UD. Totuși, în lunile ce urmează ne propunem să obținem o variantă a treebank-ului nostru complet aliniată la UD, pe care să o și distribuim comunității de cercetare prin intermediul acestei inițiative.

Așa cum ne-am propus în planul inițial al proiectului, am urmărit **automatizarea extensivă a etapelor** sale: de la prelucrarea automată a corpusului sursă, pe baza căruia s-a construit treebank-ul, la adnotarea automată cu un parser statistic, la utilizarea unui instrument cu interfață grafică pentru corectarea manuală a erorilor de adnotare automată, la evaluarea rezultatelor proiectului folosind instrumente consacrate la competiții din domeniu. Resursele și instrumentele folosite sunt descrise detaliat în Capitolul 3 al acestei lucrări.

Stagiul de mobilitate efectuat la Institut Universitari de Lingüística Aplicada (IULA) al universității Pompeu Fabra din Barcelona s-a dovedit foarte oportun: am beneficiat din partea echipei IULA atât de un model statistic de-lexicalizat de analiză cu dependențe antrenat pe un corpus spaniol cât și de experiența acestora în dezvoltarea unui treebank de limbă catalană pe baza acestui model. Astfel, am putut evita să pornim de la zero în adnotarea treebank-ului nostru, strategie ce ar fi presupus un consum serios de timp și resurse umane și am automatizat parțial munca de adnotare. Am folosit modelul de limbă spaniolă pentru a adnota cu parserul disponibil liber MaltParser un set de 500 de propoziții din treebank-ul nostru și am corectat aceste propoziții manual cu instrumentul yEd. Apoi am antrenat un model statistic lexicalizat de limbă română pe cele 500 de propoziții adnotate și am trecut la adnotarea statistică cu acest model. În tranșe de câte 500 de propoziții, am corectat întreg treebank-ul, re-antrenând modelul statistic românesc după fiecare tranșă corectată. **Evoluția performanței de adnotare a acestui model este grăitoare, de la 0,58 pentru prima antrenare la 0,87 pentru ultima**. De altfel, dificultatea muncii de corectare a scăzut în mod evident pe parcursul procesului. În acest moment, cu un model statistic care reduce substanțial munca de corectare manuală, este fezabilă **perspectiva extinderii treebank-ului dezvoltat dincolo de limita de 5.000 de propoziții** pe care ne-am propus-o, mai ales că se urmărește integrarea nivelului de analiză sintactică în corpusul computațional de referință pentru limba română contemporană, CoRoLa, proiect prioritar al Academiei Române.

De asemenea, într-o etapă ulterioară finalizării acestui proiect, intenţionăm să folosim metodologia evaluării distorsionate pe întreg treebank-ul pentru a identifica eventualele erori de adnotare umană și a le corecta. Chiar dacă posibilitatea existenței acestui tip de eroare în treebank-ul nostru a fost redusă datorită implicării în munca de corectare a doi specialiști (cel de-al doilea revizuind munca de corectare a primului), metodologia menționată ne poate ajuta să eliminăm complet eroarea din nucleul de treebank pe care l-am dezvoltat.

Principalele contribuţii ale acestui proiect sunt:

- Dezvoltarea unui nucleu de bancă de arbori pentru limba română divers şi reprezentativ, alcătuit din 5.000 de propoziţii analizate sintactic automat cu relaţii de dependenţă şi corectate manual de către doi specialişti lingvişti;
- Dezvoltarea unui set de relaţii de dependenţă specific limbii române dar aliniabil standardelor internaţionale în domeniu;
- Dezvoltarea unui ghid de adnotare cu exemple corespunzător setului de relaţii de dependenţe stabilit;

- Antrenarea unui model statistic de limbă română cu performanţe bune în raport cu dimensiunea corpusului de antrenare (0,87 scor LAS pentru 4500 de propoziţii de antrenare); folosit cu instrumentul statistic MaltParser, acest model poate servi la adnotarea ulterioară a altor corpusuri de limbă română.

Mulțumiri

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

Autoarea mulțumește călduros domnului academician Florin Gheorghe Filip pentru îndrumarea atentă, pentru sfaturile utile și pentru încurajările oferite în momentele cele mai dificile ale ducerii la bun sfârșit a acestui proiect. De asemenea, sunt recunoscătoare domnului academician Dan Tufiș pentru entuziasmul cu care a îmbrățișat această întreprindere și pentru sprijinul acordat. Colegei mele Dr. Verginica Barbu Mititelu îi datorez curajul de a mă lansa în acest proiect, precum și colaborarea și susținerea la fiecare pas, de la expertiza lingvistică la îndrumările care au ușurat mult îndeplinirea termenilor contractului postdoctoral.

Mulțumiri speciale doamnei Profesoare Nuria Bel de la IULA, Universitatea Pompeu Fabra, Barcelona, care m-a primit cu căldură în echipa sa pe perioda stagiului de mobilitate și, prin ideile și îndrumarea atentă, a ajustat de mai multe ori acest proiect. Experiența de lucru împreună se va prelungi și prin redactarea în perioada următoare a unui articol comun în care vom aborda comparativ dezvoltarea treebank-ului catalan și a celui românesc.

REFERINŢE BIBLIOGRAFICE

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

Academia Română. (2009). *DGLR: Dicţionarul General al Literaturii Române*. Editura Univers Enciclopedic. Vol I-VII. 1993-2009

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

Barbero, C., Lesmo, L., Lombardo, V. and Merlo, P. (1998). Integration of syntactic and lexical information in a hierarchical dependency grammar. *In Kahane, S. and Polguere, A. (eds), Proceedings of the Workshop on Processing of Dependency-Based Grammars* (ACL-COLING), pp. 58–67.

Barbu Mititelu, V., Dumitrescu, Ș.D., Tufiș, D. (2014). News about the Romanian Wordnet. *In Proceedings of the 7th International Global WordNet Conference*. Tartu, Estonia.

Barbu Mititelu, V. and Irimia, E. (2014) The Provisional Structure of the reference Corpus of the Contemporary Romanian Language (CoRoLa). *In Proceedings of the 10th International Conference "Linguistic resources and Tools for Processing the Romanian Language"* (Colhon, M., Iftene, A., Barbu Mititelu, V., Cristea, D. și Tufiș, D. (eds.)). Editura Universității „Alexandru Ioan Cuza", Iași, pp. 57–66.

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

Black, E., Jelinek, F., Lafferty, J., Magerman, D., Mercer, R. and Roukos, S. (1992). Towards history-based grammars: Using richer models for probabilistic parsing. *In Proceedings of the 5th DARPA Speech and Natural Language Workshop*, pp. 31–37.

Bloomfield, L. (1933). *Language.* The University of Chicago Press.

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), pp. 597-620.

Carroll, G. and Charniak, E. (1992). Two experiments on learning probabilistic dependency grammars from corpora, *Technical Report TR-92*, Department of Computer Science, Brown University.

Călăcean, M., Nivre, J. (2009). A Data-Driven Dependency Parser for Romanian, *In Proceedings the Seventh International Workshop on Treebanks and Linguistic Theories*, pp. 65-76.

Chang, C.-C. and Lin, C.-J. (2001). *LIBSVM: A library for support vector machines.* Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

Chen, D., Manning, C.D. (2014) A Fast and Accurate Dependency Parser using Neural Networks. *Proceedings of EMNLP 2014*.

Ciaramita, M., Attardi, G. (2010). Dependency Parsing with Second-Order Feature Maps and Annotated Semantic Information. In H. Bunt, P. Merlo, J. Nivre (eds.), *Trends in Parsing Technology*, Springer, pp. 87-104.

Colhon, M. (2012). Syntactic Translation Patterns from a Parallel Treebank, *Workshop on Computational Linguistics and Natural Language Processing of Balkan Languages*, Balkan Conference in Informatics, pp. 85-88.

Collins, M. (1996). A new statistical parser based on bigram lexical dependencies, *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pp 184-191.

Collins, M. (1999). *Head-driven statistical models for natural language parsing.* Ph.D. thesis, Computer Science Department, University of Pennsylvania.

Covington, M. A. (2001). A fundamental algorithm for dependency parsing, *Proceedings of the 39th Annual ACM Southeast Conference*, pp. 95–102.

Daelemans, W. and Van den Bosch, A. (2005). *Memory-Based Language Processing*. Cambridge University Press.

Debusmann, R. (2000). *An Introduction to Dependency Grammar*. [Online] Disponibil la: http://www.ps.uni-sb.de/~rade/}

De Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen K., Ginter, F., Nivre, J., Manning C. (2014) Universal Stanford Dependencies: A cross-linguistic typology. *In Proceedings of LREC 2014*, Reykjavik, Iceland.

Duchier, D. (1999). Axiomatizing dependency parsing using set constraints. *In Proceedings of the Sixth Meeting on Mathematics of Language*, pp. 115–126.

Duchier, D. (2003). Configuration of labeled trees under lexicalized constraints and principles. *Research on Language and Computation* 1, pp. 307–336.

Earley, J. (1970). En efficient context-free parsing algorithm. *Communications of the ACM 13*. pp. 94–102.

Eisner, J. M. (1996). *An empirical comparison of probability models for dependency grammar*, Technical Report IRCS-96-11, Institute for Research in Cognitive Science, University of Pennsylvania.

Eisner, J. M. (2000). Bilexical grammars and their cubic-time parsing algorithms. In Bunt, H. and Nijholt, A. (eds), *Advances in Probabilistic and Other Parsing Technologies*, Kluwer, pp. 29–62.

Engel, U. (1994). *Syntax der deutschen Gegenwartssprache*.3. Auflage. Berlin: Schmidt.

Engel, U. (1996). Tesniere missverstanden. In *Lucien Tesniere - Syntaxe Structurale et Operation Mentales*. Akten des deutsch-franzosischen Kolloquiums anlasslich der 100 Wiederkehr seines Geburstages, Strasbourg 1993, volume 348 of Linguistische Abeiteu, pp. 53-61. Niedermeyer, Tubingen.

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

Gaifman, H. (1965). Dependency systems and phrase-structure systems, *Information and Control* 8(3), pp. 304-337.

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

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.

Hajičová, Eva, Hana Skoumalová and Petr Sgall. (1995). An Automatic Procedure for Topic-Focus Identification. *In Computational Linguistics* 21(1), pp. 81-94.

Harper, M. P. and Helzerman, R. A. (1995). Extensions to constraint dependency parsing for spoken language processing. *Computer Speech and Language* 9, pp.187–234.

Harper, M. P., Helzermann, R. A., Zoltowski, C. B., Yeo, B. L., Chan, Y., Steward, T. and Pellom, B. L. (1995). Implementation issues in the development of the PARSEC parser. *Software: Practice and Experience* 25, pp. 831–862.

Hays, D. G. (1964). Dependency theory: A formalism and some observations, *Language 40*, pp. 511-525.

Helbig, G. (1992). *Probleme der Valenz- und Kasustheorie. Konzepte der Sprach- und Literaturwissenschaft.* Tübingen: Niemeyer.

Hellwig, P. (1986). Dependency unification grammar. *In Proceedings of the 11th International Conference on Computational Linguistics (COLING)*, pp. 195–198.

Hellwig, P. (2003). Dependency unification grammar. In Agel, V., Eichinger, L.M., Eroms, H.-W., Hellwig, P., Heringer, H. J. and Lobin, H. (eds), *Dependency and Valency*, Walter de Gruyter, pp. 593–635.

Hristea, F. and Popescu, M. (2003). A Dependency Grammar Approach to Syntactic Analysis with Special Reference to Romanian. F. Hristea și M. Popescu (coord.), *Building Awareness in Language Technology*, București, Editura Universității din București, pp. 9-16.

Hudson, R.(1990). *English Word Grammar*.Oxford: Basil Blackwell.

Ion, R. (2007). *Word Sense Disambiguation Methods Applied to English and Romanian*. PhD thesis (in Romanian). Romanian Academy. Bucharest. 138 p.

Ion, R., Irimia, E., Ștefănescu D., Tufiș, D.(2012). ROMBAC: The Romanian Balanced Annotated Corpus, *In Procedings of LREC 2012*, Istanbul, Turkey.

Irimia, E. (2009). EBMT experiments for the English-Romanian Language Pair. *In Recent Advances in Intelligent Information Systems* (Klopotek et al.). Springer, Warsaw, pp. 91-102.

Jarvinen, T. and Tapanainen, P. (1998). Towards an implementable dependency grammar. In Kahane, S. and Polguere, A. (eds), *Proceedings of the Workshop on Processing of Dependency-Based Grammars*, pp. 1–10.

Karlsson, F. (1990). Constraint Grammar as a Framework for Parsing Unrestricted Text. H. Karlgren, ed., *Proceedings of the 13th International Conference of Computational Linguistics*, Vol. 3. Helsinki, pp. 168-173.

Karlsson, F., Voutilainen, A., Heikkilä, J: and Anttila, A (eds.) 1995. Constraint Grammar: A Language-Independent System for Parsing Running Text. *Natural Language Processing*, No 4. Mouton de Gruyter, Berlin and New York. ISBN 3-11-014179-5.

Kasami, T. (1965). *An efficient recognition and syntax algorithm for context-free languages*, Technical Report AF-CRL-65-758, Air Force Cambridge Research Laboratory.

Klein, D., Manning, C.D. (2003). Fast Exact Inference with a Factored Model for Natural Language Parsing. *In Advances in Neural Information Processing Systems* 15 (NIPS 2002), Cambridge, MA: MIT Press, pp. 3-10.

Korhonen, J., (1977). Studien zur Dependenz, Valenz und Satzmodell, Teil 1. *Theorie und Beschreibung der deutschen Gegenwartssprache*. Dokumentation, kritische Besprechung, Vorschläge.Bern: Peter Lang.

Kromann, M. T. (2004). Optimality parsing and local cost functions in Discontinuous Grammar. *Electronic Notes of Theoretical Computer Science* 53, pp. 163–179.

Krujiff, G.-J. M. (2002). *Formal and computational aspects of dependency grammar: History and development of DG*, Technical report, ESSLLI-2002.

Kudo, T. and Matsumoto, Y. (2000). Japanese dependency structure analysis based on support vector machines. *In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora* (EMNLP/VLC), pp. 18–25.

Lombardo, V. and Lesmo, L. (1996). An Earley-type recognizer for Dependency Grammar. *In Proceedings of the 16th International Conference on Computational Linguistics (COLING)*, pp. 723–728.

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

Marimon, M. (2013). The Spanish DELPHIN Grammar. *Language Resources and Evaluation*, 47(2), pp. 371–397

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

Maruyama, H. (1990). Structural disambiguation with constraint propagation. *In Proceedings of the 28th Meeting of the Association for Computational Linguistics* (ACL), Pittsburgh, PA, pp. 31–38.

Mărănduc C. and Perez. A.-C. (2015). A Romanian dependency treebank, CICLing 2015, Cairo, 14-20 April.

Mel'čuk, I. (1988). *Dependency Syntax: Theory and Practice*. State University of New York Press.

Nikula, H. (1986). *Dependensgrammatik*. Liber.

Nivre, J. (2003). An efficient algorithm for projective dependency parsing. In Van Noord, G. (ed.), *Proceedings of the 8th International Workshop on Parsing Technologies (IWPT)*, pp. 149–160.

Nivre, J., Hall, J. and Nilsson, J. (2004). Memory-based dependency parsing. In Ng, H. T. and Riloff, E. (eds), *Proceedings of the 8th Conference on Computational Natural Language Learning (CoNLL)*, pp. 49–56.

Nivre, J. and Hall. J. (2005). MaltParser: A language-independent system for data-driven dependency parsing. *In Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT)*, 9–10 December 2005, Barcelona, Spain.

Nivre, J. (2005). *Dependency grammar and dependency parsing*, Techreport, Växjö University.

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

Nivre, J. (2006). *Inductive Dependency Parsing*. Springer, ISBN-13: 978-1402048883, ISBN-10: 1402048882

Obrebski, T. (2003). Dependency parsing using dependency graph. In Van Noord, G. (ed.), *Proceedings of the 8th International Workshop on Parsing Technologies (IWPT)*, pp. 217–218.

Och, F.-J., Tillmann, Ch., Ney, H. (1990). Improved Alignment Models for Statistical Machine Translation. *Proceedings of the Joint Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora*, College Park, MD, June, pp. 20–28.

Padró, L., Collado M., Reese S., Lloberes M., Castellón, I. (2010). FreeLing 2.1: Five Years of Open-Source Language Processing Tools, *In Proceedings of 7th Language Resources and Evaluation Conference* (LREC 2010), ELRA, La Valletta, Malta. May, 2010.

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

Popescu, M. (2003). Dependency Grammar Annotator. F. Hristea și M. Popescu (coord.), *Building Awareness in Language Technology*, București, Editura Universității din București, pp. 17-34.

Punyakanok, V., Roth, D., Yih W-T. (2008). The Importance of Syntactic Parsing and Inference in Semantic Role Labeling, *Computational Linguistics*, 34(2), pp. 257-287.

Robinson, J. J. (1970). Dependency structures and transformation rules, *Language* 46, 259-285.

Sampson, G. (2003). Thoughts on Two Decades of Drawing Trees, In Abeillé, A. (ed.) *Treebanks. Building and Using Parsed Corpora*, pp. 23-41, Text, Speech and Language Technologies, Volume 20, Springer Netherlands, ISBN 1-4020-1334-5.

Samuelsson, C. (2000). A statistical theory of dependency syntax. *In Proceedings of the 18th International Conference on Computational Linguistics* (COLING).

Sgall, P., Hajičová E., Panevová J. (1986). *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Dordrecht: Reidel.

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

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

Sleator, D. and Temperley, D. (1991). *Parsing English with a link grammar*, Technical Report CMU-CS-91-196, Carnegie Mellon University, Computer Science.

Sleator, D. and Temperley, D. (1993). Parsing English with a link grammar. *Third International Workshop on Parsing Technologies* (IWPT), pp. 277–292.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C.,Erjavec, T., Tufiş, D., and Varga, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. *In Proceedings of the 5th International Conference on Language Resources and Evaluation* (LREC'2006). Genoa. Italy.

Tapanainen, P. and Järvinen, T. (1997). A non-projective dependency parser. *In Proceedings of the 5th Conference on Applied Natural Language Processing.* Washington, DC: Association for Computational Linguistics.

Tarvainen, K. (1981). Einführung in die Dependenzgrammatik. *Reihe Germanistische Linguistik* 35. Tübingen: Niemeyer.

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

Tesnière L. (1959). *Éléments de syntaxe structurale*. Editions Klincksieck

Tiedemann, J. (2009). News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In N. Nicolov and K. Bontcheva and G. Angelova and R. Mitkov (eds.) *Recent Advances in Natural Language Processing* (vol V). pp. 237--248.

Trandabăț, D., Irimia, E., Barbu Mititelu, V., Cristea, D., Tufiș, D. (2012). *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.

Tufiș, D., and Cristea, D. (2002). Methodological issues in building the Romanian Wordnet and consistency checks in BalkaNet. *In Proceedings of LREC 2002 Workshop on Wordnet Structures and Standardisation* (Christodoulakis, D.-N., Kunze, C. and Lemnitzer, L.). Las Palmas, Spain, pp. 35-41.

Tufiș D., and Irimia E. (2006). RoCo_News - A Hand Validated Journalistic Corpus of Romanian. *In Proceedings of the 5th LREC Conference*, Genoa, Italy, 22-28 May, pp. 869-872, ISBN 2-9517408-2-4.

Tufiş, D., Ion R., Ceauşu A., Ştefănescu D. (2008). RACAI's Linguistic Web Services. *In Proceedings of the 6th Language Resources and Evaluation Conference*, LREC'08. Marrakech, Morocco. ELRA -European Language Resources Association.

Tufiş, D., Ion R., Dumitrescu, Ş.D. (2013a). Wikipedia as an SMT Training Corpus. *In Proceedings of the International Conference on Recent Advances on Language Technology* (RANLP 2013). Hissar, Bulgaria.

Tufiş, D., Boroş, T., Dumitrescu, Ş.D. (2013b). The RACAI Speech Translation System. *In Proceedings of the 7th International Conference on Speech Technology and Human-Computer Dialogue* (SPED 2013). Cluj-Napoca.

Wang, W. and Harper, M. P. (2004). A statistical constraint dependency grammar (CDG) parser. In Keller, F., Clark, S., Crocker, M. and Steedman, M. (eds), *Proceedings of the Workshop in Incremental Parsing: Bringing Engineering and Cognition Together* (ACL), pp. 42

Weber, Heinz J. (1996). Translation und Rekursivität bei Lucien Tesnière. In *Lucien Tesniere - Syntaxe Structurale et Operation Mentales*. Akten des deutsch-franzosischen Kolloquiums anlasslich der 100 Wiederkehr seines Geburstages, Strasbourg 1993, volume 348 of Linguistische Abeiteu, pp. 53-61. Niedermeyer, Tubingen.

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

Yamada, H. and Matsumoto, Y. (2003). Statistical dependency analysis with support vector machines. In Van Noord, G. (ed.), *Proceedings of the 8th International Workshop on Parsing Technologies (IWPT)*, pp. 195–206.

Younger, D. H. (1967). Recognition and parsing of context-free languages in time n^3. *Information and Control* 10, pp. 189–208.

ANEXA 1. GHIDUL DE ANDOTARE CU RELAŢII SINTACTICE DE DEPENDENŢE

În exemplele din coloana 3, centrul este marcat cu caractele cursive (eng. italics) iar dependentul este marcat cu caractere îngroşate (eng. bold). Intrările din tabel sunt ordonate alfabetic după eticheta dependenţei (coloana 4).

centru	**dependent**	**exemplu**	**eticheta dependenței**
substantiv	verb ne-predicativ	Mă impresionează o *fată* **plângând.**	*acl*
substantiv	verb în propoziția subordonată	*fata* care **locuiește** la Paris /*locul* unde nu s-a **întâmplat** nimic / *dorința* să **învingă**	*acl*
adjectiv	verb în propoziția subordonată	E *nervos*, fiindcă nu a **terminat** lucrarea./E atât de *nervos*, încât va **rămâne** acasă.	*advcl*
adjectiv	verb în propoziția subordonată	Ion e mai *mic* decât **e** Ana./ E mai *tânăr* de_cum îl **știau** ei./*Curată* cum **e** zăpada./ Era *trist* de_parcă/ca_și_cum i se **înecaseră** corăbiile.	*advcl*
adverb	conjuncție subordonatoare	Nu e *bine*, **pentru_că** redactarea este neclară.	*advcl*
adverb	conjuncție subordonatoare sau verb în subordonată introdusă de un relativ	Desenează mai *repede* **decât** desenezi tu./ Scrie la fel de *bine* cum **scria** acum un an.	*advcl*
interjecție	conjuncție subordonatoare sau verb în subordonată introdusă de un relativ	Și *pleosc!* o palmă, **fiindcă** se enervase.	*advcl*
verb	verb nepredicativ	**Ajungând** la serviciu/**Ajunsă** la serviciu, *a observat* că-i lipsea cheia de la birou.	*advcl*
verb	conjuncție subordonatoare	**Deși** nimeni nu se aștepta, *a venit*.	*advcl*
oricare	conjuncție cu valoare adverbială	Efectuez **și** *lucrări* de înlocuire.	*advmod*
adjectiv	adverb	Mi-l amintesc **mereu** *alert*.	*advmod*
adjectiv	adverb	O călătorie mai *frumoasă* **ca**/**decât** în ianuarie/la fel de *frumoasă* **ca** în ianuarie.	*advmod*
interjecție	adverb	*Iată*-l **acolo** pe Ion./*Hai* **mâine** la munte.	*advmod*
substantiv	adverb	*Cititul* **noaptea** nu-i sănătos.	*advmod*
verb	adverb	*Am văzut* trenul **acolo**./**Totuși** *a venit*./Ion *vine* **iarna**./ **Totuși** mi-*a adus* cartea, *deși nu* mai *speram* să o primesc înapoi./**Parcă** *au venit*.	*advmod*
adjectiv	prepoziție	Teoria e *acceptabilă* **de_către** toți cercetătorii.	*agc*
verb ne-predicativ	prepoziție	Aceasta este calea *de urmat* **de_către** orice om integru.	*agc*
verb	prepoziție	Teoria *e acceptată* **de_către** toți cercetătorii.	*agc*
substantiv	adjectiv	*Casă* **frumoasă**/ **doi** *copii* / **aceste** *case* / *mere* **coapte** / *ordine* **crescândă** / **halal**	*amod*

		treabă / *bărbați* **bine**	
pronume, numeral	adjectiv	*Acela* **roșu** / *ceva* **temeinic** / *nimic* **remarcabil** / vreau 3 garoafe: *una* **albă** și *două* **roșii**	*amod*
adjectiv	adjectiv	*Mână-spartă* (adică **risipitor**), așa îl știau toți.	*appos*
adverb	conjuncție subordonatoare	Muncește *continuu*, adică **fără** să facă pauze.	*appos*
adverb	verb în subordonată introdusă de un relativ	Ne-am întâlnit *aici*, unde **am stabilit.**	*appos*
verb ne-predicativ	verb ne-predicativ	A reușind *muncind*, adică **asudând**...	*appos*
verb ne-predicativ	verb în subordonată introdusă de un relativ	Muncește *delocalizat*, adică unde i se **cere**...	*appos*
verb ne-predicativ	verb ne-predicativ	Vocea *vibrândă*, adică **tremurândă** de emoție	*appos*
substantiv	substantiv	Numai *soldatul*, **fratele** meu	*appos*
substantiv	numeral	*Elena*, a **doua**, avea cele mai mari șanse.	*appos*
substantiv	pronume	*Venerabilul* (adică **eu**) merge diseară la întrunire.	*appos*
substantiv	conjuncție subordonatoare	Asta e *calea*: **să** fiu deștept.	*appos*
substantiv	verb în subordonată relativă	*Mama*, care acum a **auzit**	*appos*
prepoziție	prepoziție	Este vorba *de* ceva foarte simplu, anume **de** a insufla...	*appos*
pronume	substantiv	Numai *tu*, **fratele** meu	*appos*
pronume	substantiv	Vine *el* **tata** imediat.	*appos*
pronume	numeral	*Tu*, a **doua**, ai cele mai mari șanse.	*appos*
pronume	pronume	*Tu* (adică **eu**) vei merge diseară la întrunire!	*appos*
pronume	verb în subordonată relativă	*Eu*, care acum am **înțeles**.	*appos*
substantiv	substantiv	Ana Ionescu, *Str.* **Rozelor,** nr 3	*appos*
verb ne-predicativ	auxiliar	**Am** *stabilit*	*aux*
verb ne-predicativ	auxiliar „a fi” în diateza pasivă	A sfârșit prin a **fi** *spânzurat*.	*auxpass*
primul conjunct	conjuncție coordonatoare	*Maria* **și** Ion	*cc*
numeral	numeral	Am **patru** *mii* de lei.	*compound*
primul element al unei coordonări	al n-lea element al unei coordonări, unde n ≠1	*Maria* și **Ion**	*conj*
verb	pronume neaccentuat	Am *văzut*-**o** pe Maria. **I**-am *spus* lui Ion.	*dblclitic*
substantiv	articol hotărât	Spune-i **lui** *Ion*.	*det*
substantiv	articol nehotărât	Am văzut **un** *okapi*.	*det*
substantiv	articol demonstrativ	**cel** de-al doilea *copil*	*det*
substantiv	articol posesiv/genitiv	o rochie de **a** *Mariei*	*det*
numeral	pronume semi-independent	Cartea **celui** de-al *doilea* e nouă.	*det*
verb	element dislocat	Am *băut*-o, **cafeaua.**	*dislocated*
interjecție	substantiv	*Iată* o **nuntă**./*Na* **țigări**!	*dobj*

interjecție	prepoziție	*Iat*-o **pe** Maria**.**	*dobj*
interjecție	pronume	*Iată*-**l**.	*dobj*
verb	conjuncție subordonatoare	*Înțeleg* **că** ești obosit.	*dobj*
verb	substantiv	*Văd* **fata**.	*dobj*
verb	numeral	Am *citit* **două.**	*dobj*
verb	prepoziție	O *văd* **pe** Mara.	*dobj*
verb	pronume	**O** *văd*.	*dobj*
		I-a zis *good* **bye** și a plecat.	*foreign*
		L-am *bă*...**nuit** că nu e sincer.	*goeswith*
adjectiv	substantiv	Concurs *deschis* **elevilor** din clasele a patra.	*iobj*
adjectiv	pronume	Conjunctura nu-**mi** este *favorabilă*.	*iobj*
adverb	substantiv	Fiul se comportă *aidoma* **tatălui**.	*iobj*
adverb	pronume	Fiul se comportă *aidoma* **lui**.	*iobj*
interjecție	substantiv	*Bravo* **mamei!**	*iobj*
interjecție	pronume	*Na*-**ți** cartea pe care mi-ai cerut-o./*Bravo* **lor**!/*Vai* **mie!**	*iobj*
verb	verb în subordonată relativă	*Dau* cui **are** nevoie.	*iobj*
verb	substantiv	I-am *spus* **Mariei.**	*iobj*
verb	numeral	**Primei** i-a *mers* bine./Am *dat* prăjituri **amândurora.**	*iobj*
verb	prepoziție	Am *dat* prăjituri **la** trei dintre ei./ *Zis*-a el **către** mine.	*iobj*
verb	pronume	**I**-am *spus*.	*iobj*
		Ana Ionescu, **Str.** Rozelor, **nr** 3	*list*
centrul unei apoziții	adverb	Venerabilul **(adică** *eu*) merge diseară la întrunire	*mark*
substantiv/ pronume	Conjuncție	Vine **și** *el* imediat.	*mark*
verb	"să" în regentă	**Să** *poftiți*!	*mark*
verb (infinitiv)	"a"	Arta de **a** *vorbi* manierat.	*mark*
adjectiv	adverb	Copilul este **mai** *afectuos* acum. / Copilul este **la_fel_de** / **tot_așa_de**/ **tot_atât_de** *afectuos.* / Copilul este **mai_puțin** *afectuos.* / Copilul este **foarte** *afectuos.* / Copilul este **cel_mai** *afectuos.* / Copilul este **cel_mai_puțin** *afectuos.*	*mark*
adverb	adverb	Copilul vorbește **mai** *afectuos* acum. / Copilul vorbește **la_fel_de** / **tot_așa_de**/ **tot_atât_de** *afectuos.* / Copilul vorbește **mai_puțin** *afectuos.* / Copilul vorbește **foarte** *afectuos.* / Copilul vorbește **cel_mai** *afectuos.* / Copilul vorbește **cel_mai_puțin** *afectuos.*	*mark*
verb (supin)	prepoziție	Apa este **de** *băut.*	*mark*
substantiv propriu	substantiv propriu	*Elena* **Irimia** e la Barcelona./ *Tîrgu* **Mureș** e un oraș frumos.	*name*
verb	adverb	**Nu** *știe* nimic.	*neg*
substantiv	substantiv	*cămașa* **fetei** / *orașul* **București**/ *acordarea* de ajutoare **sinistraților** / *numirea* acestui ins **ministru**	*nmod*
substantiv	pronume	*strigătul* **lui**/ *mâna*-**ți**/ *cartea* **fiecăruia**	*nmod*

verb	substantiv	*A stat* la Paris toată **vara/vara** următoare./ *Stai* **locului**!/ Pepenele c*ântărește* două **kilograme**.	*nmod*
verb	pronume	*Costă* **ceva**./Nu *costă* **nimic.**	*nmod*
verb	verb	Am *ajuns* târziu, a **spus** el.	*parataxis*
verb	pronume reflexive	**Se** *bat* albușurile cu zahăr	*passmark*
adjectiv	prepoziție	Fata era *albă* **la** față/*înaltă* **de** trei metri./ **De** *frumoasă*, e frumoasă.	*pmod*
adjectiv	prepoziție	Este mai *înalt* **decât** tine./ E *timid*, **asemenea** ție./ E tot așa de *nehotărâtă* **ca** prietena ei./Băiatul cel mai *cuminte* **din** lume/**dintre** toți copiii.	*pmod*
adjectiv	prepoziție	Gardul e **cât** casa de *înalt*.	*pmod*
adverb	prepoziție	E *bine* **cu** sănătatea.	*pmod*
adverb	prepoziție	Strigă la fel de *tare* **ca** mine.	*pmod*
interjecție	prepoziție	*Iată*-l **la** ușă pe Ion.	*pmod*
substantiv	prepoziție	*Fața* **de** pernă	*pmod*
verb	prepoziție	**De** poet, *e* poet.	*pmod*
verb	prepoziție	Am *văzut* trenul **în** gară./A mai *venit* și altcineva **decât** tine?	*pmod*
adjective	conjuncție subordonatoare	Omul era *gata* **să** înceapă lucrul.	*pobj*
adjective	prepoziție	Presa este *străină* **de** jocurile de culise./ Fata era *hotărâtă* **a** lupta până la capăt.	*pobj*
adverb	prepoziție	Locuiesc *aproape* **de** gară.	*pobj*
interjecție	conjuncție subordonatoare	*Mersi* **că** ai venit la timp.	*pobj*
interjecție	prepoziție	*Vai* **de** ei!/ *Mulțumesc*/*Mersi* **pentru** masă./ *Halal* **de** ei.	*pobj*
verb	conjuncție subordonatoare	Mă *tem* că nu voi **reuși**.	*pobj*
verb	prepoziție	L-au *angajat* **ca** grădinar./M-au *luat* **drept** tine./Ion s-a *prezentat* **drept** mine la examen./Mă *gândesc* **la** Ion./Ion *are* **în** tine un prieten devotat./Ion și Maria *contează* unul **pe** altul./**Din** îmbujorată/bucătăreasă, fata s-a *făcut* albă/doctoriță./ Ion se *gândește* **la** ea./ Mă *tem* **a** spune adevărul.	*pobj*
interjecție	pronume neaccentuat	*Uite*-**ți** paltonul! *Iată*-**ți** fratele! *Na*-**ți** caietul înapoi.	*poss*
verb	substantiv	Lui **Ion** i-am *auzit* vocea.	*poss*
verb	pronume neaccentuat	**I**-am *auzit* vocea.	*poss*
verb	clitic	Lui Ion **i**-am *auzit* vocea.	*possclitic*
adjective	prepoziție	Gardul e cât casa **de** *înalt.*	*post*
verb	adjectiv	Fata *este* **înaltă.**	*pred*
verb	adverb	*Este* **atât**. Este ***ciudat*** că...	*pred*
verb	interjecție	*Este* **vai** de capul lor!	*pred*
verb	verb ne-predicativ	Apa *este* de **băut.**	*pred*
verb	substantiv	Fata *este* **prietena** mea.	*pred*
verb	substantiv/pronume	Cartea *este* a **Mariei.**	*pred*
verb	numeral	El *este* **primul.**	*pred*
verb	prepoziție	Caroseria *e* **din** metal.	*pred*

verb	pronume	Fratele meu *este* **el**.	*pred*
verb	conjuncție subordonatoare	Credința lor *este* **că** vor ajunge...	*pred*
verb	conjuncție subordonatoare sau verb în subordonată introdusă de un relativ	Întrebarea *este* **dacă** mă ajută. / Întrebarea *este* cine mă **ajută.**	*pred*
prepoziție	adverb	*În_afară_de* **mâine**, altădată nu mai pot veni.	*prep*
prepoziție	verb ne-predicativ	L-au desemnat *ca* **reprezentând** România la NATO. / S-a apucat *de* **gătit**. / A sfârșit *prin* a fi **spânzurat**./Se teme *a* **spune** adevărul.	*prep*
prepoziție	substantiv	M-au luat *drept* **inspector. /** L-am văzut *în* **gară**. / Mă gândesc *la* **Ion**.	*prep*
prepoziție	numeral	Stătea *lângă* **doi** mai voinici. / Muncește *cât* **doi**.	*prep*
prepoziție	pronume	M-au luat *drept* **tine**. / A pus cartea *sub* **el**. / Mă gândesc *la* **tine**. / Nu se moare *din* **asta**.	*prep*
verb	pronume reflexiv	**M**-am *răzgândit*.	*reflclitic*
		Maria a mers la Berlin, **Elena** la Barcelona. / *Îți* telefonează mai des decât **surorii** lui / la fel de des ca **prietenei** lui.	*remnant*
		Mergi **la stânga**... *la dreapta*.	*reparandum*
adverb		**Jos** mafia!	*ROOT*
interjecție		**Marș** în camera ta!	*ROOT*
verb		**Vorbește** tare.	*ROOT*
conjuncție	verb	Vreau *să* **ies** la soare.	*sc*
verb	verb ne-predicativ	L-am *convins* a **spune** adevărul.	*secobj*
verb	conjuncție subordonatoare	Te *anunț* **că** hotărârea mea e nestrămutată.	*secobj*
verb	substantiv	L-a *învățat* **poezia**.	*secobj*
verb	pronume	L-a *învățat* **asta**. / Nu-l *întreabă* **nimic.**	*secobj*
substantiv	pronume	*Fetele* n-au venit **niciuna** la examen.	*spe*
pronume	adverb	*I*-ai văzut **împreună?**	*spe*
verb	adjectiv	Femeia se *arăta* **simpatică**.	*spe*
verb	adverb	**Cum** l-au *denumit*? / L-au *denumit* **așa.**	*spe*
verb	substantiv	L-au *botezat* **Ion**. / L-au *ales* **deputat.**	*spe*
verb	pronume relativ	L-au *ales* **ceva.**	*spe*
verb	interjecție	M-a *lăsat* **paf.**	*spe*
verb	verb ne-predicativ	Pe Maria o *vedeam* **trecând** în fiecare dimineață.	*spe*
verb	substantiv	S-a *întors* **bărbat.**	*spe*
verb	numeral	Ea a *ieșit* **a doua** pe județ.	*spe*
verb	conjuncție subordonatoare	Te *știu* **că** minți.	*spe*
adverb	substantiv	*Jos* **mafia**!	*subj*
interjecție	substantiv	**Vecinul** *hop* la masă.	*subj*

interjecție	pronume	Fato, *marș* și **tu** în camera ta!	*subj*
verb ne-predicativ	substantiv	E greu de *manevrat* **substanțe** periculoase (de către persoane fără echipamentul necesar).	*subj*
verb	verb ne-predicativ	A **greși** *e* omenesc. / Se *aude* **tunând**. / *Trebuie* **făcut** acest lucru. / *Este* important de **citit** cartea.	*subj*
verb	conjuncție subordonatoare	Se *cuvine* **să** salutați. / Se *crede* **că** el a câștigat.	*subj*
verb	verb în subordonată introdusă de un relativ	Cine a **încălcat** legea a fost *pedepsit* de instanță.	*subj*
verb	adverb	Mi-*e* **bine.**	*subj*
verb	substantiv	**Copilul** *citește*.	*subj*
verb	numeral	**Doi** *aleargă*.	*subj*
verb	pronume	**El** *citește*.	*subj*
verb	pronume semi-independent	**A** mea *spală* bine./ **Cel** înalt *vorbește* politicos.	*subj*
verb	substantiv	**Cântecul** e *compus* de mama.	*subj*
verb	numeral	**Două** sunt *compuse* de mama.	*subj*
verb	pronume	**Acesta** e *compus* de mama.	*subj*
verb	pronume semi-independent	**Al** Mariei a fost *croit* de mama.	*subj*
verb	substantiv	**Ioana,** *vino* la mine!	*voc*
verb	adjectiv	**Bucuros** că a câștigat, și-*a invitat* prietenii la o cină. / **Singură**, *nu se duce* în excursie.	*xcomp*

ANEXA 2. CORESPONDENŢA ETICHETELOR MORFO-LEXICALE ÎNTRE ROMBAC (RO) ŞI IULA LSP (SP)

Etichete RO	Etichete SP	Etichete RO	Etichete SP
Afcfp-n	AQCFP0	Pi3-pr	PI30PN00
Afcfson	AQCFS0	Pi3--r	PI300N00
Afcfsrn	AQCFS0	Pi3-sr	PI30SN00
Afcms-n	AQCMS0	PLUS	Fz
Afp	AQ0000	Pp1-pa--------w	PP10PA00
Afpf--n	AQ0F00	Pp1-pa--y-----w	PP10PA00
Afpfp-n	AQ0FP0	Pp1-pd--------w	PP10PD00
Afpfpoy	AQ0FP0	Pp1-pd--y-----w	PP10PD00
Afpfpry	AQ0FP0	Pp1-pr--------s	PP10PN00
Afpfson	AQ0FS0	Pp1-sa--------s	PP10SA00
Afpfsoy	AQ0FS0	Pp1-sa--------w	PP10SA00
Afpfsrn	AQ0FS0	Pp1-sa--y-----w	PP10SA00
Afpfsry	AQ0FS0	Pp1-sd--------w	PP10SD00
Afpmp-n	AQ0MP0	Pp1-sd--y-----w	PP10SD00
Afpmpoy	AQ0MP0	Pp1-sn--------s	PP10SN00
Afpmpry	AQ0MP0	Pp2-pa--------w	PP20PA00
Afpms-n	AQ0MS0	Pp2-pa--y-----w	PP20PA00
Afpmsoy	AQ0MS0	Pp2-pd--------w	PP20PD00
Afpmsry	AQ0MS0	Pp2-pd--y-----w	PP20PD00
Afp-p-n	AQ00P0	Pp2----------s	PP200000
Afp-poy	AQ00P0	Pp2-sa--------s	PP20SA00
Afsfsrn	AQSFS0	Pp2-sa--------w	PP20SA00
AMPER	AMPER0	Pp2-sa--y-----w	PP20SA00
BULLET	Fg	Pp2-sd--------w	PP20SD00
Cccsp	CC	Pp2-sd--y-----w	PP20SD00
Ccssp	CC	Pp2-sn--------s	PP20SN00
COLON	Fd	Pp2-sr--------s	PP20SN00
COMMA	Fc	Pp3fpa--------w	PP3FPA00
Crssp	CC	Pp3fpa--y-----w	PP3FPA00
Cscsp	CS	Pp3fpr--------s	PP3FPN00
Csssp	CS	Pp3fsa--------w	PP3FSA00
Cssspy	CS	Pp3fsa--y-----w	PP3FSA00
DASH	Fg	Pp3fso--------s	PP3FSO00
DBLQ	Fe	Pp3fsr--------s	PP3FSN00
Dd3fpo	DD3FP	Pp3mpa--------w	PP3MPA00
Dd3fpr	DD3FP	Pp3mpa--y-----w	PP3MPA00
Dd3fpr---e	DD3FP0	Pp3mpr--------s	PP3MPN00
Dd3fso	DD3FS	Pp3msa--------w	PP3MSA00
Dd3fso---e	DD3FS0	Pp3msa--y-----w	PP3MSA00

Dd3fsr	DD3FS	Pp3mso--------s	PP3MSO00
Dd3fsr---e	DD3FS0	Pp3msr--------s	PP3MSN00
Dd3fsr---o	DD3FS0	Pp3-pd--------w	PP30PD00
Dd3mpo	DD3MP	Pp3-pd--y-----w	PP30PD00
Dd3mpr	DD3MP	Pp3-p---------s	PP30P000
Dd3mpr---e	DD3MP0	Pp3-sd--------w	PP30SD00
Dd3mso---e	DD3MS0	Pp3-sd--y-----w	PP30SD00
Dd3msr---e	DD3MS0	Ps1fsrp	PX1FSNP0
Dd3msr---o	DD3MS0	Ps1fsrs	PX1FSNS0
Dd3-po---e	DD30P0	Ps1mp-p	PX1MP0P0
Dd3-po---o	DD30P0	Ps1mp-s	PX1MP0S0
Dh1ms	D01MS0	Ps3fp-s	PX3FP0S0
Dh3fsr	D03FS	Ps3fsrs	PX3FSNS0
Dh3ms	D03MS0	Ps3ms-s	PX3MS0S0
Di3	DI3000	Pw3fpr	PR3FPN00
Di3fp	DI3FP0	Pw3fso	PR3FSO00
Di3fpr	DI3FP	Pw3mpr	PR3MPN00
Di3fpr---e	DI3FP0	Pw3mso	PR3MSO00
Di3fso---e	DI3FS0	Pw3msr	PR3MSN00
Di3fsr	DI3FS	Pw3-po	PR30PO00
Di3fsr---e	DI3FS0	Pw3--r	PR300N00
Di3mp	DI3MP0	Px3--a--------s	PX300A00
Di3mpr	DI3MP	Px3--a--------w	PX300A00
Di3mpr---e	DI3MP0	Px3--a--y-----w	PX300A00
Di3ms----e	DI3MS0	Px3--d--------w	PX300D00
Di3mso---e	DI3MS0	Px3--d--y-----w	PX300D00
Di3msr	DI3MS	Pz3fsr	PZ3FSN00
Di3msr---e	DI3MS0	Pz3mso	PZ3MSO00
Di3-po	DI30P	Pz3msr	PZ3MSN00
Di3-po---e	DI30P0	Pz3-sr	PZ30SN00
Di3--r---e	DI3000	QUEST	Fit
Di3-sr---e	DI30S0	Rc	RG
Ds1fp-p	DP1FPP	Rgc	RG
Ds1fp-s	DP1FPS	Rgp	RG
Ds1fsop	DP1FSP	Rgpy	RG
Ds1fsos	DP1FSS	Rgs	RG
Ds1fsrp	DP1FSP	Rp	RG
Ds1fsrs	DP1FSS	RPAR	Fpt
Ds1mp-p	DP1MPP	RSQR	Fct
Ds1mp-s	DP1MPS	Rw	RG
Ds1ms-p	DP1MSP	Rz	RZ
Ds1ms-s	DP1MSS	SCOLON	Fx
Ds2fsrs	DP2FSS	SLASH	Fh
Ds2ms-s	DP2MSS	Spca	SPS00

Ds3fp-s	DP3FPS	Spcg	SPS00
Ds3fsos	DP3FSS	Spsa	SPS00
Ds3fsrs	DP3FSS	Spsay	SPS00
Ds3mp-s	DP3MPS	Spsd	SPS00
Ds3ms-s	DP3MSS	Spsg	SPS00
Ds3---p	DP300P	STAR	STA00
Ds3---s	DP300S	Tdfpr	DA0FP0
Dw3fso---e	DT3FS0	Tdfsr	DA0FS0
Dw3fsr	DT3FS	Tdmpr	DA0MP0
Dw3mso---e	DT3MS0	Tdmso	DA0MS0
Dw3-po---e	DT30P0	Tdmsr	DA0MS0
Dw3--r---e	DT3000	Td-po	DA00P0
Dz3fso---e	D03FS0	Tffs-y	DA0FS0
Dz3fsr---e	D03FS0	Tfmsoy	DA0MS0
Dz3msr---e	D03MS0	Tfmsry	DA0MS0
EQUAL	Fz	Tfms-y	DA0MS0
EXCL	Fat	Tf-so	DA00S0
HELLIP	Fs	Tifso	DA0FS0
I	I	Tifsr	DA0FS0
LPAR	Fpa	Timso	DA0MS0
LSQR	Fca	Timsr	DA0MS0
Mc	Z	Ti-po	DA00P0
Mcfp-l	Z	Tsfp	DA0FP0
Mcfp-ln	Z	Tsfs	DA0FS0
Mcfprln	Z	Tsmp	DA0MP0
Mcfsrln	Z	Tsms	DA0MS0
Mcmp-l	Z	Va--1	VA00100
Mcmsrl	Z	Va--1p	VA001P0
Mc-p-d	Z	Va--1s	VA001S0
Mc-p-l	Z	Va--2p	VA002P0
Mlfpr	Z	Va--2s	VA002S0
Mlmpr	Z	Va--3	VA00300
Mmfsr-n	AQFSR0	Va--3p	VA003P0
Mofpoly	AO0FP0	Va--3p----y	VA003P0
Mofprly	AO0FP0	Va--3s	VA003S0
Mofs-l	AO0FS0	Va--3s----y	VA003S0
Mofsrly	AO0FS0	Vag	VAG0000
Mo---l	AO0000	Vaii1	VAII100
Mompoly	AO0AP0	Vaii3p	VAII3P0
Moms-l	AO0AS0	Vaii3s	VAII3S0
Moms-ln	AO0AS0	Vail3p	VAI03P0
Momsoly	AO0AS0	Vail3s	VAI03S0
Momsrly	AO0AS0	Vaip1p	VAIP1P0
Mo-s-r	AO00S0	Vaip2p	VAIP2P0

Nc	NC00000	Vaip2s	VAIP2S0
Ncf--n	NCF0000	Vaip3p	VAIP3P0
Ncfp-n	NCFP000	Vaip3s	VAIP3S0
Ncfpoy	NCFP000	Vais3s	VAIS3S0
Ncfpry	NCFP000	Vanp	VANP000
Ncfson	NCFS000	Vap--sm	VAP00SM
Ncfsoy	NCFS000	Vasp1p	VASP1P0
Ncfsrn	NCFS000	Vasp1s	VASP1S0
Ncfsry	NCFS000	Vasp2s	VASP2S0
Ncfsvy	NCFS000	Vasp3	VASP300
Ncm--n	NCM0000	Vmg	VMG0000
Ncmp-n	NCMP000	Vmg-------y	VMG0000
Ncmpoy	NCMP000	Vmii1	VMII100
Ncmpry	NCMP000	Vmii2p	VMII2P0
Ncmpvy	NCMP000	Vmii2s	VMII2S0
Ncms-n	NCMS000	Vmii3p	VMII3P0
Ncmsoy	NCMS000	Vmii3s	VMII3S0
Ncmsrn	NCMS000	Vmil1	VMI0100
Ncmsry	NCMS000	Vmil1p	VMI01P0
Ncmsvn	NCMS000	Vmil3p	VMI03P0
Ncmsvy	NCMS000	Vmil3s	VMI03S0
Nc---n	NC00000	Vmip1p	VMIP1P0
Np	NP00000	Vmip1s	VMIP1S0
Npfpoy	NPFP000	Vmip2p	VMIP2P0
Npfson	NPFS000	Vmip2s	VMIP2S0
Npfsoy	NPFS000	Vmip3	VMIP300
Npfsry	NPFS000	Vmip3p	VMIP3P0
Npmpoy	NPMP000	Vmip3s	VMIP3S0
Npmsoy	NPMS000	Vmip3s----y	VMIP3S0
Npmsry	NPMS000	Vmis1p	VMIS1P0
Pd3fpr	PD3FPN00	Vmis1s	VMIS1S0
Pd3fso	PD3FSO00	Vmis3p	VMIS3P0
Pd3fsr	PD3FSN00	Vmis3s	VMIS3S0
Pd3fsr--y	PD3FSN00	Vmm-2p	VMM02P0
Pd3mpr	PD3MPN00	Vmm-2s	VMM02S0
Pd3mso	PD3MSO00	Vmm-2s----y	VMM02S0
Pd3msr	PD3MSN00	Vmnp	VMNP000
Pd3-po	PD30PO00	Vmp--pf	VMP00PF
PERIOD	Fp	Vmp--pm	VMP00PM
Pi3fpr	PI3FPN00	Vmp--sf	VMSP294
Pi3fso	PI3FSO00	Vmp--sm	VMSP295
Pi3fsr	PI3FSN00	Vmp--sm---y	VMSP296
Pi3mpr	PI3MPN00	Vmsp1p	VMSP297
Pi3mso	PI3MSO00	Vmsp2	VMSP298

Pi3msr	PI3MSN00	Vmsp3	VMSP299
Pi3-po	PI30PO00	Vmsp3-----y	VMSP300

ANEXA 3: FORMATUL CONLL ȘI FORMATUL GRAPHML PENTRU PROPOZIȚIA: “ARE 52 DE ANI, ESTE CĂSĂTORIT ȘI ARE O FIICĂ.”

Formatul CONLL

1	Are	ara	v	Vmsp3	_	0	ROOT	_	_	
2	52	52	e	Eni	_	4	amod	_	_	
3	de	de	s	Spsa	_	4	post	_	_	
4	ani	an	n	Ncmp-n	_	1	dobj	_	_	
5	,	,	c	COMMA		_	1	punct	_	_
6	este	fi	v	Vaip3s	_	1	conj	_	_	
7	căsătorit	căsători	v	Vmp--sm		_	6	pred	_	_
8	și	și	c	Crssp	_	1	cc	_	_	
9	are	avea	v	Vmip3s	_	1	conj	_	_	
10	o	un	t	Tifsr	_	11	det	_	_	
11	fiică	fiică	n	Ncfsrn	_	9	dobj	_	_	
12	.	.	p	PERIOD		_	1	punct	_	_

Formatul GRAPHML

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<graphml xmlns="http://graphml.graphdrawing.org/xmlns" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:y="http://www.yworks.com/xml/graphml" xmlns:yed="http://www.yworks.com/xml/yed/3" xsi:schemaLocation="http://graphml.graphdrawing.org/xmlns http://www.yworks.com/xml/schema/graphml/1.1/ygraphml.xsd">
  <!--Created by yEd 3.14-->
  <key for="port" id="d0" yfiles.type="portgraphics"/>
  <key for="port" id="d1" yfiles.type="portgeometry"/>
  <key for="port" id="d2" yfiles.type="portuserdata"/>
  <key attr.name="conllID" attr.type="string" for="node" id="d3"/>
  <key attr.name="label" attr.type="string" for="node" id="d4"/>
  <key attr.name="POS" attr.type="string" for="node" id="d5"/>
  <key attr.name="LEMA" attr.type="string" for="node" id="d6"/>
  <key attr.name="func" attr.type="string" for="node" id="d7"/>
  <key attr.name="url" attr.type="string" for="node" id="d8"/>
  <key attr.name="description" attr.type="string" for="node" id="d9"/>
  <key for="node" id="d10" yfiles.type="nodegraphics"/>
  <key for="graphml" id="d11" yfiles.type="resources"/>
  <key attr.name="label" attr.type="string" for="edge" id="d12"/>
  <key attr.name="sourceID" attr.type="string" for="edge" id="d13"/>
  <key attr.name="targetID" attr.type="string" for="edge" id="d14"/>
  <key attr.name="url" attr.type="string" for="edge" id="d15"/>
  <key attr.name="description" attr.type="string" for="edge" id="d16"/>
  <key for="edge" id="d17" yfiles.type="edgegraphics"/>
  <graph edgedefault="directed" id="G">
    <node id="n0">
      <data key="d3"><![CDATA[1]]></data>
      <data key="d4"><![CDATA[Are]]></data>
      <data key="d5"><![CDATA[Vmsp3]]></data>
      <data key="d6"><![CDATA[ara]]></data>
      <data key="d7"/>
      <data key="d9"/>
      <data key="d10">
```

```
      <y:ShapeNode>
        <y:Geometry height="49.40234375" width="51.34765625" x="225.6946831597222" y="54.701171875"/>
        <y:Fill color="#CCCCFF" transparent="false"/>
        <y:BorderStyle color="#000000" type="line" width="1.0"/>
        <y:NodeLabel alignment="center" autoSizePolicy="content" fontFamily="Dialog" fontSize="12" fontStyle="plain" hasBackgroundColor="false" hasLineColor="false" height="33.40234375" modelName="internal" modelPosition="c" textColor="#000000" visible="true" width="41.34765625" x="5.0" y="8.0">1) Are
Vmsp3</y:NodeLabel>
        <y:Shape type="rectangle"/>
      </y:ShapeNode>
    </data>
  </node>
  <node id="n1">
    <data key="d3"><![CDATA[2]]></data>
    <data key="d4"><![CDATA[52]]></data>
    <data key="d5"><![CDATA[Eni]]></data>
    <data key="d6"><![CDATA[52]]></data>
    <data key="d7"/>
    <data key="d9"/>
    <data key="d10">
      <y:ShapeNode>
        <y:Geometry height="49.40234375" width="41.3515625" x="86.01792844742064" y="243.505859375"/>
        <y:Fill color="#CCCCFF" transparent="false"/>
        <y:BorderStyle color="#000000" type="line" width="1.0"/>
        <y:NodeLabel alignment="center" autoSizePolicy="content" fontFamily="Dialog" fontSize="12" fontStyle="plain" hasBackgroundColor="false" hasLineColor="false" height="33.40234375" modelName="internal" modelPosition="c" textColor="#000000" visible="true" width="31.3515625" x="5.0" y="8.0">2) 52
Eni</y:NodeLabel>
        <y:Shape type="rectangle"/>
      </y:ShapeNode>
    </data>
  </node>
  <node id="n2">
    <data key="d3"><![CDATA[3]]></data>
    <data key="d4"><![CDATA[de]]></data>
    <data key="d5"><![CDATA[Spsa]]></data>
    <data key="d6"><![CDATA[de]]></data>
    <data key="d7"/>
    <data key="d9"/>
    <data key="d10">
      <y:ShapeNode>
        <y:Geometry height="49.40234375" width="41.3515625" x="157.36951574900792" y="243.505859375"/>
        <y:Fill color="#CCCCFF" transparent="false"/>
        <y:BorderStyle color="#000000" type="line" width="1.0"/>
        <y:NodeLabel alignment="center" autoSizePolicy="content" fontFamily="Dialog" fontSize="12" fontStyle="plain" hasBackgroundColor="false" hasLineColor="false" height="33.40234375" modelName="internal" modelPosition="c" textColor="#000000" visible="true" width="31.3515625" x="5.0" y="8.0">3) de
Spsa</y:NodeLabel>
        <y:Shape type="rectangle"/>
      </y:ShapeNode>
    </data>
  </node>
  <node id="n3">
    <data key="d3"><![CDATA[4]]></data>
    <data key="d4"><![CDATA[ani]]></data>
    <data key="d5"><![CDATA[Ncmp-n]]></data>
    <data key="d6"><![CDATA[an]]></data>
    <data key="d7"/>
    <data key="d9"/>
```

```
    <data key="d10">
      <y:ShapeNode>
        <y:Geometry height="49.40234375" width="56.005859375" x="92.69236731150792" y="164.103515625"/>
        <y:Fill color="#CCCCFF" transparent="false"/>
        <y:BorderStyle color="#000000" type="line" width="1.0"/>
        <y:NodeLabel alignment="center" autoSizePolicy="content" fontFamily="Dialog" fontSize="12" fontStyle="plain" hasBackgroundColor="false" hasLineColor="false" height="33.40234375" modelName="internal" modelPosition="c" textColor="#000000" visible="true" width="46.005859375" x="5.0" y="8.0">4) ani
Ncmp-n</y:NodeLabel>
        <y:Shape type="rectangle"/>
      </y:ShapeNode>
    </data>
  </node>
  <node id="n4">
    <data key="d3"><![CDATA[5]]></data>
    <data key="d4"><![CDATA[,]]></data>
    <data key="d5"><![CDATA[COMMA]]></data>
    <data key="d6"><![CDATA[,]]></data>
    <data key="d7"/>
    <data key="d9"/>
    <data key="d10">
      <y:ShapeNode>
        <y:Geometry height="49.40234375" width="59.99609375" x="178.6984406001984" y="164.103515625"/>
        <y:Fill color="#CCCCFF" transparent="false"/>
        <y:BorderStyle color="#000000" type="line" width="1.0"/>
        <y:NodeLabel alignment="center" autoSizePolicy="content" fontFamily="Dialog" fontSize="12" fontStyle="plain" hasBackgroundColor="false" hasLineColor="false" height="33.40234375" modelName="internal" modelPosition="c" textColor="#000000" visible="true" width="49.99609375" x="5.0" y="8.0">5) ,
COMMA</y:NodeLabel>
        <y:Shape type="rectangle"/>
      </y:ShapeNode>
    </data>
  </node>
  <node id="n5">
    <data key="d3"><![CDATA[6]]></data>
    <data key="d4"><![CDATA[este]]></data>
    <data key="d5"><![CDATA[Vaip3s]]></data>
    <data key="d6"><![CDATA[fi]]></data>
    <data key="d7"/>
    <data key="d9"/>
    <data key="d10">
      <y:ShapeNode>
        <y:Geometry height="49.40234375" width="50.69140625" x="268.694831969246" y="164.103515625"/>
        <y:Fill color="#CCCCFF" transparent="false"/>
        <y:BorderStyle color="#000000" type="line" width="1.0"/>
        <y:NodeLabel alignment="center" autoSizePolicy="content" fontFamily="Dialog" fontSize="12" fontStyle="plain" hasBackgroundColor="false" hasLineColor="false" height="33.40234375" modelName="internal" modelPosition="c" textColor="#000000" visible="true" width="40.69140625" x="5.0" y="8.0">6) este
Vaip3s</y:NodeLabel>
        <y:Shape type="rectangle"/>
      </y:ShapeNode>
    </data>
  </node>
  <node id="n6">
    <data key="d3"><![CDATA[7]]></data>
    <data key="d4"><![CDATA[căsătorit]]></data>
    <data key="d5"><![CDATA[Vmp--sm]]></data>
    <data key="d6"><![CDATA[căsători]]></data>
    <data key="d7"/>
```

```
    <data key="d9"/>
    <data key="d10">
     <y:ShapeNode>
      <y:Geometry height="49.40234375" width="73.35546875" x="257.362800719246" y="243.505859375"/>
      <y:Fill color="#CCCCFF" transparent="false"/>
      <y:BorderStyle color="#000000" type="line" width="1.0"/>
      <y:NodeLabel alignment="center" autoSizePolicy="content" fontFamily="Dialog" fontSize="12" fontStyle="plain" hasBackgroundColor="false" hasLineColor="false" height="33.40234375" modelName="internal" modelPosition="c" textColor="#000000" visible="true" width="63.35546875" x="5.0" y="8.0">7) căsătorit
Vmp--sm</y:NodeLabel>
      <y:Shape type="rectangle"/>
     </y:ShapeNode>
    </data>
   </node>
   <node id="n7">
    <data key="d3"><![CDATA[8]]></data>
    <data key="d4"><![CDATA[și]]></data>
    <data key="d5"><![CDATA[Crssp]]></data>
    <data key="d6"><![CDATA[și]]></data>
    <data key="d7"/>
    <data key="d9"/>
    <data key="d10">
     <y:ShapeNode>
      <y:Geometry height="49.40234375" width="45.3359375" x="349.38645523313494" y="164.103515625"/>
      <y:Fill color="#CCCCFF" transparent="false"/>
      <y:BorderStyle color="#000000" type="line" width="1.0"/>
      <y:NodeLabel alignment="center" autoSizePolicy="content" fontFamily="Dialog" fontSize="12" fontStyle="plain" hasBackgroundColor="false" hasLineColor="false" height="33.40234375" modelName="internal" modelPosition="c" textColor="#000000" visible="true" width="35.3359375" x="5.0" y="8.0">8) și
Crssp</y:NodeLabel>
      <y:Shape type="rectangle"/>
     </y:ShapeNode>
    </data>
   </node>
   <node id="n8">
    <data key="d3"><![CDATA[9]]></data>
    <data key="d4"><![CDATA[are]]></data>
    <data key="d5"><![CDATA[Vmip3s]]></data>
    <data key="d6"><![CDATA[avea]]></data>
    <data key="d7"/>
    <data key="d9"/>
    <data key="d10">
     <y:ShapeNode>
      <y:Geometry height="49.40234375" width="54.013671875" x="1.001953125" y="164.103515625"/>
      <y:Fill color="#CCCCFF" transparent="false"/>
      <y:BorderStyle color="#000000" type="line" width="1.0"/>
      <y:NodeLabel alignment="center" autoSizePolicy="content" fontFamily="Dialog" fontSize="12" fontStyle="plain" hasBackgroundColor="false" hasLineColor="false" height="33.40234375" modelName="internal" modelPosition="c" textColor="#000000" visible="true" width="44.013671875" x="5.0" y="8.0">9) are
Vmip3s</y:NodeLabel>
      <y:Shape type="rectangle"/>
     </y:ShapeNode>
    </data>
   </node>
   <node id="n9">
    <data key="d3"><![CDATA[10]]></data>
    <data key="d4"><![CDATA[o]]></data>
    <data key="d5"><![CDATA[Tifsr]]></data>
    <data key="d6"><![CDATA[un]]></data>
```

```
    <data key="d7"/>
    <data key="d9"/>
    <data key="d10">
     <y:ShapeNode>
      <y:Geometry height="49.40234375" width="41.3515625" x="7.3330078125" y="312.908203125"/>
      <y:Fill color="#CCCCFF" transparent="false"/>
      <y:BorderStyle color="#000000" type="line" width="1.0"/>
      <y:NodeLabel alignment="center" autoSizePolicy="content" fontFamily="Dialog" fontSize="12" fontStyle="plain" hasBackgroundColor="false" hasLineColor="false" height="33.40234375" modelName="internal" modelPosition="c" textColor="#000000" visible="true" width="31.3515625" x="5.0" y="8.0">10) o
Tifsr</y:NodeLabel>
      <y:Shape type="rectangle"/>
     </y:ShapeNode>
    </data>
   </node>
   <node id="n10">
    <data key="d3"><![CDATA[11]]></data>
    <data key="d4"><![CDATA[fiică]]></data>
    <data key="d5"><![CDATA[Ncfsrn]]></data>
    <data key="d6"><![CDATA[fiică]]></data>
    <data key="d7"/>
    <data key="d9"/>
    <data key="d10">
     <y:ShapeNode>
      <y:Geometry height="49.40234375" width="56.017578125" x="0.0" y="243.505859375"/>
      <y:Fill color="#CCCCFF" transparent="false"/>
      <y:BorderStyle color="#000000" type="line" width="1.0"/>
      <y:NodeLabel alignment="center" autoSizePolicy="content" fontFamily="Dialog" fontSize="12" fontStyle="plain" hasBackgroundColor="false" hasLineColor="false" height="33.40234375" modelName="internal" modelPosition="c" textColor="#000000" visible="true" width="46.017578125" x="5.0" y="8.0">11) fiică
Ncfsrn</y:NodeLabel>
      <y:Shape type="rectangle"/>
     </y:ShapeNode>
    </data>
   </node>
   <node id="n11">
    <data key="d3"><![CDATA[12]]></data>
    <data key="d4"><![CDATA[.]]></data>
    <data key="d5"><![CDATA[PERIOD]]></data>
    <data key="d6"><![CDATA[.]]></data>
    <data key="d7"/>
    <data key="d9"/>
    <data key="d10">
     <y:ShapeNode>
      <y:Geometry height="49.40234375" width="60.0078125" x="424.72273995535716" y="164.103515625"/>
      <y:Fill color="#CCCCFF" transparent="false"/>
      <y:BorderStyle color="#000000" type="line" width="1.0"/>
      <y:NodeLabel alignment="center" autoSizePolicy="content" fontFamily="Dialog" fontSize="12" fontStyle="plain" hasBackgroundColor="false" hasLineColor="false" height="33.40234375" modelName="internal" modelPosition="c" textColor="#000000" visible="true" width="50.0078125" x="5.0" y="8.0">12) .
PERIOD</y:NodeLabel>
      <y:Shape type="rectangle"/>
     </y:ShapeNode>
    </data>
   </node>
   <node id="n12">
    <data key="d3"><![CDATA[0]]></data>
    <data key="d4"><![CDATA[Are 52 de ani , este căsătorit şi are o fiică . ]]></data>
    <data key="d5"><![CDATA[-]]></data>
```

```
    <data key="d6"><![CDATA[-]]></data>
    <data key="d7"/>
    <data key="d9"/>
    <data key="d10">
     <y:ShapeNode>
      <y:Geometry height="34.701171875" width="244.115234375" x="129.3108940972222" y="0.0"/>
      <y:Fill color="#FFCC00" transparent="false"/>
      <y:BorderStyle hasColor="false" type="line" width="1.0"/>
      <y:NodeLabel alignment="center" autoSizePolicy="content" fontFamily="Dialog" fontSize="12" fontStyle="plain" hasBackgroundColor="false" hasLineColor="false" height="18.701171875" modelName="internal" modelPosition="c" textColor="#000000" visible="true" width="234.115234375" x="5.0" y="8.0">Are 52 de ani , este căsătorit și are o fiică . </y:NodeLabel>
      <y:Shape type="rectangle"/>
     </y:ShapeNode>
    </data>
   </node>
   <edge id="e0" source="n1" target="n3">
    <data key="d12"><![CDATA[amod]]></data>
    <data key="d13"><![CDATA[2]]></data>
    <data key="d14"><![CDATA[4]]></data>
    <data key="d16"/>
    <data key="d17">
     <y:PolyLineEdge>
      <y:Path sx="0.0" sy="-24.701171875" tx="-14.00146484375" ty="24.701171875"/>
      <y:LineStyle color="#000000" type="line" width="1.0"/>
      <y:Arrows source="none" target="standard"/>
      <y:EdgeLabel alignment="center" distance="2.0" fontFamily="Dialog" fontSize="12" fontStyle="plain" hasBackgroundColor="false" hasLineColor="false" height="18.701171875" modelName="six_pos" modelPosition="tail" preferredPlacement="anywhere" ratio="0.5" textColor="#000000" visible="true" width="34.017578125" x="2.0000987676085487" y="-24.3212890625">amod<y:PreferredPlacementDescriptor angle="0.0" angleOffsetOnRightSide="0" angleReference="absolute" angleRotationOnRightSide="co" distance="-1.0" frozen="true" placement="anywhere" side="anywhere" sideReference="relative_to_edge_flow"/>
      </y:EdgeLabel>
      <y:BendStyle smoothed="false"/>
     </y:PolyLineEdge>
    </data>
   </edge>
   <edge id="e1" source="n2" target="n3">
    <data key="d12"><![CDATA[post]]></data>
    <data key="d13"><![CDATA[3]]></data>
    <data key="d14"><![CDATA[4]]></data>
    <data key="d16"/>
    <data key="d17">
     <y:PolyLineEdge>
      <y:Path sx="0.0" sy="-24.701171875" tx="14.00146484375" ty="24.701171875">
       <y:Point x="178.04529699900792" y="228.505859375"/>
       <y:Point x="134.69676184275792" y="228.505859375"/>
      </y:Path>
      <y:LineStyle color="#000000" type="line" width="1.0"/>
      <y:Arrows source="none" target="standard"/>
      <y:EdgeLabel alignment="center" distance="2.0" fontFamily="Dialog" fontSize="12" fontStyle="plain" hasBackgroundColor="false" hasLineColor="false" height="18.701171875" modelName="six_pos" modelPosition="tail" preferredPlacement="anywhere" ratio="0.5" textColor="#000000" visible="true" width="26.681640625" x="-35.01509423634366" y="-12.970703125">post<y:PreferredPlacementDescriptor angle="0.0" angleOffsetOnRightSide="0" angleReference="absolute" angleRotationOnRightSide="co" distance="-1.0" frozen="true" placement="anywhere" side="anywhere" sideReference="relative_to_edge_flow"/>
      </y:EdgeLabel>
      <y:BendStyle smoothed="false"/>
     </y:PolyLineEdge>
    </data>
   </edge>
   <edge id="e2" source="n3" target="n0">
```

```
    <data key="d12"><![CDATA[dobj]]></data>
    <data key="d13"><![CDATA[4]]></data>
    <data key="d14"><![CDATA[1]]></data>
    <data key="d16"/>
    <data key="d17">
     <y:PolyLineEdge>
      <y:Path sx="0.0" sy="-24.701171875" tx="-12.8369140625" ty="24.701171875">
       <y:Point x="120.69529699900792" y="134.103515625"/>
       <y:Point x="238.5315972222222" y="134.103515625"/>
      </y:Path>
      <y:LineStyle color="#000000" type="line" width="1.0"/>
      <y:Arrows source="none" target="standard"/>
      <y:EdgeLabel alignment="center" distance="2.0" fontFamily="Dialog" fontSize="12" fontStyle="plain" hasBackgroundColor="false" hasLineColor="false" height="18.701171875" modelName="six_pos" modelPosition="tail" preferredPlacement="anywhere" ratio="0.5" textColor="#000000" visible="true" width="26.6875" x="45.57439986940412" y="-27.970703125">dobj<y:PreferredPlacementDescriptor angle="0.0" angleOffsetOnRightSide="0" angleReference="absolute" angleRotationOnRightSide="co" distance="-1.0" frozen="true" placement="anywhere" side="anywhere" sideReference="relative_to_edge_flow"/>
      </y:EdgeLabel>
      <y:BendStyle smoothed="false"/>
     </y:PolyLineEdge>
    </data>
   </edge>
   <edge id="e3" source="n4" target="n0">
    <data key="d12"><![CDATA[punct]]></data>
    <data key="d13"><![CDATA[5]]></data>
    <data key="d14"><![CDATA[1]]></data>
    <data key="d16"/>
    <data key="d17">
     <y:PolyLineEdge>
      <y:Path sx="0.0" sy="-24.701171875" tx="-4.278971354166668" ty="24.701171875">
       <y:Point x="208.6964874751984" y="149.103515625"/>
       <y:Point x="247.0895399305555" y="149.103515625"/>
      </y:Path>
      <y:LineStyle color="#000000" type="line" width="1.0"/>
      <y:Arrows source="none" target="standard"/>
      <y:EdgeLabel alignment="center" distance="2.0" fontFamily="Dialog" fontSize="12" fontStyle="plain" hasBackgroundColor="false" hasLineColor="false" height="18.701171875" modelName="six_pos" modelPosition="tail" preferredPlacement="anywhere" ratio="0.5" textColor="#000000" visible="true" width="33.35546875" x="40.39305250379772" y="-42.2587890625">punct<y:PreferredPlacementDescriptor angle="0.0" angleOffsetOnRightSide="0" angleReference="absolute" angleRotationOnRightSide="co" distance="-1.0" frozen="true" placement="anywhere" side="anywhere" sideReference="relative_to_edge_flow"/>
      </y:EdgeLabel>
      <y:BendStyle smoothed="false"/>
     </y:PolyLineEdge>
    </data>
   </edge>
   <edge id="e4" source="n5" target="n0">
    <data key="d12"><![CDATA[conj]]></data>
    <data key="d13"><![CDATA[6]]></data>
    <data key="d14"><![CDATA[1]]></data>
    <data key="d16"/>
    <data key="d17">
     <y:PolyLineEdge>
      <y:Path sx="0.0" sy="-24.701171875" tx="4.278971354166664" ty="24.701171875">
       <y:Point x="294.040535094246" y="149.103515625"/>
       <y:Point x="255.64748263888887" y="149.103515625"/>
      </y:Path>
      <y:LineStyle color="#000000" type="line" width="1.0"/>
      <y:Arrows source="none" target="standard"/>
      <y:EdgeLabel alignment="center" distance="2.0" fontFamily="Dialog" fontSize="12" fontStyle="plain" hasBackgroundColor="false" hasLineColor="false" height="18.701171875" modelName="six_pos"
```

modelPosition="tail" preferredPlacement="anywhere" ratio="0.5" textColor="#000000" visible="true" width="26.013671875" x="-36.393044704861154" y="-42.2587890625">conj<y:PreferredPlacementDescriptor angle="0.0" angleOffsetOnRightSide="0" angleReference="absolute" angleRotationOnRightSide="co" distance="-1.0" frozen="true" placement="anywhere" side="anywhere" sideReference="relative_to_edge_flow"/>
</y:EdgeLabel>
<y:BendStyle smoothed="false"/>
</y:PolyLineEdge>
</data>
</edge>
<edge id="e5" source="n6" target="n5">
<data key="d12"><![CDATA[pred]]></data>
<data key="d13"><![CDATA[7]]></data>
<data key="d14"><![CDATA[6]]></data>
<data key="d16"/>
<data key="d17">
<y:PolyLineEdge>
<y:Path sx="0.0" sy="-24.701171875" tx="0.0" ty="24.701171875"/>
<y:LineStyle color="#000000" type="line" width="1.0"/>
<y:Arrows source="none" target="standard"/>
<y:EdgeLabel alignment="center" distance="2.0" fontFamily="Dialog" fontSize="12" fontStyle="plain" hasBackgroundColor="false" hasLineColor="false" height="18.701171875" modelName="six_pos" modelPosition="tail" preferredPlacement="anywhere" ratio="0.5" textColor="#000000" visible="true" width="28.017578125" x="2.0000077504960245" y="-24.3212890625">pred<y:PreferredPlacementDescriptor angle="0.0" angleOffsetOnRightSide="0" angleReference="absolute" angleRotationOnRightSide="co" distance="-1.0" frozen="true" placement="anywhere" side="anywhere" sideReference="relative_to_edge_flow"/>
</y:EdgeLabel>
<y:BendStyle smoothed="false"/>
</y:PolyLineEdge>
</data>
</edge>
<edge id="e6" source="n7" target="n0">
<data key="d12"><![CDATA[cc]]></data>
<data key="d13"><![CDATA[8]]></data>
<data key="d14"><![CDATA[1]]></data>
<data key="d16"/>
<data key="d17">
<y:PolyLineEdge>
<y:Path sx="0.0" sy="-24.701171875" tx="12.8369140625" ty="24.701171875">
<y:Point x="372.05442398313494" y="134.103515625"/>
<y:Point x="264.2054253472222" y="134.103515625"/>
</y:Path>
<y:LineStyle color="#000000" type="line" width="1.0"/>
<y:Arrows source="none" target="standard"/>
<y:EdgeLabel alignment="center" distance="2.0" fontFamily="Dialog" fontSize="12" fontStyle="plain" hasBackgroundColor="false" hasLineColor="false" height="18.701171875" modelName="six_pos" modelPosition="tail" preferredPlacement="anywhere" ratio="0.5" textColor="#000000" visible="true" width="16.0" x="-61.9244881766183" y="-27.970703125">cc<y:PreferredPlacementDescriptor angle="0.0" angleOffsetOnRightSide="0" angleReference="absolute" angleRotationOnRightSide="co" distance="-1.0" frozen="true" placement="anywhere" side="anywhere" sideReference="relative_to_edge_flow"/>
</y:EdgeLabel>
<y:BendStyle smoothed="false"/>
</y:PolyLineEdge>
</data>
</edge>
<edge id="e7" source="n8" target="n0">
<data key="d12"><![CDATA[conj]]></data>
<data key="d13"><![CDATA[9]]></data>
<data key="d14"><![CDATA[1]]></data>
<data key="d16"/>
<data key="d17">
<y:PolyLineEdge>
<y:Path sx="0.0" sy="-24.701171875" tx="-21.394856770833332" ty="24.701171875">

```
          <y:Point x="28.0087890625" y="119.103515625"/>
          <y:Point x="229.97365451388885" y="119.103515625"/>
        </y:Path>
        <y:LineStyle color="#000000" type="line" width="1.0"/>
        <y:Arrows source="none" target="standard"/>
        <y:EdgeLabel alignment="center" distance="2.0" fontFamily="Dialog" fontSize="12" fontStyle="plain" hasBackgroundColor="false" hasLineColor="false" height="18.701171875" modelName="six_pos" modelPosition="tail" preferredPlacement="anywhere" ratio="0.5" textColor="#000000" visible="true" width="26.013671875" x="87.97559678819442" y="-42.9560546875">conj<y:PreferredPlacementDescriptor angle="0.0" angleOffsetOnRightSide="0" angleReference="absolute" angleRotationOnRightSide="co" distance="-1.0" frozen="true" placement="anywhere" side="anywhere" sideReference="relative_to_edge_flow"/>
        </y:EdgeLabel>
        <y:BendStyle smoothed="false"/>
      </y:PolyLineEdge>
    </data>
  </edge>
  <edge id="e8" source="n9" target="n10">
    <data key="d12"><![CDATA[det]]></data>
    <data key="d13"><![CDATA[10]]></data>
    <data key="d14"><![CDATA[11]]></data>
    <data key="d16"/>
    <data key="d17">
      <y:PolyLineEdge>
        <y:Path sx="0.0" sy="-24.701171875" tx="0.0" ty="24.701171875"/>
        <y:LineStyle color="#000000" type="line" width="1.0"/>
        <y:Arrows source="none" target="standard"/>
        <y:EdgeLabel alignment="center" distance="2.0" fontFamily="Dialog" fontSize="12" fontStyle="plain" hasBackgroundColor="false" hasLineColor="false" height="18.701171875" modelName="six_pos" modelPosition="tail" preferredPlacement="anywhere" ratio="0.5" textColor="#000000" visible="true" width="20.681640625" x="2.0" y="-19.3115234375">det<y:PreferredPlacementDescriptor angle="0.0" angleOffsetOnRightSide="0" angleReference="absolute" angleRotationOnRightSide="co" distance="-1.0" frozen="true" placement="anywhere" side="anywhere" sideReference="relative_to_edge_flow"/>
        </y:EdgeLabel>
        <y:BendStyle smoothed="false"/>
      </y:PolyLineEdge>
    </data>
  </edge>
  <edge id="e9" source="n10" target="n8">
    <data key="d12"><![CDATA[dobj]]></data>
    <data key="d13"><![CDATA[11]]></data>
    <data key="d14"><![CDATA[9]]></data>
    <data key="d16"/>
    <data key="d17">
      <y:PolyLineEdge>
        <y:Path sx="0.0" sy="-24.701171875" tx="0.0" ty="24.701171875"/>
        <y:LineStyle color="#000000" type="line" width="1.0"/>
        <y:Arrows source="none" target="standard"/>
        <y:EdgeLabel alignment="center" distance="2.0" fontFamily="Dialog" fontSize="12" fontStyle="plain" hasBackgroundColor="false" hasLineColor="false" height="18.701171875" modelName="six_pos" modelPosition="tail" preferredPlacement="anywhere" ratio="0.5" textColor="#000000" visible="true" width="26.6875" x="2.0" y="-24.3212890625">dobj<y:PreferredPlacementDescriptor angle="0.0" angleOffsetOnRightSide="0" angleReference="absolute" angleRotationOnRightSide="co" distance="-1.0" frozen="true" placement="anywhere" side="anywhere" sideReference="relative_to_edge_flow"/>
        </y:EdgeLabel>
        <y:BendStyle smoothed="false"/>
      </y:PolyLineEdge>
    </data>
  </edge>
  <edge id="e10" source="n11" target="n0">
    <data key="d12"><![CDATA[punct]]></data>
    <data key="d13"><![CDATA[12]]></data>
    <data key="d14"><![CDATA[1]]></data>
```

```
    <data key="d16"/>
    <data key="d17">
     <y:PolyLineEdge>
      <y:Path sx="0.0" sy="-24.701171875" tx="21.39485677083333" ty="24.701171875">
       <y:Point x="454.72664620535716" y="119.103515625"/>
       <y:Point x="272.76336805555553" y="119.103515625"/>
      </y:Path>
      <y:LineStyle color="#000000" type="line" width="1.0"/>
      <y:Arrows source="none" target="standard"/>
      <y:EdgeLabel alignment="center" distance="2.0" fontFamily="Dialog" fontSize="12" fontStyle="plain" hasBackgroundColor="false" hasLineColor="false" height="18.701171875" modelName="six_pos" modelPosition="tail" preferredPlacement="anywhere" ratio="0.5" textColor="#000000" visible="true" width="33.35546875" x="-107.65938129727806" y="-42.9560546875">punct<y:PreferredPlacementDescriptor angle="0.0" angleOffsetOnRightSide="0" angleReference="absolute" angleRotationOnRightSide="co" distance="-1.0" frozen="true" placement="anywhere" side="anywhere" sideReference="relative_to_edge_flow"/>
      </y:EdgeLabel>
      <y:BendStyle smoothed="false"/>
     </y:PolyLineEdge>
    </data>
   </edge>
   <edge id="e11" source="n0" target="n12">
    <data key="d12"/>
    <data key="d13"/>
    <data key="d14"/>
    <data key="d16"/>
    <data key="d17">
     <y:PolyLineEdge>
      <y:Path sx="0.0" sy="-24.701171875" tx="0.0" ty="17.3505859375"/>
      <y:LineStyle color="#000000" type="line" width="1.0"/>
      <y:Arrows source="none" target="standard"/>
      <y:EdgeLabel alignment="center" configuration="AutoFlippingLabel" distance="2.0" fontFamily="Dialog" fontSize="12" fontStyle="plain" hasBackgroundColor="false" hasLineColor="false" height="18.701171875" modelName="custom" preferredPlacement="anywhere" ratio="0.5" textColor="#000000" visible="true" width="38.6640625" x="10.66796502007378" y="-19.811523437500007">ROOT<y:LabelModel>
       <y:SmartEdgeLabelModel autoRotationEnabled="false" defaultAngle="0.0" defaultDistance="10.0"/>
      </y:LabelModel>
      <y:ModelParameter>
       <y:SmartEdgeLabelModelParameter angle="0.0" distance="30.0" distanceToCenter="true" position="right" ratio="0.5" segment="0"/>
      </y:ModelParameter>
      <y:PreferredPlacementDescriptor angle="0.0" angleOffsetOnRightSide="0" angleReference="absolute" angleRotationOnRightSide="co" distance="-1.0" frozen="true" placement="anywhere" side="anywhere" sideReference="relative_to_edge_flow"/>
      </y:EdgeLabel>
      <y:BendStyle smoothed="false"/>
     </y:PolyLineEdge>
    </data>
   </edge>
  </graph>
  <data key="d11">
   <y:Resources/>
  </data>
</graphml>
```

ANEXA 4. VERBELE SELECTATE PENTRU A FACE PARTE DIN TREEBANK. FRECVENȚELE LOR ÎN ROMBAC ȘI ÎN TREEBANK, ÎN SUBSECȚIUNILE CORESPUNZĂTOARE

Verbe Jurnalistic	Frecvența		Verbe Literar	Frecvența		Verbe Academic	Frecvența	
	R	T		R	T		R	T
putea	25444	70	putea	30664	37	apărea	6388	16
face	12082	26	face	29568	57	face	6030	27
vinde	9707	10	vedea	16894	27	publica	4448	8
privi	8946	19	spune	16207	20	scrie	4191	14
oferi	8819	14	da	15438	35	putea	4155	30
afla	8159	28	ști	15213	20	deveni	3596	14
organiza	7825	9	trebui	12721	29	urma	3571	11
începe	7466	23	zice	12561	10	semna	3235	8
trebui	7259	24	lua	9826	10	colabora	3028	4
avea_loc	6693	14	veni	9461	24	da	2803	13
urma	6687	11	crede	8372	12	debuta	2666	5
executa	6426	12	vrea	8311	5	rămâne	2416	22
desfășura	6313	5	părea	7963	10	începe	2265	14
prezenta	6024	19	pune	7719	19	afla	1810	10
obține	5980	6	rămâne	6995	13	lua	1804	13
realiza	5535	4	trece	6791	10	susține	1769	9
cuprinde	4691	8	sta	6647	11	trece	1674	10
primi	4549	15	lăsa	6618	15	pune	1664	8
participa	4462	7	începe	6447	19	traduce	1635	10
asigura	4437	9	vorbi	6315	8	părea	1538	8
stabili	4306	8	uita	6084	8	vedea	1528	5
acorda	4129	6	privi	6083	6	cuprinde	1464	6
efectua	4060	4	duce	6015	16	continua	1423	4
anunța	4022	20	simți	5601	9	încerca	1419	7
beneficia	3959	8	ajunge	5488	5	reprezenta	1372	5
pune	3939	8	găsi	5422	8	tipări	1370	8
reprezenta	3752	17	înțelege	5307	8	aduce	1354	4
lua	3586	15	auzi	5253	6	privi	1347	9
prevedea	3585	11	afla	5133	8	absolvi	1333	4
repara	3559	8	ține	4984	9	lucra	1281	3
da	3556	8	aduce	4885	9	veni	1257	4
exista	3499	9	întreba	4781	2	intra	1240	6
declara	3404	16	arăta	4421	2	edita	1220	6
deschide	3294	2	ieși	4371	13	conduce	1186	4
cumpăra	3175	6	trăi	4361	3	aparține	1173	6
susține	3100	4	exista	4318	3	realiza	1166	5
ajunge	3087	15	ridica	4274	9	ajunge	1153	7
spune	3085	3	deveni	4256	2	propune	1124	7
dori	3008	10	cunoaște	4143	8	trebui	1115	8
deveni	2912	13	intra	4129	12	alcătui	1110	3

informa	2893	3	scrie	4032	10	ține	1097	3
înregistra	2890	7	aștepta	3916	17	lipsi	1091	5
veni	2833	8	gândi	3903	6	trăi	1086	8
încheia	2721	2	căuta	3779	6	cunoaște	1072	8
aduce	2709	11	pleca	3778	9	constitui	1072	7
solicita	2706	10	merge	3738	5	considera	1053	7
trece	2701	8	întoarce	3571	7	prezenta	1049	3
depune	2647	10	cere	3537	4	spune	1037	7
include	2618	3	răspunde	3473	5	numi	1035	5
rămâne	2548	7	iubi	3259	4	înscrie	1008	6
intra	2528	6	muri	3059	3	lăsa	991	7
găsi	2504	6	primi	3038	9	include	985	5
aproba	2493	8	cădea	2994	6	urmări	983	10
situa	2386	6	pierde	2968	4	obține	977	4
monta	2283	6	scoate	2776	20	vrea	964	5
lansa	2267	6	voi	2732	7	oferi	954	2
preciza	2210	6	urma	2672	11	impune	902	2
constitui	2209	9	opri	2632	5	vorbi	896	3
conduce	2067	10	încerca	2611	4	exista	891	3
continua	2000	3	citi	2577	5	găsi	891	2
propune	1995	6	arunca	2561	5	reveni	881	4
închiria	1973	2	apărea	2428	3	duce	881	2
vedea	1941	5	deschide	2397	6	scoate	874	5
naște	1902	11	bate	2377	2	afirma	869	3
aplica	1883	5	purta	2354	7	dovedi	847	5
adresa	1867	4	așeza	2258	9	stabili	837	7
folosi	1861	8	numi	2257	4	ocupa	819	3
înscrie	1857	6	plăcea	2227	3	remarca	813	5
apărea	1831	4	însemna	2225	4	sta	808	6
ridica	1816	7	trage	2192	6	conține	808	3
derula	1760	4	asculta	2185	2	ilustra	807	4
lucra	1742	8	râde	2156	2	funcționa	789	5
ocupa	1681	4	ruga	2086	5	trimite	780	3
lega	1671	7	lipsi	1978	5	reuși	779	9
plăti	1659	8	prinde	1970	3	participa	774	4
elibera	1647	5	striga	1966	3	primi	770	6
ține	1647	2	trimite	1939	6	păstra	765	4
trimite	1599	7	continua	1873	4	arăta	759	6
deține	1579	2	dori	1861	2	folosi	759	3
cunoaște	1540	6	apropia	1818	4	căuta	743	2
permite	1540	5	petrece	1816	6	purta	737	4
găzdui	1531	4	scăpa	1804	3	ști	731	6
funcționa	1525	5	chema	1797	7	distinge	722	2
considera	1524	2	schimba	1788	2	întoarce	715	8
căuta	1517	2	recunoaște	1764	3	citi	708	9
viza	1512	7	naște	1719	3	deschide	703	3
pleca	1489	8	învăța	1703	5	relua	698	3

emite	1469	6	apuca	1668	7	alege	697	2
dispune	1461	5	juca	1616	2	aduna	692	3
decide	1434	9	întinde	1589	7	acorda	682	2
reuși	1402	16	coborî	1575	2	pleca	679	4
semna	1375	4	cuprinde	1537	3	lega	678	4
ști	1371	6	atinge	1533	2	marca	676	2
invita	1368	2	ascunde	1526	2	exprima	675	4
duce	1365	4	dovedi	1502	6	defini	658	2
alege	1364	3	plânge	1487	2	opri	654	6
răspunde	1363	4	strânge	1471	7	manifesta	651	3
pregăti	1356	7	crește	1469	2	ieși	636	2
scrie	1353	2	observa	1442	3	porni	629	4
câștiga	1333	3	porni	1429	5	postfața	626	2
crește	1331	6	reprezenta	1427	3	adăuga	623	2
depăși	1331	3	trezi	1425	2	cultiva	620	3
cere	1327	14	fugi	1424	2	frecventa	611	3
reveni	1309	2	mânca	1421	4	însoți	599	3
sta	1306	4	dormi	1407	2	descoperi	597	2
achita	1306	2	produce	1400	2	redacta	596	4
schimba	1296	6	plăti	1398	10	crede	595	2
publica	1287	4	descoperi	1377	2	muri	593	4
transmite	1287	3	lega	1357	7	desfășura	589	4
utiliza	1287	3	alege	1352	4	contribui	577	7
produce	1280	4	cânta	1347	2	învăța	573	2
implica	1264	4	suferi	1328	5	crea	572	5
comercializa	1258	2	povesti	1324	6	transpune	572	4
recomanda	1252	2	tăcea	1321	3	anunța	572	4
purta	1229	7	hotărî	1315	6	dedica	566	3
înființa	1220	5	închide	1315	2	pierde	566	3
programa	1210	4	urca	1291	2	atrage	564	3
apela	1206	2	zâmbi	1288	4	evoca	563	2
arăta	1196	3	păstra	1272	2	înțelege	559	5
împlini	1194	3	căra	1256	3	aborda	559	5
încerca	1174	10	acoperi	1250	2	simți	556	4
îndeplini	1172	3	aminti	1233	3	naște	552	3
întâlni	1165	6	crea	1218	4	construi	546	5
forma	1164	2	ocupa	1194	2	compune	543	2
prelua	1163	3	arde	1192	5	cădea	543	2
învinge	1159	3	părăsi	1192	2	surprinde	538	3
marca	1155	4	lucra	1186	3	întâlni	534	6
construi	1148	3	pricepe	1186	2	apropia	533	3
aparține	1137	4	explica	1177	5	consacra	525	3
verifica	1129	6	reuși	1165	8	angaja	515	4
urmări	1121	2	bea	1165	2	juca	497	2
finaliza	1116	3	dispărea	1161	2	reproduce	493	5
menționa	1112	6	urmări	1154	2	produce	492	3
vizita	1079	2	pătrunde	1145	2	forma	491	2

achiziționa	1070	2	prezenta	1141	3	figura	489	3
adăuga	1062	2	declara	1141	2	aminti	484	2
întocmi	1046	2	forma	1127	2	studia	483	2
însoți	1045	4	câștiga	1110	2	schimba	479	4
impune	1032	2	mișca	1103	2	releva	473	2
interesa	1024	4	convinge	1098	2	situa	472	2
crea	1022	6	adăuga	1085	2	observa	467	3
întreba	1019	2	repeta	1084	3	comenta	453	4
reține	1017	4	mira	1075	3	domina	452	7
juca	1010	4	ajuta	1067	12	reține	450	2
părea	999	7	susține	1061	2	însemna	447	3
încadra	998	4	folosi	1054	5	transforma	446	2
conține	984	4	apăra	1050	2	vădi	442	4
pierde	982	5	interesa	1042	4	înființa	437	2
aștepta	981	4	cumpăra	1036	6	preda	435	3
rezulta	967	4	sări	1032	3	preocupa	433	3
costa	966	3	sili	1026	3	asigura	433	2
lăsa	962	6	visa	1020	2	viza	432	2
ajuta	961	2	termina	1011	6	reuni	425	2
constata	952	4	mulțumi	999	5	depăși	423	3
referi	949	2	îmbrăca	981	3	prelua	422	4
petrece	934	8	îndrepta	972	2	organiza	421	2
afecta	931	6	pregăti	971	7	petrece	415	2
autoriza	925	4	propune	949	4	analiza	413	3
completa	923	5	închipui	945	2	număra	410	3
trăi	915	4	semăna	942	2	cere	406	2
modifica	914	6	lovi	925	2	dezvălui	405	7
circula	913	2	umple	923	2	provoca	404	2
accepta	911	3	reveni	922	2	interesa	399	6
numi	892	3	aduna	913	2	zice	398	2
aloca	892	3	considera	911	2	sugera	397	6
confecționa	886	2	suna	908	3	merge	395	3
introduce	884	4	stinge	905	4	regăsi	389	2
dovedi	880	2	ședea	904	2	caracteriza	389	2
dota	877	2	constata	901	2	termina	387	7
proveni	876	5	oferi	898	2	încheia	387	4
finanța	874	3	șopti	897	2	recunoaște	380	6
furniza	871	3	tăia	896	3	părăsi	376	8
adopta	868	6	permite	894	3	promova	376	2
promova	848	3	zări	893	2	consemna	375	3
scoate	847	5	băga	891	4	practica	362	4
recunoaște	843	2	bucura	890	4	răspunde	361	2
demara	840	3	săruta	888	4	implica	356	2
merge	838	5	teme	886	2	menționa	354	2
deplasa	834	3	supăra	884	4	cânta	351	2
elabora	829	7	sosi	864	6	trata	350	2
comunica	828	2	omorî	860	4	îndrepta	349	4

intenționa	825	7	aprinde	858	3	așeza	348	3
însemna	821	5	despărți	855	4	explica	347	5
evolua	816	4	vinde	848	2	căpăta	347	2
amenaja	816	3	spera	847	3	intitula	347	2
obliga	810	2	rupe	840	2	aprecia	346	2
disputa	808	2	renunța	834	7	semnala	345	5
locui	806	5	retrage	827	2	identifica	344	2
respecta	803	3	asigura	826	2	renunța	342	3
iniția	797	2	accepta	823	2	dezvolta	342	2
consta	785	2	feri	823	2	contura	337	4
reuni	781	3	pomeni	821	2	iniția	337	2
aminti	779	3	umbla	821	2	strânge	336	2
număra	773	2	lupta	817	2	mărturisi	336	2
crede	769	5	socoti	815	3	inspira	331	3
atrage	766	2	atrage	803	2	atinge	328	2
inaugura	764	2	servi	800	2	plasa	328	2
descoperi	763	3	ierta	789	2	deține	327	3
reduce	747	4	constitui	787	3	accepta	325	3
vorbi	747	2	căpăta	780	2	tinde	325	2
păstra	741	4	alerga	776	2	determina	324	3
aprecia	739	2	admite	766	3	descrie	324	2
specializa	735	2	mărturisi	765	4	prinde	321	5
mulțumi	730	3	înceta	765	2	fixa	321	2
hotărî	718	4	pretinde	756	2	întemeia	320	2
ieși	709	7	amesteca	753	2	tălmăci	316	2
angaja	699	2	rosti	750	2	declara	314	2
închide	689	5	conduce	749	2	ridica	311	2
acoperi	685	4	publica	747	5	culege	309	7
opri	672	2	da_seama	745	2	uita	309	3
prinde	671	5	grăbi	744	2	crește	306	3
determina	671	3	întrerupe	742	2	înregistra	306	2
plăcea	671	2	refuza	737	3	proiecta	305	5
discuta	668	5	atârna	735	2	demonstra	303	2
sosi	659	2	merita	734	2	elabora	303	2
acuza	658	7	tremura	729	2	ascunde	302	5
sprijini	648	2	încheia	727	2	adopta	300	2
suna	644	3	impune	725	3	bucura	299	2
afirma	637	7	discuta	707	2	pregăti	299	2
interveni	637	4	supune	705	3	iubi	297	2
practica	635	6	speria	700	3	refuza	297	2
rezolva	635	2	realiza	692	2	completa	295	2
învăța	631	4	sfârși	690	2	proveni	294	3
cuceri	623	3	bănui	685	4	sprijini	292	3
bucura	622	3	înșela	683	2	întâmpla	292	2
acumula	622	2	presupune	677	2	introduce	291	4
porni	619	2	locui	672	3	insera	291	2
contribui	612	3	preface	671	5	concepe	290	3

analiza	610	6	adresa	669	3	asuma	288	3
expune	610	5	ucide	669	3	reconstitui	284	2
desemna	605	4	judeca	669	2	înfățișa	284	2
înlocui	604	6	împărți	668	2	dori	282	2
preda	597	4	anunța	660	4	condamna	281	3
afișa	596	2	împinge	660	2	muta	279	5
relua	591	5	exprima	658	3	reflecta	276	2
muri	590	5	șterge	650	2	apăra	272	4
întoarce	586	3	stăpâni	647	5	socoti	272	2
distribui	585	2	amenința	646	2	preciza	271	2
califica	580	2	îndrăzni	646	2	pătrunde	270	3
studia	577	4	obține	645	2	balota	270	2
sărbători	575	2	izbucni	640	4	discuta	269	2
amplasa	573	3	izbuti	631	3	interpreta	269	2
cânta	572	3	culca	617	2	sublinia	268	2
identifica	572	2	introduce	615	2	îngriji	267	2
exprima	565	2	răsări	613	2	căsători	266	2
lipsi	563	3	plimba	610	6	suferi	264	2
provoca	562	5	stabili	608	2	aștepta	263	2
uita	553	3	relua	602	2	conferi	263	2
coordona	549	3	călca	593	2	respinge	262	2
dezvolta	541	3	provoca	593	2	închide	261	4
atesta	536	2	împiedica	589	5	chema	258	3
suferi	535	2	avea_nevoie	589	3	lansa	256	2
vui	533	3	transforma	585	2	indica	255	3
extinde	530	4	însoți	582	2	abandona	255	3
repartiza	530	2	muta	582	2	trage	255	2
înțelege	529	6	curge	581	2	alătura	254	3
spăla	527	2	lipi	581	2	gândi	252	3
citi	524	3	construi	580	4	utiliza	251	2
opta	524	2	afirma	577	4	recurge	249	4
contacta	523	4	compune	574	4	dispărea	249	2
supune	520	3	scula	570	2	transfera	246	4
cădea	520	3	potrivi	568	2	retrage	245	2
menține	519	3	străbate	568	2	câștiga	244	4
părăsi	518	4	fura	557	2	adresa	243	5
ruga	516	4	comunica	549	2	bate	243	2
conta	514	2	îndoi	548	2	evita	241	3
parcurge	511	3	prefera	547	2	găzdui	241	2
simți	500	5	ivi	546	4	beneficia	240	4
activa	496	2	mirosi	546	3	îmbina	239	6
datora	496	2	adormi	546	2	întreprinde	238	2
explica	495	3	decide	543	2	coborî	236	2
fora	495	2	suporta	543	2	baza	232	5
aborda	494	2	munci	541	3	evidența	231	2
presupune	493	5	aparține	541	2	supune	229	2
condamna	492	2	ataca	539	2	schița	223	5

opera	487	5	sui	537	5	aplica	221	2
majora	486	2	împlini	536	2	evolua	221	2
reaminti	484	3	smulge	521	2	expune	220	2
depista	472	3	dispune	520	3	formula	220	2
scădea	464	4	urî	520	3	asocia	220	2
sancționa	462	5	pieri	516	2	menține	219	3
calcula	461	3	țipa	511	2	povesti	218	2
edita	460	2	invita	507	2	prefața	218	2
sustrage	457	2	lumina	507	2	îndeplini	217	3
clasa	456	2	liniști	504	2	referi	217	2
interzice	454	3	costa	497	2	concentra	216	2
întâmpla	454	2	strica	496	2	întreține	215	4
oficia	451	2	reduce	495	4	orienta	215	2
sublinia	446	2	surprinde	494	2	înlocui	214	5
acționa	445	2	manifesta	492	2	activa	214	2
suporta	439	6	trata	490	2	dobândi	212	5
reglementa	439	2	pluti	489	2	avea_loc	212	3
semnala	438	4	reproduce	486	3	acoperi	210	4
dobândi	433	2	îngădui	484	2	ucide	210	2
renunța	429	2	rezulta	482	3	izbuti	209	3
dura	426	2	aplica	480	4	atesta	208	2
interpreta	424	2	cita	479	3	pleda	207	5
estima	422	3	sparge	479	2	plânge	207	2
atinge	422	3	dura	474	3	genera	206	3
răsări	421	2	ghici	474	2	străbate	206	3
debuta	419	2	vota	473	6	întrerupe	205	3
confirma	417	3	înainta	472	4	elibera	203	6
transforma	416	4	domni	472	3	dubla	203	2
alcătui	416	3	determina	471	5	prefera	203	2
instala	415	3	zăcea	469	2	grupa	202	3
remarca	412	5	opune	468	3	aresta	201	5
gândi	411	4	chinui	464	2	acuza	201	2
încasa	410	4	strecura	464	2	coace	201	2
elimina	410	2	uni	463	4	recomanda	200	2
percepe	409	6	imagina	461	2	întocmi	198	5
indica	409	2	descrie	457	2	reduce	198	4
consemna	409	2	zbura	457	2	ajuta	197	5
retrage	408	2	sprijini	457	2	auzi	197	2
arde	397	5	minți	456	2	scăpa	197	2
comite	397	2	străluci	454	5	stârni	196	3
peria	395	2	surâde	452	3	salva	196	2
necesita	391	5	inspira	451	4	selecta	196	2
ucide	391	4	interveni	449	3	refugia	195	5
lovi	390	3	apăsa	449	2	critica	194	2
suspenda	388	4	dezvolta	449	2	desprinde	193	4
absolvi	387	3	obișnui	447	4	integra	192	4
cerceta	387	2	repezi	446	3	stinge	192	4

accesa	385	3	risipi	445	5	coordona	192	2
baza	380	3	desface	445	2	opune	191	4
premia	379	3	clătina	445	2	influența	191	4
consulta	379	2	izbi	443	2	valorifica	191	2
urca	378	5	înconjura	443	2	comunica	190	4
refuza	377	4	rezista	441	2	cerceta	189	2
termina	376	2	durea	441	2	consta	189	2
spera	374	2	evita	440	5	atribui	188	4
garanta	373	3	cerceta	436	2	configura	188	2
împărți	370	2	încărca	435	2	relata	187	3
trage	369	4	învinge	433	4	locui	186	3
fabrica	369	2	fixa	429	2	permite	185	3
servi	366	2	înghiți	428	2	parcurge	185	2
figura	365	2	traduce	427	5	apela	184	2
prelungi	364	2	tresări	427	2	varia	184	2
promite	362	5	înălța	427	2	invoca	184	2
admite	360	3	conține	426	6	servi	181	4
controla	359	3	turbura	426	5	iscăli	181	2
călători	359	2	întrebuința	426	2	alterna	179	3
compune	357	2	obliga	424	2	arde	179	3
zice	357	2	promite	423	2	constata	179	2
aduna	351	2	respinge	423	2	imprima	178	6
auzi	350	4	lăuda	420	3	fugi	177	2
asista	342	2	mângâia	419	3	sfârși	177	2
sesiza	341	4	spăla	414	5	despărți	176	2
destina	340	2	întemeia	410	5	exclude	176	2
transporta	339	4	număra	410	2	data	175	2
evita	339	2	salva	409	2	rezulta	173	4
consuma	338	4	răsturna	408	4	arunca	172	2
solda	333	4	rătăci	402	2	presupune	171	3
anula	331	2	alcătui	399	2	reface	171	2
înnora	330	2	scutura	398	2	datora	171	2
confrunta	329	2	depăși	396	2	încadra	170	3
administra	326	3	păzi	394	2	inaugura	170	2
întreprinde	325	4	prevedea	393	5	trezi	168	4
deceda	324	2	lămuri	393	4	insista	168	3
tapița	324	2	cuveni	391	2	imagina	168	2
răni	323	4	acorda	389	4	spori	167	5
visa	323	2	distruge	389	2	stăpâni	167	2
dispărea	322	3	înlocui	388	3	trăda	167	2
colabora	322	3	distinge	387	4	încredința	166	3
demonstra	321	3	murmura	387	2	confirma	166	2
apropia	319	2	combate	385	3	împinge	166	2
muta	318	5	deosebi	385	2	axa	165	3
distruge	318	3	rezolva	383	4	urca	164	5
pronunța	318	2	studia	383	2	miza	164	2
regăsi	317	2	recomanda	382	4	structura	164	2

convinge	315	3	conta	379	2	anima	163	2
observa	315	3	alunga	378	4	ignora	161	3
manifesta	310	3	obosi	376	3	opera	161	2
recupera	310	3	consista	375	7	decerna	160	4
povesti	309	5	exclama	374	2	încărca	160	3
ataca	309	3	condamna	372	2	percepe	160	2
prefera	306	3	atribui	371	4	visa	159	2
respinge	303	8	stârni	368	3	fonda	158	4
dezlega	303	2	vărsa	367	2	rosti	157	2
selecta	301	4	spori	366	2	plăcea	156	5
trata	299	5	răspândi	360	5	lupta	156	2
proteja	299	4	desprinde	360	2	modifica	155	5
bate	299	2	pronunța	360	2	combate	155	2
scăpa	297	4	aproba	359	2	milita	154	4
strânge	296	2	urla	359	2	intenționa	153	2
evalua	295	4	cuceri	358	2	ataca	152	6
salva	295	4	încredința	356	2	înceta	152	2
dedica	294	2	asista	355	5	transmite	151	2
celebra	289	3	apleca	355	2	profesa	150	2
sufla	288	4	scădea	354	2	exercita	149	2
înceta	288	3	reflecta	352	5	rupe	149	2
îndrepta	286	5	pofti	352	2	călători	147	2
întruni	286	2	pasa	352	2	respecta	146	4
prilejui	281	4	protesta	351	3	republica	146	2
mări	280	4	ceda	350	2	cita	146	2
chema	280	3	desfășura	350	2	adapta	146	2
întrece	280	2	înscrie	349	3	recenza	145	2
surprinde	278	2	întîmpla	349	2	convinge	145	2
pătrunde	276	3	depune	348	2	hotărî	144	4
implementa	271	3	sugera	348	2	premia	144	4
dona	271	2	cugeta	347	4	împiedica	142	2
arunca	267	4	mușca	344	2	prilejui	142	2
coborî	267	2	depărta	343	3	judeca	141	3
vota	265	3	confirma	343	2	interveni	141	2
corespunde	264	2	prăbuși	343	2	examina	140	2
lupta	263	2	comite	342	4	anticipa	140	2
diversifica	263	2	inventa	342	3	ruja	140	2
evidenția	262	2	îndemna	342	2	obliga	139	4
preconiza	261	5	împărtăși	342	2	accentua	139	3
curge	259	2	admira	341	4	revela	139	3
declanșa	258	2	concepe	341	3	străbui	138	4
lichida	258	2	înfige	339	2	împrumuta	138	3
îmbunătăți	257	2	hrăni	337	4	împlini	136	2
genera	257	2	înlătura	336	4	asculta	135	5
amenința	255	3	avea_loc	335	4	mișca	135	4
institui	255	2	plictisi	334	2	declanșa	135	4
rașcheta	255	2	referi	333	4	recupera	135	3

presta	254	3	ferici	333	2	reapărea	134	4
releva	254	3	topi	332	3	nutri	134	2
oscila	253	2	aluneca	332	2	avansa	134	2
întrerupe	251	2	împăca	331	3	amesteca	133	2
întârzia	249	2	conveni	331	2	închina	133	2
diferi	248	4	jura	331	2	nota	131	5
transfera	248	3	datori	329	5	detașa	131	4
prepara	247	6	insista	328	2	pensiona	131	4
rezerva	247	4	mărgini	324	5	specializa	131	2
structura	245	2	preda	324	3	amplifica	130	4
înainta	244	3	încurca	324	2	extrage	130	3
așeza	243	5	angaja	324	2	exploata	130	2
fura	242	5	pica	324	2	opta	129	2
întâmpina	241	2	înregistra	323	3	diferi	127	2
grupa	240	2	răsuci	322	2	îmbogăți	126	3
totaliza	239	2	începea	321	3	traversa	126	2
concepe	238	4	îndeplini	320	4	preceda	126	2
fixa	238	2	organiza	318	3	circula	124	3
aresta	237	2	sfătui	318	3	prelucra	124	3
iubi	235	2	baza	316	4	înrudi	123	4
modera	233	3	trăda	315	5	rezista	122	5
atribui	233	2	risca	313	4	proceda	122	2
colecta	231	2	întări	310	2	rezuma	121	2
stinge	230	5	nimeri	309	2	înzestra	121	2
întinde	230	4	căsători	309	2	aspira	120	3
sigila	229	2	vârî	309	2	extinde	120	2
xerocopia	229	2	împrumuta	307	7	admite	120	2
ascunde	228	4	porunci	307	2	ceda	120	2
investi	228	2	reține	307	2	corespunde	119	3
facilita	226	2	ocoli	307	2	risipi	118	3
conferi	225	4	justifica	306	2	răspândi	117	5
formula	222	2	turna	305	3	îndepărta	117	3
recepționa	222	2	mărita	305	2	prețui	117	2
asculta	221	4	sorbi	303	2	inventa	116	3
diminua	220	5	regăsi	302	5	suporta	116	2
căsători	219	3	odihni	302	3	solicita	116	2
merita	219	3	trânti	302	3	prefigura	116	2
asuma	219	2	saluta	301	2	merita	115	2
cota	219	2	însărcina	300	4	uni	115	2
relata	218	6	însuși	300	2	dormi	115	2
restitui	217	4	exercita	299	2	limita	114	4
axa	217	3	răpi	298	4	ivi	114	4
tăia	217	3	respecta	297	3	instala	114	3
cifra	217	2	greși	295	4	stimula	114	3
onora	216	3	scuza	295	2	deosebi	114	2
defini	216	2	poseda	293	2	repeta	114	2
ara	215	5	respira	292	2	confrunta	113	4

redacta	214	3	comanda	290	2	transcrie	113	3
stipula	214	2	culege	289	2	însuși	113	2
specifica	213	3	luci	287	3	concretiza	113	2
readuce	212	3	transmite	287	2	compara	112	4
scuti	212	2	domina	286	4	împărți	112	4
apune	211	2	corespunde	285	4	impresiona	112	3
încuraja	211	2	zugrăvi	283	3	invita	112	3
mărturisi	210	2	îndepărta	283	2	reproșa	112	2
ilustra	210	2	consuma	282	2	lumina	111	5
desfunda	209	3	preocupa	281	6	dispune	111	5
vopsi	209	3	vizita	280	2	îmbrățișa	111	4
expedia	209	2	ploua	280	2	trasa	111	3
expira	209	2	dărui	279	2	contesta	111	3
cauza	208	3	îngriji	277	9	desemna	111	2
reieși	208	2	închina	277	6	îndemna	111	2
raporta	208	2	slăbi	277	2	suna	110	2
conveni	207	3	nega	276	2	pronunța	110	2
limita	207	2	indica	276	2	raporta	110	2
integra	207	2	răsuna	274	2	converti	109	3
depinde	207	2	îngropa	274	2	sluji	109	2
dărui	207	2	demonstra	273	4	predomina	108	5
acredita	205	3	așterne	273	2	justifica	108	3
cuveni	205	2	aranja	272	2	bântui	108	2
proceda	205	2	contribui	271	6	interzice	108	2
livra	204	4	mări	271	6	difuza	107	3
difuza	204	2	ofta	271	2	conta	107	2
asocia	203	3	elibera	271	2	îmbrăca	106	6
admira	203	2	deprinde	270	2	împleti	106	4
alătura	202	5	îmbrățișa	269	4	investiga	106	3
înmâna	202	2	abate	268	2	încuraja	105	3
dezbate	202	2	confunda	267	2	feri	105	2
avertiza	201	3	întrece	267	2	decide	104	5
vindeca	200	2	depinde	266	5	ocoli	104	3
apăra	199	4	sufla	266	3	tulbura	104	2
grăbi	198	2	întuneca	265	3	denunța	104	2
ciobăni	198	2	regreta	264	2	institui	104	2
comanda	197	2	dezvălui	262	2	întregi	104	2

Verbe	Frecvența		Verbe	Frecvența	
medical	R	T	Juridic	R	T
trebui	63053	114	trebui	29291	119
putea	55097	164	privi	27136	94
administra	26502	52	putea	25681	158
utiliza	26420	32	prevedea	20267	79
avea	22049	37	avea	14601	31
vedea	19766	27	aplica	13445	50

conține	16088	17	stabili	12844	46
observa	15529	26	menționa	10031	35
trata	15185	13	adopta	9523	29
privi	15176	42	prezenta	8638	27
apărea	15071	25	face	8435	31
lua	14362	64	utiliza	7780	17
prezenta	13174	19	efectua	6399	32
crește	12424	25	urma	6168	10
recomanda	11402	23	modifica	6153	13
păstra	10204	4	conține	5976	13
folosi	9543	22	anexa	5727	15
determina	9418	18	asigura	5661	36
raporta	9178	24	lua	5526	35
efectua	8602	36	include	5271	18
include	8413	7	depăși	5139	6
exista	8259	24	considera	5011	18
asocia	7443	23	înlocui	4764	11
face	6862	2	permite	4415	9
reduce	6274	15	obține	4287	14
scădea	5515	5	indica	4215	21
spune	5361	13	intra	4105	23
primi	4889	3	acorda	4006	21
conduce	4642	16	furniza	3897	27
indica	4629	18	ține	3734	10
ruga	4602	13	îndeplini	3669	13
produce	4461	8	cuprinde	3628	9
demonstra	4345	22	autoriza	3484	11
citi	4212	11	proveni	3225	8
afecta	4207	13	exista	2997	16
obține	3992	2	reprezenta	2994	18
evalua	3954	19	aduce	2936	25
controla	3938	7	folosi	2923	10
urma	3821	5	solicita	2921	18
adresa	3490	7	defini	2720	18
arăta	3461	3	pune	2625	5
începe	3450	5	respecta	2562	12
lega	3332	6	produce	2541	11
continua	3319	9	informa	2466	17
necesita	3252	9	realiza	2446	10
lăsa	3151	2	constitui	2429	14
elibera	3121	2	adăuga	2372	8
atinge	3099	6	determina	2355	8
studia	3078	7	vedea	2194	3
elimina	3064	9	comunica	2169	14
injecta	3029	8	lega	2085	15
cunoaște	2983	15	înregistra	2052	3
deveni	2973	8	însemna	2042	7

evidenția	2903	5	calcula	2038	6
rămâne	2830	18	transmite	2034	7
cuprinde	2787	2	decide	1990	2
întrerupe	2786	7	supune	1932	8
depăși	2730	12	afla	1919	6
ține	2726	13	elibera	1913	5
asigura	2605	5	recunoaște	1904	7
stabili	2526	3	publica	1891	16
duce	2492	7	primi	1883	5
menține	2370	5	reglementa	1873	13
proteja	2363	3	beneficia	1829	7
provoca	2299	2	rezulta	1782	4
modifica	2231	2	evalua	1735	3
prescrie	2225	8	începe	1733	7
considera	2208	4	desemna	1733	6
pune	2181	6	încheia	1730	13
filma	2120	2	deține	1722	2
evita	2061	2	baza	1695	7
informa	2051	10	preciza	1695	6
ajuta	2036	9	descrie	1632	12
dezvolta	2030	2	specifica	1608	13
monitoriza	1995	3	acoperi	1594	3
aștepta	1906	4	însoți	1591	6
congela	1906	2	constata	1582	5
suferi	1904	6	referi	1575	8
induce	1862	7	elimina	1563	9
discuta	1861	11	emite	1561	7
proveni	1816	13	introduce	1544	3
verifica	1789	8	depune	1534	9
acționa	1728	17	garanta	1503	15
răspunde	1712	6	corespunde	1487	3
reprezenta	1661	7	enumera	1470	8
scoate	1656	8	institui	1446	11
afla	1656	4	notifica	1438	14
inhiba	1634	12	oferi	1433	5
arunca	1631	2	situa	1402	2
îndemâna	1602	2	impune	1401	6
realiza	1580	7	reduce	1378	2
excreta	1535	4	afecta	1350	8
aproba	1499	2	rămâne	1333	18
lua_în_considerare	1485	3	aproba	1332	6
varia	1482	2	declara	1315	8
metaboliza	1473	7	coopera	1310	6
influența	1447	9	plăti	1306	10
preveni	1445	2	exprima	1299	2
contacta	1436	5	crește	1298	2
sugera	1434	4	limita	1281	6

alăpta	1377	6	desfășura	1266	7
comercializa	1364	2	implica	1260	6
defini	1352	2	da	1242	2
înregistra	1284	2	verifica	1240	12
întreba	1275	9	confirma	1231	7
furniza	1271	3	cere	1226	4
schimba	1271	3	clasifica	1171	2
trece	1268	3	identifica	1153	6
implica	1261	14	participa	1150	16
permite	1256	15	accepta	1142	8
doza	1253	13	încorpora	1142	8
compara	1252	5	scădea	1114	2
da	1193	4	încadra	1103	2
limita	1172	7	dispune	1099	5
iniția	1150	2	funcționa	1097	10
stimula	1143	13	vinde	1093	2
exclude	1142	12	întocmi	1084	7
descrie	1130	7	adresa	1079	4
găsi	1110	4	exercita	1069	12
uita	1083	7	demonstra	1067	2
baza	1075	15	continua	1023	13
clorura	1074	3	contribui	1020	10
infecta	1063	6	păstra	1004	4
anexa	1053	2	atribui	989	5
impune	1044	2	exclude	985	4
estima	1023	5	înțelege	982	3
măsura	1017	5	examina	969	8
înscrie	1007	3	semna	959	8
manifesta	998	3	răspunde	959	6
identifica	997	2	completa	959	5
îndepărta	981	2	aparține	956	7
destina	978	3	trimite	950	6
simți	964	5	menține	950	4
constata	963	2	arăta	946	2
mânca	928	3	diferi	943	5
diferi	920	11	importa	942	11
contraindica	885	2	măsura	901	7
introduce	877	4	exporta	895	2
susține	870	4	compune	889	2
dispărea	861	4	raporta	882	4
reveni	856	3	susține	869	5
aplica	855	7	avea_loc	868	4
dovedi	852	3	prepara	854	2
adecva	850	4	deveni	823	4
consulta	849	6	apărea	821	2
părea	839	6	justifica	820	8
decide	811	14	ridica	816	2

dori	810	10	consta	809	2
ajunge	802	6	forma	800	3
diminua	794	5	genera	798	4
însoți	786	2	destina	796	3
repeta	782	6	angaja	793	6
opri	779	8	conveni	791	5
modera	774	2	elabora	786	6
trage	770	5	viza	780	2
datora	751	7	litera	779	2
media	746	9	dovedi	778	6
semnala	741	2	conduce	778	3
ajusta	740	5	evita	776	4
cere	733	2	propune	770	7
fixa	730	2	figura	764	3
înceta	726	4	acționa	761	3
confirma	718	6	fabrica	749	4
explica	717	10	interzice	743	2
concentra	710	2	adapta	735	2
exprima	687	2	adecva	726	4
adăuga	682	11	hotărî	713	2
ști	676	6	controla	703	4
expune	674	5	atinge	688	2
funcționa	672	2	insera	688	2
apăsa	670	4	necesita	685	6
cauza	668	2	deschide	682	7
consta	664	3	retrage	677	7
absorbi	662	4	trata	673	4
detecta	661	2	certifica	672	8
depinde	656	7	ajunge	668	4
forma	647	4	împiedica	666	14
prelungi	646	2	concepe	666	2
randomiza	646	2	estima	662	3
amesteca	642	4	formula	661	2
corespunde	628	3	transfera	651	8
reconstitui	627	2	cauza	649	2
comprima	616	2	fixa	629	3
însemna	611	2	reflecta	610	2
înghiți	609	2	suporta	606	5
crede	596	6	duce	605	2
îmbunătăți	594	2	obliga	604	6
înrola	590	2	facilita	602	2
urmări	585	4	revizui	600	3
agita	584	3	transporta	594	2
pierde	582	2	înscrie	587	6
menționa	582	2	conferi	576	7
dizolva	573	2	expedia	567	10
referi	570	3	cunoaște	566	9

acorda	560	9	abroga	552	8
extrage	553	2	aloca	552	7
acoperi	552	10	veni	548	7
agrava	545	3	concluziona	548	2
marca	541	2	alege	545	2
investiga	539	7	recomanda	534	2
enumera	533	2	invita	532	10
relua	532	2	afirma	531	2
testa	518	7	proceda	530	4
programa	514	2	dori	525	5
spăla	507	4	asocia	525	2
actualiza	504	2	practica	523	8
rezulta	501	2	preleva	521	2
intenționa	498	6	decurge	520	2
caracteriza	493	2	reveni	518	6
atașa	487	3	admite	513	3
concepe	480	2	crea	507	2
prevedea	478	13	presupune	507	2
constitui	478	3	proteja	507	2
înlocui	475	3	consulta	504	9
avea_loc	474	4	enunța	500	4
aparține	472	7	deriva	496	2
dura	467	5	urmări	492	2
emite	455	10	purta	489	4
depune	453	2	extinde	488	7
întârzia	451	2	suferi	486	5
respecta	450	2	atesta	479	2
ridica	448	2	respinge	478	3
justifica	441	3	analiza	476	4
autoriza	439	4	conforma	476	3
calcula	437	2	finanța	474	6
deschide	429	6	scrie	468	11
recunoaște	428	2	prelungi	466	4
sfătui	425	3	asuma	462	9
avea_nevoie	419	2	condiționa	462	2
potența	417	5	prelucra	462	2
tolera	416	4	naște	458	4
împinge	413	3	numi	457	6
compensa	411	6	suspenda	453	3
bloca	398	7	reține	437	4
întâmpla	391	2	tăcea	436	2
inscripționa	387	3	recurge	435	10
încerca	387	2	separa	435	2
anunța	387	2	asista	433	9
cronica	384	2	contabiliza	433	2
împiedica	383	2	încuraja	428	3
contribui	382	6	refuza	421	5

prefera	381	2	achiziționa	420	2
omite	379	4	trece	418	2
răsuci	378	2	întreprinde	414	7
instrui	377	4	explica	410	2
consuma	369	6	organiza	408	2
diagnostica	369	2	înceta	406	5
dilua	368	2	sprijini	400	7
corela	367	4	actualiza	393	6
examina	356	8	observa	388	3
închide	356	5	depinde	383	2
oferi	355	2	expira	377	8
ameliora	355	2	echipa	372	2
îndrepta	354	2	opera	368	3
bea	350	5	tăia	364	2
alege	350	4	iniția	363	8
participa	349	10	îmbunătăți	361	2
detalia	349	2	percepe	358	2
analiza	348	2	colecta	357	3
intra	347	2	integra	357	2
activa	346	4	gestiona	356	7
persista	345	2	schimba	354	4
dializa	341	2	ajusta	351	7
reacționa	339	2	intenționa	351	2
desfășura	338	2	executa	343	2
împărți	336	2	consolida	340	2
roti	331	2	înființa	338	5
anemia	329	7	dobândi	333	4
omogeniza	326	3	expune	333	3
regăsi	325	2	preveni	333	2
mări	323	4	preceda	326	3
ambala	319	2	usca	321	3
remite	317	5	pune_în_aplicare	320	6
anticipa	315	2	monitoriza	320	2
planifica	314	2	monta	319	8
izola	314	2	fi_vorba	319	2
interacționa	312	6	găsi	317	3
selecta	311	2	promova	314	7
prepara	310	5	lucra	313	3
corecta	307	2	părea	310	3
purta	307	2	selecta	307	7
solicita	305	2	livra	306	2
surveni	305	2	interpreta	301	8
traversa	305	2	interveni	301	6
încrucișa	304	2	ocupa	299	4
transporta	300	2	pregăti	296	3
distribui	298	2	transforma	294	4
hotărî	297	2	influența	293	3

presupune	294	2	datora	291	6
intensifica	288	3	marca	291	2
aminti	283	2	înainta	290	5
întâlni	282	2	redacta	288	6
supraveghea	278	5	comercializa	284	4
avertiza	277	2	scuti	282	3
converti	276	2	deduce	278	4
situa	275	3	concentra	275	5
nota	274	2	provoca	273	2
clasifica	272	2	proiecta	268	4
reflecta	271	3	repartiza	266	2
predispune	271	2	scoate	265	6
transfera	271	2	pescui	264	3
curăța	268	2	corecta	263	2
suspecta	267	2	testa	262	6
reciti	267	2	depozita	260	2
lipsi	260	2	satisface	256	2
reapărea	257	3	dezvolta	256	2
încadra	249	2	negocia	255	4
retrage	249	2	închide	255	2
așeza	246	2	recupera	253	3
acumula	246	2	împărți	251	2
aduce	243	2	pondera	247	2
întoarce	241	2	mări	246	4
beneficia	239	2	sacrifica	245	4
înrăutăți	237	2	preconiza	245	2
atenționa	231	5	detalia	243	3
specializa	227	2	instala	242	10
regula	225	2	excepta	241	4
perfora	225	2	interesa	241	4
deriva	224	4	compara	239	2
pătrunde	224	2	însărcina	238	5
atribui	223	2	combina	238	2
extinde	217	2	extrage	237	2
sigila	215	13	servi	233	6
supune	215	2	ambala	233	2
avansa	214	2	alcătui	231	3
lovi	213	2	lăsa	231	2
compune	211	4	corobora	231	2
urina	211	2	ataşa	230	5
tulbura	210	2	citi	229	8
combina	208	2	avea_dreptul	228	8
apropia	206	2	alia	226	2
numi	202	2	lipsi	225	5
rupe	199	2	presta	224	2
adsorbi	196	2	compensa	224	2
economisi	195	2	finaliza	223	5

scurge	195	2	rezerva	223	2
deteriora	193	2	cumpăra	218	2
reîncepe	193	2	recolta	216	2
veni	193	2	exploata	215	2
merge	192	2	sublinia	215	2
auzi	192	2	coordona	214	3
favoriza	191	2	ruga	207	7
vaccina	189	2	anunța	207	2
pregăti	186	2	armoniza	206	2
diviza	183	2	culege	205	2
vizita	178	2	defalca	205	2
îndeplini	178	2	reieși	204	2
reînnoi	176	5	denumi	203	6
releva	175	3	construi	198	5
dezinfecta	175	2	rambursa	197	2
stabiliza	175	2	achita	195	2
obișnui	174	5	amesteca	192	2
facilita	169	5	reînnoi	191	4
încetini	167	2	părăsi	190	4
potrivi	167	2	plasa	190	2
amâna	166	2	surveni	186	4
accelera	165	10	reexamina	186	3
fi_nevoie	165	2	varia	186	2
înrudi	165	2	pronunța	185	3
încheia	161	2	lua_măsuri	184	2
vorbi	160	2	distribui	183	2
fabrica	159	2	reaminti	183	2
șterge	158	2	conserva	182	2
îngrijora	157	2	pierde	182	2
umple	157	2	înmulți	180	6
tinde	155	5	ajuta	180	3
îndulci	155	2	cântări	179	2
exercita	154	3	renunța	179	2
capsula	154	2	delimita	179	2
vindeca	153	2	compromite	178	2
recolta	151	2	detecta	177	2
angaja	150	2	converti	176	5
încălzi	149	2	sta	175	2
înțelege	149	2	contesta	173	2
avea_grijă	149	2	reuni	173	2
colecta	148	4	diminua	169	2
adera	148	2	cultiva	168	2
reține	147	2	invoca	168	2
debuta	146	3	detașa	167	6
sta	146	2	porni	167	4
lupta	145	2	remarca	166	2
preceda	145	2	strânge	165	5

distruge	144	2	favoriza	165	3
îmbolnăvi	143	2	aștepta	164	6
călători	142	2	încredința	164	2
instala	142	2	supraveghea	163	2
alcătui	139	2	contamina	162	3
institui	138	2	audia	162	2
depista	137	2	captura	161	4
interfera	137	2	majora	160	2
usca	134	2	întruni	157	2
număra	133	2	anula	157	2
compromite	133	2	evidenția	156	4
defecta	133	2	califica	155	2
grava	133	2	dota	154	7
accentua	132	2	semnala	154	5
clăti	131	5	confecționa	154	2
puți	131	2	încălca	153	2
juca	130	2	consuma	152	2
precipita	130	2	prelua	151	2
minimaliza	128	2	administra	151	2
amplifica	126	2	spori	150	2
perfuza	126	2	coincide	148	2
generaliza	126	2	axa	148	2
uni	126	2	dilua	147	7
recupera	125	2	curăța	147	4
conferi	124	2	întemeia	147	4
genera	124	2	încerca	146	2
prinde	124	2	congela	145	3
regla	123	3	valida	145	3
dona	122	2	poseda	145	2
specifica	121	5	imprima	144	3
îngroșa	121	2	conecta	143	5
excepta	120	2	lua_în_considerare	143	2
îmbuna	120	2	vărsa	143	2
deosebi	119	2	aglomera	143	2
naște	119	2	arbora	142	2
transforma	118	2	tranzacționa	142	2
deplasa	118	2	grupa	142	2
strânge	117	3	investi	142	2
separa	117	2	opune	141	3
denumi	117	2	deplasa	141	2
deșuruba	117	2	uni	140	2
micșora	116	3	atrage	138	2
autoadministra	116	2	suspecta	136	2
leșina	116	2	amâna	135	2
eșua	116	2	activa	134	5
imprima	114	11	restricționa	134	3
revizui	114	2	spăla	133	6

securiza	114	2	deteriora	133	4
evolua	113	4	vaccina	133	2
progresa	113	4	apropia	132	3
livra	113	2	adera	131	2
alterna	113	2	rectifica	130	2
rezolva	112	3	risca	129	3
înlătura	112	2	divulga	128	3
cumula	110	2	lamina	128	2
familiariza	110	2	motiva	127	8
suprima	109	2	șterge	127	3
recruta	108	4	ieși	127	2
transmite	108	3	încasa	127	2
secreta	108	2	relua	126	2
interveni	108	2	confrunta	126	2
descoperi	107	4	comite	126	2
apăra	107	2	plea	126	2
concluziona	107	2	caracteriza	124	2
declanșa	107	2	izola	124	2
aspira	106	2	denunța	123	14
pronunța	105	2	delega	123	3
accepta	105	2	dubla	123	2
reintroduce	105	2	îndepărta	122	4
finaliza	103	2	juca	122	4
vărsa	103	2	ceda	122	2
degrada	102	2	reproduce	122	2
transpira	102	2	simplifica	121	2
adapta	101	2	omologa	120	2
respira	101	2	căuta	119	5
arma	97	2	clarifica	119	2
atenua	97	2	deconta	117	2
alinia	96	2	rotunji	116	2
cântări	96	2	nota	116	2
conjuga	95	2	specializa	115	2
întinde	95	2	antrena	113	2
goli	95	2	imputa	113	2
lua_măsuri	94	2	soluționa	112	7
masca	93	3	trage	112	4
eticheta	92	2	cofinanța	112	2
ieși	92	2	întrerupe	112	2
reevalua	92	2	inspecta	111	2
reuși	92	2	remedia	109	2
spori	91	4	absorbi	108	2
mișca	91	2	descoperi	107	2
altera	91	2	distila	107	2
răsturna	88	4	opri	106	5
inactiva	88	2	ameliora	106	3
decela	87	2	sesiza	106	3

îngriji	87	2	înmatricula	106	2
mesteca	87	2	acumula	105	2
neutraliza	86	2	aprecia	105	2
exacerba	85	2	prevala	105	2
feri	85	2	parafa	104	5
gândi	84	2	aborda	104	3
cataliza	83	2	alimenta	103	4
încărca	83	2	pleca	103	2
reumple	83	2	carbura	102	2
termina	82	2	refugia	101	6
echilibra	82	2	amenința	101	2
deprima	82	2	orienta	99	2
reieși	82	2	cădea	99	2
subția	82	2	induce	99	2
dobândi	81	6	lansa	97	5
localiza	81	2	stoca	97	3
compărea	81	2	dizolva	97	2
comporta	81	2	manifesta	97	2
regresa	81	2	aproviziona	95	2
agrea	80	2	manipula	95	2
renunța	79	2	planifica	95	2
ocupa	79	2	încărca	93	2
pleca	79	2	consemna	92	5
desemna	78	2	restrânge	91	4

ANEXA 5: DISTRIBUŢIA ERORILOR DE ADNOTARE SINTACTICĂ AUTOMATĂ ÎN CADRUL PROCESULUI ITERATIV DE ADNOTARE/CORECTARE/RE-ANTRENARE

Acurateţea totală şi distribuţia acurareţei de adnotare pe etichete morfo-lexicale (în ordine descrescătoare a acurateţei de identificare corectă a centrului şi dependenţei: vezi ultima coloană din tabel). Au fost luate în considerare doar etichetele morfo-lexicale care apar de cel puţin 10 ori în setul analizat.

1) **Setul Literar 1**

Acurateţe	Acurateţe centru corect %	Acurateţe dependenţă corectă	Acurateţe ambele corecte
total	**68%**	**74%**	**59%**
Va--3	100%	100%	100%
Tifsr	99%	99%	99%
Timsr	98%	98%	98%
Di3fpr	95%	100%	95%
Va--3s	97%	94%	94%
Va--3p	93%	96%	93%
Ds3---p	92%	92%	92%
Dd3fsr---e	90%	90%	90%
Va--1	91%	94%	89%
Ncfsoy	91%	93%	88%
Qn	94%	94%	87%
Ncfpoy	91%	95%	86%
Pw---r	89%	88%	84%
Rp	98%	83%	83%
Tsfs	89%	89%	82%
Tsms	94%	88%	82%
Ncmpoy	91%	82%	82%
Qz	78%	96%	77%
Vmsp3	78%	84%	77%
Afpfson	92%	85%	77%
Ncfsrn	84%	78%	76%
Ncfp-n	84%	80%	75%
Pp3msr--------s	88%	75%	75%
Afpfp-n	79%	79%	74%
Afpf--n	82%	91%	73%
Np	80%	78%	72%
Vmip1s	74%	74%	71%
Vmip1p	71%	76%	71%
Afpfsry	71%	79%	71%
Vmip3p	74%	70%	70%

Ncmsoy	80%	73%	69%
Tf-so	75%	69%	69%
Ncfsry	79%	76%	68%
Ncfpry	77%	76%	68%
Ds3---s	84%	74%	68%
Ncms-n	79%	71%	67%
DBLQ	67%	100%	67%
Px3--d--y-----w	83%	78%	67%
Pd3fsr	75%	67%	67%
Di3fsr---e	67%	92%	67%
Vmnp	71%	69%	66%
Pp3fsa--------w	94%	65%	65%
Ncmsry	76%	71%	63%
Afpfsrn	76%	71%	63%
PERIOD	62%	96%	62%
Vmp--sm	67%	66%	62%
Rw	65%	85%	62%
Spca	66%	89%	61%
Vmii3s	65%	62%	60%
total	68%	74%	59%
DASH	64%	91%	59%
Vmip3s	60%	63%	58%
Vmis3s	66%	64%	58%
Mc-p-l	63%	58%	58%
Vmp--pf	58%	74%	58%
Pp3msa--------w	83%	67%	58%
Pi3-sr	92%	58%	58%
Rz	67%	75%	58%
Spsa	71%	77%	57%
Px3--a--------w	97%	57%	57%
Afpms-n	70%	66%	57%
Px3--a--y-----w	95%	59%	57%
Vanp	100%	57%	57%
Pi3mpr	71%	57%	57%
Pp3mpa--y-----w	83%	56%	56%
Ncfson	55%	64%	55%
Ncmp-n	65%	57%	54%
Ncmpry	67%	69%	51%
Vmp--sf	50%	55%	50%
Pp1-sd--y-----w	85%	50%	50%
Vmg-------y	58%	67%	50%
Px3--d--------w	92%	50%	50%
Vaip3s	50%	50%	50%
Pp2-sa--------w	100%	50%	50%
Pi3msr	60%	50%	50%
Rgp	54%	81%	49%

COMMA	48%	99%	47%
Pp3-sd--y-----w	88%	53%	47%
Pp3msa--y-----w	76%	48%	45%
Afpmp-n	67%	56%	44%
Vmil3s	53%	53%	43%
Va--2s	100%	43%	43%
Spcg	46%	75%	42%
Crssp	44%	78%	40%
Qs	52%	50%	40%
Vmis3p	55%	60%	40%
Vmp--pm	40%	47%	40%
Dd3msr---e	50%	90%	40%
Ccssp	55%	70%	38%
Csssp	57%	47%	36%
Spsay	46%	61%	32%
Vmip2s	36%	41%	32%
Vmg	41%	44%	30%
Pp3-sd--------w	83%	30%	30%
Vmii3p	35%	35%	29%
Vmil3p	41%	29%	29%
Pp1-pa--------w	94%	29%	29%
Vmis1s	36%	36%	29%
Pp1-sa--y-----w	86%	29%	29%
Pd3msr	27%	36%	27%
Vmii1	29%	34%	26%
Vmip3	28%	25%	25%
Pp1-sn--------s	37%	37%	21%
Cscsp	22%	44%	19%
Rc	35%	39%	16%
Di3fsr	15%	15%	15%
Qz-y	76%	14%	14%
Pp3fpa--------w	64%	14%	14%
Pp1-sd--------w	62%	8%	8%
Pp1-sa--------w	58%	4%	4%

2) Setul Academic 1

Acurateţe	**Acurateţe centru corect %**	**Acurateţe dependenţă corectă**	**Acurateţe ambele corecte**
total	**81%**	**85%**	**74%**
Va--3s	100%	100%	100%
Timsr	100%	100%	100%
Qz	100%	100%	100%
Px3--d--------w	100%	100%	100%
Afp-p-n	100%	100%	100%

Di3fpr	100%	100%	100%
Pw---r	98%	96%	96%
Tifsr	97%	98%	95%
Tsfs	98%	98%	95%
Va--3p	94%	94%	94%
Qn	94%	100%	94%
Ncmpry	94%	100%	94%
Tsms	93%	100%	93%
LPAR	93%	100%	93%
Tf-so	96%	92%	92%
Vaip3p	92%	92%	92%
Tsfp	92%	100%	92%
Pp3msr--------s	100%	92%	92%
Mc-p-l	100%	92%	92%
Vmsp3	91%	91%	91%
Rp	93%	90%	90%
Vmg-------y	90%	100%	90%
Vaip3s	89%	89%	89%
Afpfson	95%	93%	88%
Ncfsoy	89%	93%	87%
Ncmsoy	88%	90%	83%
PERIOD	82%	100%	82%
Ncfsry	88%	85%	81%
Ncfsrn	91%	84%	81%
Ncfpry	86%	89%	81%
Etd	90%	86%	80%
Ncfpoy	89%	82%	80%
Ncfson	80%	90%	80%
Afpfsrn	89%	83%	79%
Rw	79%	100%	79%
Ncmsry	85%	82%	78%
Np	88%	81%	77%
Afpfsry	77%	77%	77%
Vmip3	78%	77%	76%
Ncfp-n	83%	82%	75%
Ncms-n	84%	78%	74%
Vmis3s	76%	75%	74%
Px3--a--y-----w	100%	74%	74%
Px3--d--y-----w	91%	82%	73%
Afpms-n	85%	78%	72%
Vmp--sm	78%	74%	72%
Ed	84%	78%	72%
RPAR	76%	90%	72%
Afpfp-n	78%	80%	71%
Afpmp-n	81%	76%	71%
Ncmpoy	93%	79%	71%

Spsa	79%	83%	68%
Vmip3s	71%	71%	68%
Qs	91%	68%	68%
Ncmp-n	77%	74%	68%
Rgp	71%	86%	67%
Pp3msa--------w	100%	67%	67%
Px3--a--------w	99%	68%	66%
Vmip3p	72%	66%	66%
Vmp--sf	79%	67%	64%
DBLQ	65%	94%	63%
COMMA	62%	99%	62%
Vmg	61%	83%	61%
Spca	72%	81%	61%
Vmp--pf	64%	61%	58%
Vmii3s	57%	57%	57%
Pp3-sd--------w	100%	57%	57%
Vmnp	85%	63%	55%
Ccssp	66%	90%	55%
Afpmsry	55%	64%	55%
Csssp	64%	55%	55%
Ds3---s	53%	53%	53%
Crssp	55%	84%	52%
Rc	59%	67%	52%
Spsay	52%	78%	43%
Spcg	50%	92%	42%
DASH	30%	80%	30%
SLASH	36%	29%	25%

3) Setul Medical 1

Acurateţe	**Acurateţe centru corect %**	**Acurateţe dependenţă corectă**	**Acurateţe ambele corecte**
total	**82%**	**87%**	**77%**
Tifsr	100%	100%	100%
Vap--sm	100%	100%	100%
Dd3msr---e	100%	100%	100%
Tsfp	100%	100%	100%
Dd3fsr---e	100%	100%	100%
Va--3s	98%	100%	98%
Timsr	97%	100%	97%
Qz	96%	100%	96%
Vanp	96%	96%	96%
Tsfs	95%	100%	95%
Rp	100%	95%	95%
Spcg	94%	100%	94%
Qn	98%	95%	93%

Ncmsoy	93%	97%	92%
Pd3fsr	91%	91%	91%
Ncfsoy	91%	95%	90%
LPAR	90%	100%	90%
Ncmpoy	90%	100%	90%
Afpfsrn	96%	90%	89%
Ncfpoy	88%	100%	88%
Vmis3s	88%	88%	88%
Ncmp-n	92%	88%	88%
Pw---r	92%	90%	87%
Vmp--pm	93%	87%	87%
Ncfsrn	90%	90%	86%
Afpfp-n	91%	90%	86%
Vmip3	85%	85%	85%
Va--3p	85%	94%	85%
PERIOD	84%	100%	84%
Ncfp-n	89%	85%	84%
Vmsp3	84%	92%	84%
Ncfsry	87%	91%	83%
Np	86%	88%	83%
Etd	82%	91%	82%
Ncms-n	89%	87%	81%
Afpfson	85%	81%	81%
Vmip3p	80%	86%	80%
Vmip2p	82%	84%	79%
Vaip3s	79%	82%	79%
Vmip3s	81%	82%	78%
RPAR	78%	96%	78%
Ncmsry	83%	86%	77%
Ncmpry	77%	83%	77%
Afpms-n	88%	75%	74%
Vmnp	96%	76%	74%
Ncfpry	77%	90%	74%
Afpmp-n	81%	78%	74%
Spsa	78%	88%	72%
COMMA	73%	98%	72%
Csssp	78%	78%	71%
Vmp--sm	75%	73%	70%
Px3--a--y-----w	100%	70%	70%
Ed	90%	80%	70%
Qs	89%	68%	68%
Vmp--pf	84%	71%	68%
Afp	83%	75%	67%
Px3--a--------w	97%	62%	62%
Vmp--sf	68%	71%	61%
Ccssp	60%	91%	60%

Pp2-pa--------w	100%	58%	58%
Vmii2p	58%	65%	55%
Rw	65%	90%	55%
Spsg	55%	82%	55%
Eni	58%	61%	53%
Vmg	50%	70%	50%
Rc	75%	67%	50%
Rgp	52%	81%	47%
Crssp	48%	94%	47%
Spca	57%	75%	46%
Pp2----------s	69%	50%	46%
Yn	33%	43%	24%
Ncm--n	42%	17%	17%

4) Setul juridic 1

Acurateţe	**Acurateţe centru corect %**	**Acurateţe dependenţă corectă**	**Acurateţe ambele corecte**
total	**79%**	**83%**	**71%**
Qz	100%	100%	100%
Qn	100%	100%	100%
Vaip3p	100%	100%	100%
Vanp	100%	100%	100%
Di3--r---e	100%	100%	100%
Ncfson	100%	100%	100%
Dd3msr---e	100%	100%	100%
Va--3	100%	100%	100%
LPAR	99%	100%	99%
Timsr	98%	98%	98%
Pw---r	96%	99%	96%
Tifsr	96%	100%	96%
Va--3s	94%	100%	94%
Rp	93%	93%	93%
Tsfp	91%	100%	91%
Ncfsrn	95%	93%	90%
RPAR	90%	99%	90%
Vmsp3	89%	92%	89%
Tsfs	91%	94%	87%
Ncmsry	90%	88%	85%
Ncfpoy	86%	92%	83%
Vaip3s	83%	83%	83%
Ncmpoy	91%	91%	82%
Afpfsrn	96%	82%	81%
Vmis3s	81%	86%	81%
Afpfsry	80%	80%	80%

Ncfsry	84%	87%	79%
Ncfpry	84%	84%	78%
Ncfp-n	88%	84%	78%
Rgp	80%	90%	78%
Ncms-n	83%	85%	77%
Vmip3	78%	83%	77%
Ncmsoy	76%	88%	76%
Vmip3p	78%	78%	76%
Vmp--pf	84%	86%	75%
Ncfsoy	82%	89%	75%
Rw	81%	81%	75%
Afpfson	89%	75%	71%
Ncmpry	79%	93%	71%
Afpfp-n	88%	75%	70%
Np	80%	77%	70%
Spsa	74%	88%	68%
Afpmsry	71%	68%	68%
Tsms	95%	73%	68%
Vmip3s	69%	70%	66%
Afpms-n	86%	65%	64%
DASH	64%	93%	64%
Afpmp-n	91%	64%	64%
PERIOD	62%	91%	62%
Qs	73%	58%	58%
Spcg	61%	94%	58%
Afpmsoy	67%	58%	58%
Vmp--sm	61%	62%	55%
Vmnp	87%	60%	54%
DBLQ	54%	92%	54%
Ccssp	53%	91%	53%
COMMA	52%	99%	52%
Csssp	74%	70%	52%
Vmp--sf	64%	74%	51%
Crssp	53%	88%	50%
Px3--a--------w	98%	49%	49%
Spca	50%	89%	45%
Yn	55%	45%	45%
Vmg	42%	68%	42%
Rc	50%	58%	33%
Mc	60%	33%	28%
Ncm--n	32%	25%	21%

5) Setul Jurnalistic 2

Acurateţe	Acurateţe centru corect %	Acurateţe dependenţă corectă	Acurateţe ambele corecte
total	**81%**	**86%**	**75%**
Qn	100%	100%	100%
Timso	100%	100%	100%
Vag	100%	100%	100%
Va--3	100%	100%	100%
Ti-po	100%	100%	100%
Tsmp	100%	100%	100%
Ds3---p	100%	100%	100%
Di3fpr	100%	100%	100%
Di3--r---e	100%	100%	100%
Va--3p	99%	99%	99%
Tifsr	99%	100%	99%
Va--3s	99%	98%	98%
Timsr	96%	100%	96%
Vap--sm	100%	96%	96%
Dd3msr---e	95%	100%	95%
Vaip3s	95%	95%	95%
Dd3fsr---e	95%	95%	95%
Rp	100%	94%	94%
Tsfp	97%	97%	94%
Tsms	96%	96%	93%
LPAR	93%	100%	93%
Qz	92%	97%	92%
Enr	96%	92%	92%
Px3--d--------w	100%	92%	92%
Vanp	94%	91%	91%
Vaip3p	91%	91%	91%
Tsfs	94%	94%	90%
Px3--d--y-----w	90%	95%	90%
Momsrly	90%	90%	90%
PERIOD	87%	100%	87%
Ncmsoy	90%	93%	87%
Ncfpry	90%	90%	86%
Ncfpoy	89%	92%	86%
Etd	93%	87%	86%
Afpms-n	93%	84%	84%
Ncfsoy	90%	92%	84%
Pw3--r	87%	84%	84%
Ncmsry	88%	86%	83%
Afpfp-n	91%	87%	83%
Vmsp3	83%	91%	83%

Ncfsry	88%	87%	82%
Eni	84%	85%	82%
Ncfsrn	90%	85%	81%
Eqy	81%	81%	81%
Afp-p-n	81%	100%	81%
Np	85%	85%	79%
Vmp--pf	81%	87%	79%
Afpfson	97%	80%	79%
Afpmsry	79%	79%	79%
Afpfsry	78%	78%	78%
Ncfp-n	84%	84%	77%
Ncmpry	81%	88%	77%
RPAR	77%	97%	77%
Ncms-n	86%	79%	76%
Ncfson	75%	85%	75%
Rz	75%	83%	75%
Afpfsrn	93%	75%	73%
Ed	81%	76%	73%
Pp3msa--y-----w	100%	73%	73%
Pd3mpr	73%	73%	73%
Vmp--sm	73%	78%	72%
Ncmp-n	79%	75%	72%
Afpmp-n	73%	77%	72%
Vmip3s	74%	76%	71%
Px3--a--------w	99%	71%	70%
Pd3msr	80%	80%	70%
Vmp--pm	72%	79%	69%
Yn	69%	75%	69%
Spsa	76%	87%	68%
Ncmpoy	74%	74%	67%
DBLQ	64%	99%	64%
Csssp	82%	70%	64%
Px3--a--y-----w	95%	64%	64%
Eqt	71%	64%	64%
Pp3-sd--y-----w	73%	91%	64%
Vmnp	79%	65%	63%
Vmip3p	66%	70%	63%
Cscsp	71%	71%	63%
Npfsoy	71%	79%	63%
Tdfpr	63%	63%	63%
Spca	68%	88%	62%
Qs	73%	65%	62%
Rw	67%	81%	62%
Spsg	68%	89%	61%
COMMA	60%	98%	60%
Rgp	63%	85%	59%

Vmp--sf	64%	66%	59%
Vmg	62%	76%	59%
Spcg	60%	92%	58%
Vmis3s	62%	64%	57%
Vmii3s	61%	70%	57%
Vmg-------y	71%	64%	57%
Pd3-po	64%	73%	55%
DASH	54%	86%	54%
Vmip1p	57%	50%	50%
Vmil3s	50%	50%	50%
Crssp	50%	86%	49%
Rc	67%	60%	49%
Ccssp	52%	89%	48%
Spsay	43%	62%	38%
Tdmsr	36%	27%	27%
SLASH	57%	21%	7%

6) Setul Literar 2

Acurateţe	**Acurateţe centru corect %**	**Acurateţe dependenţă corectă**	**Acurateţe ambele corecte**
total	**84%**	**86%**	**77%**
Qn	100%	100%	100%
Va--1	100%	100%	100%
Qz-y	100%	100%	100%
Pp3fsa--y-----w	100%	100%	100%
Pp1-sa--------w	100%	100%	100%
Va--3s	100%	98%	98%
Tifsr	99%	99%	98%
Tsfs	98%	98%	98%
Timsr	97%	100%	97%
Ncmsoy	97%	97%	96%
Ds3---s	96%	96%	96%
Va--3p	96%	100%	96%
Qz	95%	100%	95%
Ncmpoy	95%	95%	95%
Rw	96%	96%	94%
Ncfpoy	93%	95%	93%
Tsfp	92%	100%	92%
Di3	92%	92%	92%
Vmsp3	93%	94%	91%
Px3--d--------w	100%	91%	91%
Vmil3p	90%	90%	90%
Ncfson	90%	90%	90%
Ncfsoy	90%	93%	89%
Afpfson	95%	89%	89%

Pp3msr--------s	94%	89%	89%
Ncfsry	94%	89%	88%
Pw3--r	92%	89%	88%
Afp-p-n	88%	100%	88%
PERIOD	86%	99%	86%
Afpfsrn	92%	87%	86%
Pp3msa--------w	100%	86%	86%
Ncmsry	91%	88%	85%
Vmip3p	85%	85%	85%
Vmii1	85%	85%	85%
Px3--a--------w	100%	84%	84%
Ncfpry	90%	86%	84%
Tsms	88%	96%	84%
Px3--d--y-----w	84%	84%	84%
Ncfsrn	90%	85%	83%
Rp	97%	83%	83%
Vmip3	83%	86%	83%
Ncms-n	87%	85%	82%
Ncfp-n	89%	85%	82%
Px3--a--y-----w	100%	82%	82%
Afpfp-n	89%	84%	81%
Vmii3p	85%	82%	81%
Vmp--pf	86%	81%	81%
Tf-so	82%	82%	79%
COMMA	78%	100%	78%
Np	86%	81%	77%
Vmip3s	80%	81%	77%
Vmis3s	80%	80%	76%
Vmil3s	82%	76%	76%
Spsay	79%	93%	76%
Vmip1s	76%	76%	76%
Ncmp-n	83%	81%	74%
Pp3msa--y-----w	94%	74%	74%
Vmp--sf	86%	73%	73%
Pd3fsr	100%	73%	73%
Vmp--sm	76%	77%	72%
DBLQ	72%	100%	72%
Cscsp	78%	94%	72%
Vmii3s	74%	79%	71%
Afpms-n	82%	74%	71%
Spca	71%	93%	71%
Rgp	75%	85%	70%
Vmip1p	70%	70%	70%
Vanp	95%	70%	70%
Pi3msr	90%	70%	70%
Spsa	78%	83%	68%

Vmnp	82%	68%	66%
Ncmpry	82%	71%	65%
Vmg	73%	77%	65%
Spcg	76%	82%	65%
Pp3fpa--------w	100%	64%	64%
Ed	82%	73%	64%
Pp3fsa--------w	97%	67%	63%
Afpmp-n	66%	66%	62%
Crssp	60%	90%	59%
Qs	81%	63%	59%
Eni	67%	58%	58%
Rc	66%	66%	57%
Pp1-pa--------w	100%	57%	57%
Ccssp	65%	83%	56%
Csssp	70%	67%	55%
Vmp--pm	80%	60%	55%
Afpmsry	65%	47%	47%
Pp3mpa--y-----w	100%	46%	46%
Vmg-------y	54%	85%	46%
DASH	44%	100%	44%
Pp3-sd--------w	96%	38%	38%
Pp3-pd--------w	91%	36%	36%
Pp3-sd--y-----w	100%	35%	35%

7) Setul Academic 2:

Acurateţe	**Acurateţe centru corect %**	**Acurateţe dependenţă corectă**	**Acurateţe ambele corecte**
total	**86%**	**89%**	**81%**
Tifsr	100%	100%	100%
Tf-so	100%	100%	100%
Qn	100%	100%	100%
Tifso	100%	100%	100%
Tsfp	100%	100%	100%
Timso	100%	100%	100%
Vaip3p	100%	100%	100%
Di3fpr	100%	100%	100%
Px3--d--------w	100%	100%	100%
Mc-p-l	100%	100%	100%
Px3--d--y-----w	100%	100%	100%
Ti-po	100%	100%	100%
Mcfp-l	100%	100%	100%
Vap--sm	100%	100%	100%
Afpmsoy	100%	100%	100%
Ds3ms-s	100%	100%	100%

Pd3msr	100%	100%	100%
Tsfs	100%	99%	99%
Va--3s	98%	99%	98%
Timsr	99%	99%	98%
Vmsp3	98%	98%	98%
Qz	97%	98%	97%
Ds3---s	97%	100%	97%
Pw3--r	98%	98%	96%
Rp	99%	97%	96%
Pp3msr--------s	100%	96%	96%
LPAR	95%	100%	95%
Va--3p	95%	95%	95%
Tsms	97%	95%	94%
Vanp	94%	94%	94%
Afpmsry	94%	94%	94%
Dd3msr---e	93%	100%	93%
Afp-p-n	92%	96%	92%
DBLQ	92%	98%	91%
Vaip3s	91%	91%	91%
Tdfpr	91%	91%	91%
Ncmpry	90%	90%	90%
Ncfsoy	94%	92%	89%
Pp3msa--------w	100%	89%	89%
Ncfsrn	93%	90%	88%
Afpf--n	87%	100%	87%
PERIOD	86%	100%	86%
Np	90%	89%	86%
Afpfsrn	92%	87%	86%
Ncmsoy	87%	93%	86%
Afpfson	93%	88%	86%
Ncmsry	90%	89%	85%
Ncfp-n	90%	90%	85%
Ncfpry	90%	89%	85%
RPAR	85%	99%	85%
Afpmp-n	88%	90%	85%
Rw	93%	88%	85%
Ncfson	92%	85%	85%
Afpfp-n	88%	87%	84%
Ncfsry	88%	87%	83%
Ncmp-n	86%	87%	83%
Ed	86%	87%	82%
Etd	90%	86%	82%
Ncfpoy	85%	91%	82%

Ncms-n	90%	84%	81%
Vmg-------y	81%	88%	81%
Npfsoy	88%	94%	81%
Afpms-n	90%	82%	80%
Pp3msa--y-----w	100%	80%	80%
COMMA	79%	100%	79%
Px3--a--y-----w	97%	78%	78%
Eni	78%	83%	78%
Vmip3s	78%	80%	76%
Vmip3p	77%	80%	76%
Rgp	77%	86%	75%
Vmip3	77%	80%	75%
Vmp--pf	80%	82%	75%
Vmis3s	74%	79%	73%
Afpfsry	80%	73%	73%
Spsa	80%	85%	72%
Vmnp	86%	74%	72%
Spsay	75%	97%	72%
Px3--a--------w	100%	71%	71%
Spsg	75%	85%	70%
Spca	71%	91%	69%
Qs	87%	69%	69%
Ncmpoy	81%	78%	69%
Crssp	69%	93%	68%
Vmp--sm	74%	71%	68%
Vmp--sf	73%	71%	68%
Pp3-sd--------w	100%	68%	68%
Vmg	72%	82%	67%
DASH	67%	94%	67%
Spcg	70%	91%	65%
Rc	68%	79%	63%
Cscsp	60%	80%	60%
Csssp	70%	74%	59%
Vmii3s	64%	61%	58%
SLASH	65%	74%	58%
Ccssp	52%	86%	44%

8) Setul Medical 2:

Acurateţe	**Acurateţe centru corect %**	**Acurateţe dependenţă corectă**	**Acurateţe ambele corecte**
total	**85%**	**90%**	**82%**
Qn	100%	100%	100%

Timsr	100%	100%	100%
Rp	100%	100%	100%
Vap--sm	100%	100%	100%
Vanp	100%	100%	100%
Tsfp	100%	100%	100%
Va--2p	100%	100%	100%
Dd3msr---e	100%	100%	100%
Vaip3p	100%	100%	100%
Di3fpr---e	100%	100%	100%
Dd3fpr---e	100%	100%	100%
Tifso	100%	100%	100%
Va--3	100%	100%	100%
Timso	100%	100%	100%
Di3-sr---e	100%	100%	100%
Tifsr	100%	99%	99%
Vaip3s	100%	98%	98%
Va--3s	98%	99%	97%
Qz	98%	99%	97%
Va--3p	97%	99%	97%
Tsfs	97%	97%	97%
Pp2----------s	96%	96%	96%
Afpfsrn	97%	96%	95%
Ncfsoy	95%	99%	95%
LPAR	94%	100%	94%
Ncfpoy	96%	97%	94%
Vmsp3	94%	97%	94%
Ncmpry	94%	96%	94%
Pd3fsr	100%	94%	94%
Afpfson	96%	93%	93%
Afpfp-n	93%	95%	92%
Ncfpry	95%	93%	92%
Afp-p-n	92%	100%	92%
Pw3--r	95%	93%	91%
Tsms	91%	100%	91%
Afp	91%	91%	91%
Pp2-pa--y-----w	100%	90%	90%
DBLQ	89%	100%	89%
Ncmsry	91%	89%	88%
RPAR	88%	98%	88%
PERIOD	87%	100%	87%
Ncfsrn	90%	90%	86%
Vmip2p	86%	90%	86%
Ncmsoy	87%	93%	86%
Ncmp-n	87%	90%	85%

Vmm-2p	85%	92%	85%
Ncfsry	88%	86%	83%
Vmnp	91%	84%	83%
Px3--a--y-----w	100%	83%	83%
Vmii2p	83%	93%	83%
Pp2-pa--------w	100%	83%	83%
Ncfp-n	88%	87%	82%
Np	84%	88%	82%
Afpms-n	87%	84%	82%
Ncms-n	84%	87%	81%
Qs	90%	81%	81%
Spcg	81%	95%	81%
Vmp--pm	80%	87%	80%
COMMA	81%	97%	79%
Vmp--sm	81%	84%	79%
Px3--a--------w	98%	79%	79%
Vmp--sf	82%	86%	79%
Rw	88%	83%	79%
Vmp--pf	80%	92%	78%
Afpmp-n	86%	84%	78%
Vmip3s	78%	82%	77%
Vmip3p	79%	77%	77%
Enr	77%	77%	77%
Spsa	78%	91%	75%
Vmip3	75%	85%	75%
Mc-p-l	80%	80%	75%
Rgp	77%	86%	74%
Vmg	75%	82%	73%
Spsay	73%	96%	73%
Pi3fpr	100%	73%	73%
Eni	71%	76%	71%
Vmis3s	77%	80%	71%
Pd3fpr	93%	71%	71%
Csssp	81%	81%	70%
Ed	75%	72%	68%
Pp2-pd--------w	94%	67%	67%
Ncfson	74%	65%	65%
Ncmpoy	63%	88%	63%
Cscsp	71%	71%	62%
Ccssp	62%	98%	61%
Spsg	61%	100%	61%
Vmip2s	93%	60%	60%
Yn	68%	68%	59%
Pp3msa--y-----w	92%	58%	58%

Rc	69%	71%	56%
Crssp	54%	94%	53%
Ncm--n	67%	53%	53%
Spca	59%	82%	50%
DASH	50%	64%	43%
SLASH	61%	67%	39%

9) Juridic 2

Acurateţe	**Acurateţe centru corect %**	**Acurateţe dependenţă corectă**	**Acurateţe ambele corecte**
total	**90%**	**93%**	**87%**
RPAR	100%	100%	100%
Pw3--r	100%	100%	100%
Tsfs	100%	100%	100%
Vaip3p	100%	100%	100%
Tifsr	100%	100%	100%
Afpfsry	100%	100%	100%
Afpmsry	100%	100%	100%
Qn	100%	100%	100%
Qz	100%	100%	100%
Tsfp	100%	100%	100%
Ds3---p	100%	100%	100%
Va--3p	100%	100%	100%
Vanp	100%	100%	100%
Afpfson	100%	100%	100%
Px3--d--------w	100%	100%	100%
Timso	100%	100%	100%
Dd3fpr---e	100%	100%	100%
Dd3msr---e	100%	100%	100%
Ncfson	100%	100%	100%
Ncmpry	100%	100%	100%
Timsr	100%	100%	100%
COLON	100%	100%	100%
Di3-sr---e	100%	100%	100%
Ds3fp-s	100%	100%	100%
Rw	100%	100%	100%
Spsay	100%	100%	100%
Va--3s	100%	100%	100%
Vaip3s	100%	100%	100%
Afp	100%	100%	100%
Afpfpry	100%	100%	100%
Afpmp-n	100%	100%	100%
Dd3fso	100%	100%	100%
Dd3fsr	100%	100%	100%
Di3fpr	100%	100%	100%

Di3ms----e	100%	100%	100%
Di3--r---e	100%	100%	100%
Dz3msr---e	100%	100%	100%
Mofsoly	100%	100%	100%
Mo-s-r	100%	100%	100%
Pd3fsr	100%	100%	100%
Pd3msr	100%	100%	100%
Px3--d--y-----w	100%	100%	100%
Tifso	100%	100%	100%
LPAR	98%	100%	98%
Mc	97%	99%	97%
Ncfpry	96%	96%	96%
Ncfsry	94%	97%	94%
Afpms-n	97%	94%	94%
Vmnp	94%	94%	94%
Ncfsrn	98%	94%	93%
DBLQ	92%	100%	92%
PERIOD	91%	100%	91%
Afpfsrn	94%	91%	91%
Vmp--pf	91%	97%	91%
Vmip3s	91%	94%	91%
Vmsp3	90%	100%	90%
Ncms-n	92%	94%	89%
Ccssp	88%	100%	88%
Tsms	100%	88%	88%
Yn	88%	100%	88%
Ncmsry	90%	90%	86%
Ncfp-n	93%	86%	86%
Vmip3	86%	90%	86%
Np	90%	90%	84%
Ncfpoy	86%	95%	84%
Ncmsoy	86%	94%	83%
Vmp--sm	82%	82%	82%
Spsa	83%	94%	81%
Vmip3p	85%	88%	81%
Afpfp-n	89%	91%	80%
Spcg	80%	100%	80%
Ncfsoy	91%	82%	79%
Csssp	100%	78%	78%
COMMA	77%	98%	77%
Vmp--sf	85%	92%	77%
Ncm--n	75%	100%	75%
Px3--a--------w	100%	74%	74%
Vmis3s	75%	71%	71%
Afpmsoy	71%	71%	71%
Crssp	69%	100%	69%

Vmg	69%	85%	69%
Qs	83%	67%	67%
Cscsp	67%	100%	67%
Px3--a--y-----w	100%	67%	67%
Rc	100%	67%	67%
Rgp	66%	86%	66%
Spca	58%	85%	58%
Pd3fso	50%	75%	50%
Ncmpoy	50%	50%	50%
Pd3fpr	50%	50%	50%
Pi3fso	33%	33%	33%
Pi3fsr	100%	33%	33%

Distribuţia preciziei şi recall-ului pentru etichetă corectă și centru corect pe tipuri de relaţii de dependenţă (în ordine descrescătoare a recall-ului).

1) Setul Literar 1.

Relaţia de dependenţă	**Recall (%) (corecte sistem/corecte)**	**Precizie (%) (corecte sistem/sistem)**
prep	92.3	88.73
det	90.28	84.47
aux	89.27	93.37
auxpass	81.25	60.94
mark	77.33	65.52
name	70.83	54.84
ROOT	70.11	55.45
neg	68.46	78.07
amod	68.02	68.15
nmod	67.74	65.37
passmark	65.96	39.74
pmod	65.17	54.91
sc	61.04	82.1
agc	60	52.94
poss	59.09	60.47
punct	53.57	50.61
dobj	52.81	54.43
dblclitic	52.54	51.67
subj	50.11	44.68
advmod	49.34	45.61
reflclitic	42.57	72.41
cc	41.24	43.96
acl	37.58	45.74
narrat	37.5	33.33
conj	37.03	29.03

iobj	22.12	51.02
pred	21.05	42.55
post	18.75	37.5
mwe-dobj	14.29	50
advcl	12.73	21.88
spe	11.11	8.33
pobj	4.23	30
appos	0	0
correl	0	NaN[18]
foreign	0	NaN
mwe-acl	0	NaN
mwe-advcl	0	NaN
mwe-advmod	0	NaN
mwe-amod	0	0
mwe-cc	0	NaN
mwe-conj	0	NaN
mwe-foreign	0	0
mwe-mark	0	NaN
mwe-nmod	0	0
mwe-pmod	0	0
mwe-poss	0	NaN
mwe-prep	0	0
mwe-sc	0	NaN
mwe-subj	0	NaN
parataxis	0	0
remnant	0	NaN
secobj	0	0
voc	0	NaN
xcomp	0	NaN
list	NaN	0
mwe-pred	NaN	0

2) Setul Academic 1

Relaţia de dependenţă	**Recall (%) (corecte sistem/corecte)**	**Precizie (%) (corecte sistem/sistem)**
aux	100	97.92
list	100	50
mark	98.15	88.33
prep	97.33	92.05
neg	96.3	100
det	92.46	94.72

[18] NaN = Not a Number; provine din operația 0/0;

name	90.1	85.85
nmod	89.43	74.56
poss	86.36	90.48
amod	85.93	82.08
ROOT	80.29	80.68
auxpass	80	89.66
reflclitic	79.1	74.65
dobj	76.88	76.88
pmod	75.54	66.8
sc	75.36	91.23
subj	73.9	70.1
dblclitic	70.59	52.17
punct	69.36	69.9
advmod	64.47	69.34
mwe-mark	60	100
cc	58.18	50.26
agc	55.88	63.33
pred	51.06	46.15
admod	50	100
mwe-acl	50	100
mwe-foreign	50	100
acl	49.28	56.67
advcl	48.89	44.44
passmark	47.22	54.84
conj	45.27	44
iobj	42.86	60
parataxis	21.43	30
secobj	20	50
spe	20	20
mwe-prep	19.05	66.67
remnant	18.18	100
post	16.67	50
mwe-dobj	14.29	100
appos	12.77	20
pobj	9.26	41.67
goeswith	9.09	100
xcomp	9.09	33.33
mwe-advcl	0	NaN
mwe-amod	0	NaN
mwe-aux	0	NaN
mwe-cc	0	NaN
mwe-conj	0	NaN
mwe-name	0	NaN
mwe-nmod	0	NaN
mwe-punct	0	NaN
mwe-sc	0	NaN

nwe-amod	0	NaN
mwe-advmod	0	0
mwe-det	0	0
mwe-pmod	0	0

3) Setul Medical 1

Relaţia de dependenţă	**Recall (%) (corecte sistem/corecte)**	**Precizie (%) (corecte sistem/sistem)**
post	100	50
prep	97.45	92.59
neg	95.52	94.12
aux	95	88.79
det	94.77	92.95
amod	91.3	83.39
auxpass	88.79	93.14
mark	87.14	98.39
ROOT	82.92	82.92
dobj	82.19	74.5
nmod	80.73	80.73
name	80	72.73
punct	79.44	79.18
pmod	78.13	71.84
agc	76.47	61.9
sc	72.51	91.18
reflclitic	71.43	50.85
subj	70.53	75.62
mwe-dobj	66.67	33.33
poss	62.5	55.56
acl	61.59	60.39
pred	59.38	62.3
iobj	57.89	42.31
passmark	57.81	82.22
cc	55.33	52.87
advmod	51.2	45.7
dblclitic	50	50
conj	48.52	45.81
mwe-prep	36.84	77.78
secobj	33.33	50
advcl	33.04	54.41
mwe-amod	25	100
appos	23.08	60
pobj	20.83	50
xcomp	20	50
goeswith	5	50

parataxis	2.5	12.5
correl	0	NaN
mwe-nmod	0	NaN
mwe-pmod	0	NaN
mwe-post	0	NaN
remnant	0	NaN
spe	0	0

4) Setul Juridic 1

Relaţia de dependenţă	**Recall (%) (corecte sistem/corecte)**	**Precizie (%) (corecte sistem/sistem)**
neg	100	100
aux	100	93.18
mwe-mark	100	50
prep	97.69	91.97
mark	94.23	92.45
det	92.61	86.7
auxpass	92	98.57
pred	87.8	65.45
dblclitic	83.33	71.43
dobj	80.51	67.78
poss	80	66.67
nmod	79.94	65.76
ROOT	77.86	77.86
advmod	76.53	58.59
amod	75.86	76.38
punct	70.87	71.51
pmod	70.27	63.19
subj	70.03	69.44
reflclitic	68.57	34.29
sc	68.29	90.32
name	54.55	54.55
acl	53.75	70.83
cc	53.41	50.27
iobj	51.16	55
agc	46.15	75
conj	45.75	45.33
passmark	42.47	75.61
advcl	39.62	20.79
pobj	36.36	69.57
mwe-prep	34.78	100
post	30	60
appos	18.18	19.05
mwe-pmod	15.38	100

mwe-dobj	12.5	100
mwe-nmod	11.76	50
parataxis	7.59	63.16
mwe-amod	6.45	66.67
foreign	0	NaN
mwe-advmod	0	NaN
mwe-cc	0	NaN
mwe-conj	0	NaN
mwe-det	0	NaN
mwe-foreign	0	NaN
mwe-sc	0	NaN
partaxis	0	NaN
voc	0	NaN
goeswith	0	0
remnant	0	0
secobj	0	0
spe	0	0
xcomp	0	0

5) Setul Jurnalistic 2

Relaţia de dependenţă	**Recall (%) (corecte sistem/corecte)**	**Precizie (%) (corecte sistem/sistem)**
aux	98.7	98.06
prep	96.92	91.88
auxpass	96.43	93.75
neg	94.83	91.67
det	91.92	91.22
amod	91.34	83.79
name	87.12	84.94
ROOT	87.01	87.01
mark	84.5	87.9
nmod	83.54	75.76
post	80.22	84.88
dobj	79.01	69.73
poss	78.95	90.91
dblclitic	78.13	75.76
passmark	77.78	64.95
pmod	73.01	65.51
sc	72.92	87.94
punct	68.77	68.37
subj	68.34	73.72
reflclitic	59.09	72.22
advmod	59.09	59.51
agc	57.5	76.67

cc	52.38	49.3
pred	46.84	57.81
advcl	45.29	47.83
acl	44.29	64.37
conj	44.11	36.59
iobj	36.78	60.38
appos	34.82	52
mwe-dobj	33.33	16.67
pobj	29.55	44.83
goeswith	27.78	31.25
secobj	25	20
mwe-amod	20.99	89.47
list	16	100
spe	14.29	50
mwe-prep	14.06	100
mwe-nmod	8.51	100
mwe-pmod	7.41	57.14
parataxis	7.41	28.57
correl	0	NaN
foreign	0	NaN
mwe-advmod	0	NaN
mwe-aux	0	NaN
mwe-cc	0	NaN
mwe-conj	0	NaN
mwe-foreign	0	NaN
mwe-pred	0	NaN
mwe_neg	0	NaN
nwe-amod	0	NaN
remnant	0	NaN
xcomp	0	NaN
mwe-det	0	0
mwe-mark	0	0
mwe-name	0	0
mwe-punct	0	0

6) Setul Literar 2

Relaţia de dependenţă	**Recall (%) (corecte sistem/corecte)**	**Precizie (%) (corecte sistem/sistem)**
mwe-mark	100	60
narrat	100	50
prep	96.8	94.7
aux	96.24	98.35
det	95.82	94.3
neg	95.7	95.7

name	94.29	71.74
mark	93.23	82.67
auxpass	92.68	69.09
amod	89.28	87.98
sc	86.85	92.37
nmod	86.56	81.47
ROOT	84.88	84.88
reflclitic	83.67	90.44
poss	79.49	73.81
punct	79.44	78.8
passmark	78.38	64.44
dobj	75.57	72.05
pmod	74.6	67.92
subj	71.27	75.23
agc	68.97	54.05
advmod	68.36	72.65
dblclitic	64.44	43.28
cc	61.94	59.27
conj	60.63	51.59
acl	60.54	65.12
pred	53.41	74.6
post	50	58.33
iobj	49	60.49
advcl	44.38	53.74
mwe-prep	42.86	100
secobj	36.36	100
parataxis	26.09	33.33
appos	25.58	42.31
pobj	23.94	80.95
mwe-amod	20	50
voc	20	50
xcomp	16.67	21.43
mwe-dobj	16.67	16.67
foreign	0	NaN
list	0	NaN
mwe-advmod	0	NaN
mwe-det	0	NaN
mwe-name	0	NaN
mwe-pmod	0	NaN
mwe-sc	0	NaN
partaxis	0	NaN
remnant	0	NaN
spe	0	NaN
mwe-nmod	0	0

7) Setul Academic 2

Relaţia de dependenţă	Recall (%) (corecte sistem/corecte)	Precizie (%) (corecte sistem/sistem)
aux	99.33	96.73
det	99.04	98.09
neg	98.31	96.67
prep	97.96	96.77
auxpass	95.83	91.27
amod	94.66	88.14
poss	94.12	96.97
post	94.12	88.89
mark	93.7	95.97
name	90.84	92.31
nmod	90.41	86.39
ROOT	86.09	86.09
goeswith	85.71	66.67
sc	84.38	93.1
punct	82.59	82.44
dobj	82.4	74.48
subj	80.53	82.11
pmod	79.94	71.96
reflclitic	75.59	80
advmod	74.81	76.34
agc	71.13	74.19
dblclitic	70.59	66.67
cc	66.5	63.7
advcl	63.8	53.89
passmark	62.32	61.43
pred	60.19	70.65
conj	59.9	56.9
iobj	52.78	62.3
acl	51.3	67.3
appos	29.46	36.89
foreign	28.57	100
mwe-amod	27.27	90
mwe-prep	26.83	84.62
mwe-cc	25	100
parataxis	21.21	36.84
mwe-conj	20	100
xcomp	17.86	23.81
mwe-pmod	13.64	75
pobj	13.58	64.71
mwe-nmod	11.11	66.67
spe	5	33.33
correl	0	NaN

list	0	0
mwe-advmod	0	NaN
mwe-det	0	NaN
mwe-dobj	0	0
mwe-name	0	NaN
mwe-punct	0	NaN
remnant	0	NaN
secobj	0	NaN

8) Setul Medical 2

Relaţia de dependenţă	**Recall (%) (corecte sistem/corecte)**	**Precizie (%) (corecte sistem/sistem)**
det	98.84	97.71
neg	97.75	96.67
aux	97.65	97.27
prep	97.06	94.29
auxpass	93.72	95.72
amod	92.46	89.25
sc	91.97	96.73
mark	91.62	99.35
dobj	91.18	80
reflclitic	88.46	71.13
dblclitic	87.5	63.64
ROOT	86.07	85.89
nmod	83.68	80.1
punct	83.28	83.28
agc	81.08	61.22
subj	81.06	84.45
pred	80.87	80.87
name	80.49	73.33
pmod	77.9	73.19
advmod	75.51	72.75
passmark	74.53	90.8
post	71.43	95.24
iobj	70.89	77.78
acl	69.38	73.75
secobj	66.67	57.14
mwe-amod	65	86.67
advcl	63.1	66.25
cc	56.68	57.25
conj	51.43	54.21
goeswith	50	46.94
xcomp	50	10
mwe-dobj	42.86	75

poss	42.86	60
mwe-prep	40.74	100
mwe-pmod	40	100
mwe-nmod	33.33	50
appos	29.09	46.38
pobj	26.42	70
parataxis	17.8	43.75
spe	5.26	50
compound	0	NaN
correl	0	NaN
foreign	0	NaN
mwe-advmod	0	NaN
mwe-cc	0	NaN
mwe-det	0	NaN
remnant	0	0

9) Setul Juridic 2

Relaţia de dependenţă	**Recall (%) (corecte sistem/corecte)**	**Precizie (%) (corecte sistem/sistem)**
agc	100	60
aux	100	90
det	100	98.44
mwe-amod	100	100
mwe-dobj	100	75
mwe-pmod	100	100
mwe-prep	100	100
neg	100	100
poss	100	100
prep	100	98.9
reflclitic	100	38.89
auxpass	95.83	100
amod	93.86	92.24
sc	92.59	89.29
nmod	92.03	80.89
subj	90.52	84
punct	90.17	91.41
ROOT	90	90
dobj	89.89	88.89
mark	84.62	100
name	83.33	83.33
pred	83.33	90.91
parataxis	82.76	90.57
pmod	81.38	77.01

cc	73.61	73.61
advmod	73.33	68.75
conj	70.93	73.49
passmark	68.57	100
advcl	66.67	42.86
acl	65.12	86.15
iobj	56.52	92.86
pobj	50	80
post	50	100
spe	50	100
dblclitic	40	100
mwe-nmod	33.33	100
appos	22.22	28.57
goeswith	0	NaN
list	0	NaN
mwe-det	0	NaN
remnant	0	NaN
xcomp	NaN	0