ACCELERAREA DEZVOLTĂRII UNUI CORPUS DIGITAL ADNOTAT CU RELAȚII DE DEPENDENȚĂ PENTRU LIMBA ROMÂNĂ UTILIZÂND RESURSE ŞI INSTRUMENTE CONSTRUITE PENTRU ALTE LIMBI

Elena Irimia

elena@racai.ro

Institutul de Cercetări pentru Inteligență Artificială „Mihai Drăgănescu”, Academia Română, Bucureşti

Rezumat: Un corpus adnotat sintactic este o resursă fundamentală pentru supraviețuirea unei limbi în spațiul digital. Am construit un corpus de dimensiuni modeste (5000 de propoziții) într-un timp scurt (12 luni) şi cu resurse umane reduse, acesta urmând să funcționeze ca bază în dezvoltarea de resurse şi instrumente care să asigure suport pentru analiza sintactică a limbii române, în cadrul grupului de cercetare în Prelucrarea Limbajului Natural de la ICIA. De aceea, propozițiile selectate pentru adnotare aparțin mai multor stiluri funcționale şi domenii, au lungimi variate şi complexitate sintactică ridicată şi conțin verbe cu utilizare frecventă în limbă. Prin selecția atentă, am urmărit să asigurăm corpusului rezultat diversitate stilistică şi sintactică şi reprezentativitate lingvistică.

Cuvinte cheie: corpus, gramatică de dependențe, adnotare sintactică automată, model statistic.

Abstract: Syntactically annotated corpora are fundamental for any language's survival in the digital universe. We developed a corpus of small size (5000 sentences) in a short a period of time (12 months) and with limited work force; but it is meant to function as a base for developing more resources and instruments to support syntactic analysis for Romanian in the NLP group at ICIA. The sentences selected for annotation are representing different genres and domains, have different lengths (between 10 and 40 words), have high syntactical complexity and contain verbs that are frequently used in Romanian. By careful selection, we intended to assure the stylistic and syntactic diversity and the linguistic representativeness of the resulting corpus.

Key words: corpus, dependency grammar, parsing, treebank.

1. Introducere

Proiectul descris în cadrul acestui articol este doar un pas dintr-o strategie amplă de integrare a limbii române în spațiul digital european. Pentru ca vorbitorii săi nativi să se poată bucura neîngrădit de avantajele progresului tehnologic în viața publică şi privată la standardele la care au acces alți cetățeni europeni, limba română are nevoie de resurse şi instrumente electronice dedicate. Acest suport tehnologic îi poate asigura integrabilitatea în complexele aplicații inteligente, mobile şi web, care au devenit indispensabile.

Comisia Europeană are ca prioritate dezvoltarea unei Piețe Digitale Unice (Digital Single Market), dar, în acelaşi 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 şi unde s-a angajat la producerea şi 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 şi instrumentele lingvistice dezvoltate sunt limitate atât cantitativ cât şi calitativ (vedeți studiul "Limba română în era digitală" [1], elaborat în cadrul proiectului METANET[1]). Totuşi, anterior acestui studiu şi de atunci încoace, multe eforturi individuale, instituționale sau prin colaborarea mai multor instituții au avut loc în direcția micşorării acestor diferențe tehnologice.

La ICIA, cercetările sunt concentrate în mai multe direcții, dintre care cele mai importante:

1. dezvoltarea unei ontologii lexicale monolingve, RoWordnet [2,3], aliniată printr-un index interlingvistic la Princeton Wordnet şi, prin acesta, la o rețea globală de wordneturi, cunoscută sub numele de Global Wordnet[2]; RoWordnet este esențial în dezvoltarea a

[1] http://www.meta-net.eu/whitepapers/overview
[2] http://globalwordnet.org/

numeroase aplicații monolingve şi multilingve, precum dezambiguizarea semantică, sistemele de traducere automată, sistemele întrebare-răspuns, etc.

2. colectarea de resurse, dicționare, lexicoane, corpusuri: cel mai recent şi ambițios proiect, în care suntem angajați, împreună cu Institutul de Informatică Teoretică – Iaşi, într-un program prioritar al Academiei Române, este realizarea unui corpus computațional de referință pentru limba română contemporană, denumit CoRoLa [4]; acesta va fi o colecție de texte în format digital (scrise şi orale) de dimensiune mare (cinci sute de milioane de cuvinte); adnotate cu metainformații – precum autor, data publicării etc. – şi cu date lingvistice – precum părți de vorbire, forma din dicționar a cuvântului adnotat, dependențe sintactice etc. – documentele vor fi disponibile liber online, spre consultare şi valorificare în scopuri de cercetare.
3. traducerea automată: 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 şi tehnologii prin care corpusuri comparabile de mari dimensiuni să fie exploatate pentru creşterea performanțelor aplicațiilor de traducere automată prin metode statistice[5]; alte direcții de cercetare abordate au fost dezvoltarea unui sistem de traducere automată bazat pe exemple [6], dezvoltarea unui sistem de traducere automată pentru limbaj vorbit [7], dezvoltarea de corpusuri paralele care servesc drept resurse de antrenare pentru traducătoare statistice, oferirea online[3], spre utilizare în scopuri de cercetare, a unui sistem de traducere statistic fiabil si performant, pentru perechi de limbi precum engleză-română, germană-română, spaniolă-română.

2. Analiza sintactică automată

Nivelul de analiză sintactic este, alături de cel morfologic, cel semantic şi cel pragmatic, una dintre problemele pe care se concentrează eforturile în domeniul Prelucrarea Limbajului Natural. În context internațional, pentru multe aplicații din PLN, integrarea informației sintactice a condus la creşterea performanței față de algoritmii bazați doar pe informație morfologică sau față de cei ne-supervizați[4]. 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 [8, 9, 10]. În România însă facem abia primii paşi către valorificarea informației sintactice în aplicații de Traducere Automată: un studiu din 2012 [11] descria o metodă de extragere a unor şabloane de traducere din texte paralele Română-Engleză, adnotate cu constituenți sintactici, dar nu mergea mai departe la utilizarea şabloanelor pentru îmbunătățirea calității traducerii.

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) şi dezvoltarea de analizoare sintactice (eng. *parser*).

Corpusul Adnotat Sintactic Lancaster (eng. Lancaster Parsed Corpus) [12] şi corpusul Penn TreeBank [13] sunt resurse dedicate limbii engleze, construite în anii ’90, care au deschis drumul pentru alte corpusuri importante precum NEGRA [14], TIGER [15], TüBa[5], Prague Dependency Treebank [16]. În prezent, cele mai multe limbi europene precum şi multe alte limbi importante dispun de treebank-uri. În România, eforturile au fost limitate până în acest moment: cele câteva treebank-uri dezvoltate nu au mai mult de 5000 de propoziții adnotate, resursele sunt disponibile doar pentru căutare online sau nu sunt deloc disponibile, instrumentele folosite pentru adnotare nu sunt accesibile [17, 18, 19, 20]. Unii dintre autori [17] limitează analiza la nivel de propoziție, excluzând situațiile de subordonare prin segmentarea frazelor. Considerăm că resursa obținută este departe de a acoperi varietatea relațiilor sintactice din limba română.

[3] http://www.racai.ro/tools/translation/racai-translation-system/
[4] Învățare ne-supervizată: metodă care încearcă extragerea de şabloane ți trăsături din date complet ne-adnotate.
[5] http://www.sfs.uni-tuebingen.de/en/ascl/resources/corpora.html),

Cea mai mare provocare pentru analizoarele sintactice o constituie construcţiile ambigue, în special grupurile prepoziţionale (GP). De exemplu, în propoziţia "Citesc articolul cu interes pentru subiect" şi "Citesc articolul cu informaţii despre rezultate", ataşamentul GP-ului ce urmează după grupul nominal (GN) obiect direct "articolul" se face la verb în prima propoziţie, respectiv la GN obiect direct în a doua propoziţie. În limbi cu topică (relativ) liberă şi omonimie morfologică (precum limba română) este dificil pentru analizor să distingă între subiect şi obiectul direct al verbului, deoarece eticheta morfo-sintactică atribuită la pre-procesarea textului este aceeaşi când argumentele sunt exprimate prin substantive (cazurile sincretice nu pot fi întotdeauna diferenţiate corect).

Cele mai recente abordări ale analizei limbajului pleacă de la premisa că structura unei propoziţii depinde de semantica lexicală a verbului şi a altor predicate din propoziţie. Experimentele arată că analiza sintactică oferă informaţii utile pentru recunoaşterea entităţilor denumite, extragerea de relaţii şi etichetarea rolurilor semantice [21]. Abordarea opusă, când informaţia semantică (i.e., trăsăturile extrase de un adnotator de entităţi denumite) poate ajuta analiza sintactică, a condus de asemenea la rezultate bune [22]. Pe de altă parte, parser-ele nesupervizate, bazate doar pe grupare (en. *clustering*) de cuvinte, par să conducă, pentru limba chineză, la rezultate mai bune decât metodele bazate pe informaţie morfo-sintactică [23].

În acest moment, cercetătorii manifestă un interes susţinut pentru analiza sintactică cu dependenţe şi pe tehnicile bazate pe date, care urmăresc modelarea statistică a proprietăţilor sintactice şi lexicale ale cuvintelor, aşa cum sunt ele reflectate în corpusuri de mari dimensiuni. Primul parser statistic performant a fost dezvoltat de către Michael Collins [24, 25] şi a devenit foarte influent în NLP. I-au urmat mai multe parsere statistice dezvoltate la Stanford (un parser cu un model statistic factorizat [26], un parser pentru gramatici de dependenţe bazat pe reţele neuronale [27], etc.) precum şi MaltParser [28], un generator de parsere bazat pe date, care poate construi un parser pornind de la un treebank.

Câteva încercări de creare a unor analizoare sintactice pentru limba română au avut loc de asemenea: Călăcean şi Nivre [29] au antrenat MaltParser pe treebank-ul dezvoltat de Hristea şi Popescu [17] iar Seretan et al. [30] au adaptat analizorul bazat pe reguli Fips (http://www.latl.unige.ch/) pentru limba română. Cele două parsere nu sunt disponibile pentru descărcare şi integrare în alte aplicaţii, ci doar pentru utilizare online.

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. În continuare vom prezenta strategia de dezvoltare a acestui treebank, motivând deciziile de selecţie a resurselor şi instrumentelor folosite (Secţiunea 3), descriind gramatica de dependenţe utilizată în adnotare (Secţiunea 4) şi detaliind procesul de lucru (Secţiunea 5).

3. Resurse şi instrumente utilizate

Deoarece dispunem de resurse umane şi de timp limitate, ne-am propus să dezvoltăm o resursă ale cărei dimensiuni modeste să fie compensate de calitatea acesteia. Treebankul de 5000 de propoziţii obţinut va trebui să reprezinte un bun set de antrenare pentru un analizor sintactic statistic, facilitând 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[6] dezvoltat la ICIA [31], care cuprinde cinci secţiuni corespunzătoare la cinci domenii distincte: jurnalistic (ştiri şi editoriale), medical (scurte texte farmaceutice), juridic (texte extrase din Acquis-ul Comunitar), academic (biografii şi recenzii critice ale unor autori literari), ficţiune (romane atât româneşti cât şi traduse).

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

Corpusul este segmentat la nivel de cuvânt, adnotat morfo-sintactic, lematizat, analizat sintactic la nivel de grup sintactic şi codificat în format XCES.

Pe baza informației morfo-lexicale din ROMBAC, am putut identifica verbele predicative şi calcula frecvențele acestora în corpus. Ne-am concentrat pe cele mai frecvente 500 de verbe din fiecare dintre cele 5 secțiuni ale corpusului şi am extras din ROMBAC câte 1000 de propoziții din fiecare secțiune, astfel încât fiecare dintre cele 500 de verbe frecvente să apară în cel puțin două propoziții. În mod natural, unele verbe se vor întâlni în mai multe sau chiar toate secțiunile corpusului; în plus, cele mai multe dintre propoziții conțin mai mult de un verb predicativ. De aceea, multe dintre verbe vor avea în resursa noastră o frecvență mai mare de 2, aceasta fiind doar frecvența minimă garantată fiecărui verb. Folosirea frecvenței verbelor în corpus drept criteriu de selecție a propozițiilor, alături de balansarea domeniilor, ne asigură că structurile sintactice reprezentate în resursa noastră sunt reprezentative pentru limba română şi în acelaşi timp diverse. Cele 5000 de propoziții extrase astfel din ROMBAC vor reprezenta corpusul de lucru în continuare.

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 propozițiilor folosind un instrument grafic pentru facilitarea acesteia şi 2) adnotarea automată folosind instrumente disponibile (statistice sau bazate pe reguli) şi corectarea manuală ulterioară a soluțiilor furnizate de acestea. Am optat pentru a doua strategie bazându-ne pe rezultatele pozitive obținute în experimente similare [32, 33].

Exploatând similaritatea tipologică între limbile română, spaniolă şi catalană, am reprodus procedura folosită în (Arias et al) de adnotare a unui treebank catalan folosind un model statistic spaniol. Astfel, am adnotat corpusul nostru cu **MaltParser**[7] antrenat pe **treebank-ul de limbă spaniolă IULA LSP**[8] [34] – şi am corectat rezultatele obținute. O astfel de adnotare trans-lingvistică este posibilă deoarece MaltParser oferă opțiunea antrenării de modele statistice de-lexicalizate, bazate exclusiv pe secvențe de etichete morfo-sintactice, şi nu pe cuvinte. Ne-am bazat pe faptul că cele două limbi implicate, româna şi spaniola, împart şabloane sintactice instanțiate prin secvențe de părți de vorbire similare. În exemplul de mai jos puteți observa că cele două propoziții, traduceri reciproce în română şi spaniolă, corespund unor secvențe de părți de vorbire similare (diferențele sunt marcate cu caractere italice).

- Marți[adv], [punct] miniştrii[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].

Treebank-ul de dependențe IULA Spanish LSP este un corpus tehnic, care numără 40000 de propoziții şi este disponibil gratuit prin platforma META-SHARE[9] (ca, de altfel şi corpusul ROMBAC) printr-o licență Creative Commons. Corpusul original pe care se bazează acest treebank, *Corpus Técnic de l'IULA*, cuprinde texte scrise din domeniile: juridic, economic, ştiința calculatoarelor, mediu şi medicină. Modelul statistic antrenat cu MaltParser pe acest corpus, pe care l-am utilizat în adnotarea noastră, este un model performant, care produce adnotări cu scor LAS[10] de 94% atunci când este utilizat pe texte în limba spaniolă. În aceste condiții, este de aşteptat ca, aplicat pe limba română, modelul să producă rezultate comparabile cu cele din experimentul pentru catalană (79% scor LAS) [32].

[7] http://www.maltparser.org/
[8] http://www.iula.upf.edu/recurs01_tbk_uk.htm
[9] http://metashare.upf.edu şi http://hdl.handle.net/10230/20408.
[10] Label Atacchment Score (LAS), Scorul de Ataşare Etichetată este măsura de evaluare consacrată în domeniu şi reprezintă raportul dintre numărul de cuvinte cu centre şi etichete corect identificate şi numărul total de cuvinte din propoziție

Au fost necesare două operațiuni de armonizare a corpusurilor implicate (cel spaniol şi cel românesc ce urma a fi adnotat):

1. transformarea automată a etichetelor morfo-sintactice folosite pentru limba română[11] în etichete morfo-sintactice compatibile cu IULA LSP[12], facilitată de faptul că ambele seturi de etichete sunt derivate din specificațiile EAGLES[13].
2. Convertirea corpusului nostru din formatul XML în formatul CONLL acceptat de MaltParser:

Tabel I. Exemplu de propoziție adnotată sintactic în format CONLL, cu informațiile aşezate pe coloane separate prin tab-uri. Informațiile semnifică, în ordinea coloanelor: numărul de ordine al cuvântului în propoziție, forma cuvântului în corpus, forma cuvântului în dicționar (lema), partea de vorbire a cuvântului, eticheta morfo-sintactică a cuvântului, _ numărul de ordine în propoziție a centrului cuvântului, eticheta relației de dependență față de centru, _ , _ . Coloanele care conțin "_" au fost inițial completate cu alte tipuri de informații, irelevante pentru analiza sintactică.

1	Începetează	începeta	v	Vmip3s	_	0	ROOT	_	_
2	emisia	emisie	n	Ncfsry	_	1	subj	_	_
3	în	în	s	Spsa	_	2	pmod	_	_
4	România	România	n	Np	_	3	prep	_	_
5	a	al	t	Tsfs	_	6	det	_	_
6	postului	post	n	Ncmsoy	_	2	nmod	_	_
7	de	de	s	Spsa	_	6	pmod	_	_
8	radio	radio	n	Ncms-n	_	7	prep	_	_
9	Europa	Europa	n	Np	_	6	nmod	_	_
10	Liberă	liber	a	Afpfsrn	_	9	name	_	_
11	.	.	p	PERIOD	_	1	punct	_	_

Pentru corectura manuală am folosit **instrumentrul yEd**[14], care dispune de o interfață grafică intuitivă. Acest instrument lucrează cu formatul GRAPHML, un format uşor de utilizat dedicat grafurilor, bazat pe XML. Proprietățile structurale ale grafului codificate în GRAPHML sunt interpretate de yEd şi transpuse într-un format grafic prietenos care ajută utilizatorul să aducă modificări acestor proprietăți. Implicit, formatul CONLL a trebuit să fie convertit automat în formatul GRAPHML.

[11] http://nl.ijs.si/ME/V4/msd/html/index.html, Specificațiile morfosintactice MultText East pentru limba română
[12] http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-es.html
[13] http://www.ilc.cnr.it/EAGLES96/morphsyn/morphsyn.html
[14] http://www.yworks.com/en/products/yfiles/yed/

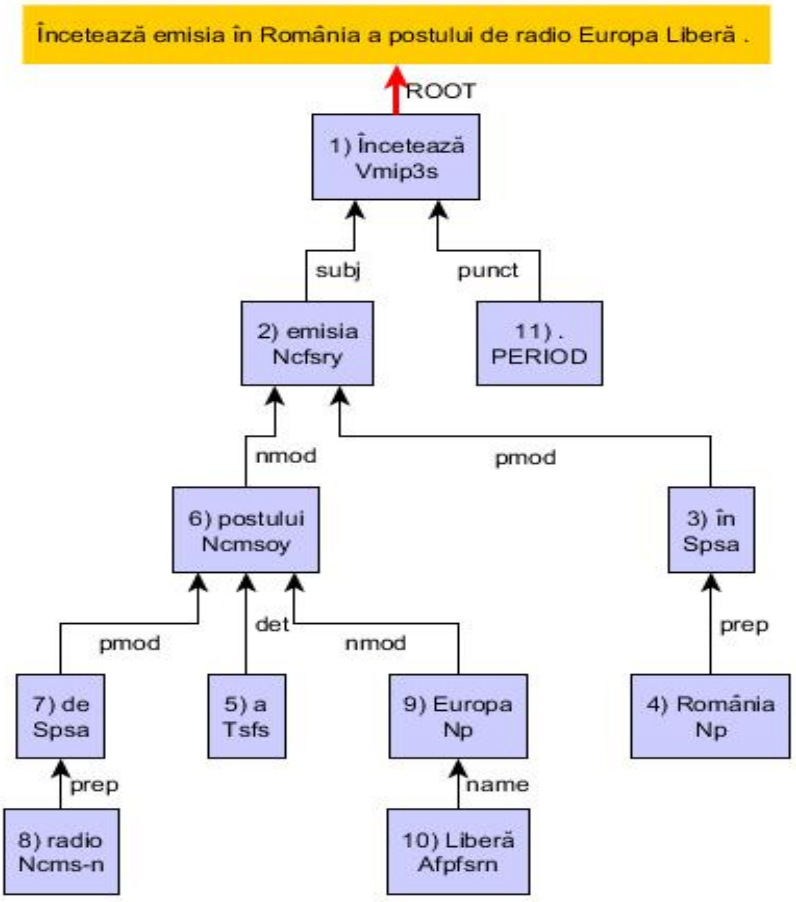

Figura 1. Formatul grafic oferit de yEd pentru propoziția din Tabelul I. Numerele de ordine ale cuvintelor, cuvintele şi etichetele morfo-sintactice sunt informații furnizate de etichetele nodurilor, în timp ce relațiile de dependență şi etichetele lor sunt reprezentate de arcele grafului. În nodul rădăcină se salvează propoziția analizată.

Pentru evaluarea rezultatelor, am folosit măsuri şi instrumente consacrate în domeniu. Competițiile CoNLL[15] 2006 şi CoNLL 2007, dedicate analizei sintactice cu dependențe şi devenite repere de evaluare a performanței parser-elor, au dezvoltat propriile scripturi Perl de evaluare, pe baza cărora s-a construit ulterior în Java **instrumentul MaltEval** [35]. MaltEval oferă, printre alte facilități suplimentare, căutare vizuală în seturile de propoziții evaluate şi sublinierea erorilor (diferențelor între fişierele gold-standard şi fişierele de test). Măsura de evaluare pe care am folosit-o este LAS, dar MaltEval oferă şi alte măsuri precum rata erorii, precizie, acurațe etc.

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

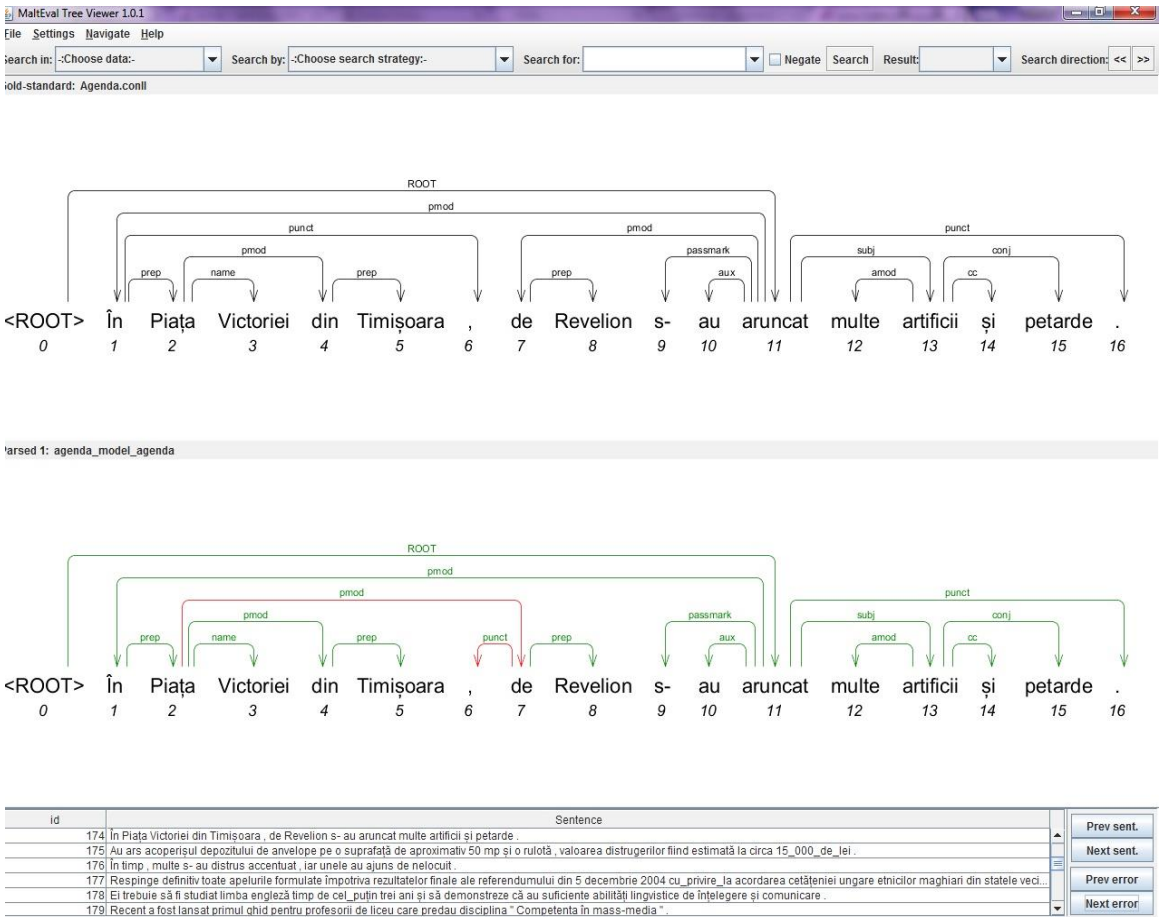

Figura 2. Opțiunea de vizualizare a erorilor (diferențele între arborele din partea superioară a imaginii şi cel din partea inferioară), căutare (butoanele de Search din partea superioară a imaginii) şi navigare între propoziții (Prev sent, Next Sent) şi erori (Prev error şi Next error) în seturile de propoziții comparate.

4. Gramatica utilizată pentru adnotare

Am ales pentru adnotare formalismul gramaticii de dependențe [36, 37], care are la bază relația de dependență între două cuvinte: un *centru* şi un *dependent*. Acest formalism oferă avantajul minimalismului (fiecare nod din structură corespunde unui cuvânt din propoziția analizată, cu excepția nodului rădăcină care este un nod artificial şi reprezintă întreaga propoziție), ceea ce face ca structurile obținute să ocupe mai puțin spațiu de reprezentare iar prelucrarea lor cu instrumente informatice să fie mai uşoară.

Setul de etichete folosit (*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. 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 *iulaLSPdep*). Astfel, avea sens să dezvoltăm pentru limba română un set de etichete în care să integrăm etichetele *iulaLSPdep*, pentru a ne uşura munca de corectură manuală. În acelaşi timp, este foarte important să producem o adnotare sintactică în concordanță cu normele internaționale, pentru a facilita utilizarea resursei noastre în proiecte multilingve viitoare. De aceea, ne-am îndreptat către o inițiativă de standardizare trans-lingvistică a metodologiei de adnotare sintactică, denumită Universal Dependency[16] (*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

[16] http://universaldependencies.github.io/docs/

mai important aspect de urmărit a fost respectarea principiilor gramaticii româneşti şi evidenţierea clară a relaţiilor sintactice specifice limbii române. Astfel, au apărut etichete noi marcate cu caractere italice în Tabelul 1, motivate de următoarele decizii:

- Clasificarea cliticelor pronominale: de dublare (*dblclitic*), posesiv (*posclitic*), reflexiv şi reciproc (ambele ca *reflclitic*).
- Diferenţierea între două argumente ale verbului care apar în aceaşi propoziţie în cazul acuzativ: obiectul direct (*dobj*, identificat prin posibilitatea dublării cliticului şi prin prepoziţia *„pe”*) şi obiectul secundar (*secobj*) (exemplu: în propoziţia *L-au ales pe Ion primar*, obiectul direct este *pe Ion* (dublat prin cliticul *L-*), iar obiectul secundar este *primar*).
- Atunci când o prepoziţie leagă cuvântul precedent de cel ce o urmează: ex. *El este teribil de timid*, unde relaţia dintre prepoziţie (*de*) şi centru (*teirbil*) va fi etichetată *post*.
- Atunci când un dependent intră într-o relaţie ternară (sintactică şi semantică) cu verbul şi cu un nominal (subiect sau obiect), acesta este legat de verb şi primeşte eticheta *spe* (element predicativ suplimentar): ex. *I-am văzut pe copii împreună*, unde *împreună* este *spe* pentru verb.
- Conjuncţiile sunt tratate în funcţie de tipul acestora: cele coordonatoare sunt ataşate primului conjunct prin relaţia *cc*, iar cele subordonatoare devin centre ale propoziţiei subordonate, al cărui verb este ataşat de conjuncţie prin relaţia *sc*: Ex. *Vreau să vii*, unde *să* este *dobj* pentru verbul *vreau*, iar *vii* este *sc* pentru conjuncţie.
- Elementele corelative primesc eticheta *correl*: *O voi chema fie pe Maria, fie pe Ana*, unde primul *fie* este *correl* pentru al doilea conjunct.

Înainte de etapa de corectare, am transferat automat din setul de etichete sintactice IULA în setul nostru de etichete de dependenţe tot ce s-a putut transfera ne-ambiguu. Etichete precum *spec* sau mod (cu mai mult de o etichetă echivalentă în setul românesc) au fost lăsate spre dezambiguizare în etapa de corectare.

Tabel II. Corespondenţele între inventarele *Rodep*, *iulaLSPdep* şi *UD*

ROdep	***iulaLSPdep***	***UD***	***ROdep***	***iulaLSPdep***	***UD***
acl		acl	name		name
advcl		advcl	neg	Neg	neg
advmod	mod	advmod	nmod	Mod	nmod
agc	byag	agc	parataxis		parataxis
amod	spec	amod	passmark	Passm	
appos	mod	appos	pmod	Mod	
aux	aux	aux	pobj	Oblc	
auxpass		auxpass	poss		poss
cc	coord	cc	*possclitic*		
compound		compound	*post*		
conj	conj	conj	pred	prd, atr	
correl			prep	Comp	case
dblclitic			punct	Punct	punct
dep	unknown	dep	*reflclitic*		
det	spec	det	remnant		remnant
discourse		discourse	reparandum		reparandum
dislocated		dislocated	root		root
dobj	do	dobj	*sc*		
foreign		foreign	*secobj*		
goeswith		goeswith	*spe*		
iobj	io	iobj	subj	Subj	nsubj, csubj,

					cubjpass
list		list	voc	Voc	vocative
mark		mark	xcomp	Oprd	xcomp
mwe		mwe			

5. Arhitectura proiectului şi evaluarea performanței modelelor statistice utilizate

Aşa cum menționam în secțiunea 3, am ales ca în dezvoltarea treebank-ului 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) şi un model antrenat pe limba spaniolă (pe treebank-ul IULA LSP) şi să corectăm adnotarea propusă pentru a corespunde standardelor gramaticii limbii române. Pentru a menține consistența adnotării, am decis să pornim, într-o primă etapă, cu prima jumătate (2500 de propoziții) a corpusului de adnotat în care am inclus propozițiile de lungime cuprinsă între 10 şi 30 de cuvinte, şi să lăsăm propozițiile mai lungi, şi implicit mai complexe sintactic, pentru adnotare şi corectare într-o etapă viitoare. În acest moment am finalizat corectarea şi adnotarea primei jumătăți din corpus şi am început corectura la cea de-a doua jumătate.

Am început adnotarea automată cu un set de 500 de propoziții din sub-corpusul jurnalistic folosind modelul statistic de-lexicalizat de limbă spaniolă, dar după corectarea acestora am decis să le folosim pentru re-antrenarea unui model lexicalizat pe limba română, intuind că modelul obținut va avea performanțe mai bune decât cel spaniol, chiar dacă este antrenat pe incomparabil mai puține propoziții: 500 versus 40000. 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 3, ciclul de lucru este: 1) adnotare cu modelul statistic cel mai performant la dispoziție (în imagine, săgețile orientate în jos indică procesul de adnotare); 2) corectura setului de propoziții adnotat la pasul 1 (în imagine, săgețile orientate în sus indică procesul de corecta); 3) adăugarea setului corectat la corpusul de antrenare şi re-antrenarea unui model extins, mai performant decât precedentul (săgeata mare orizontală din imagine simbolizează antrenarea progresivă pe seturi de date tot mai mari).

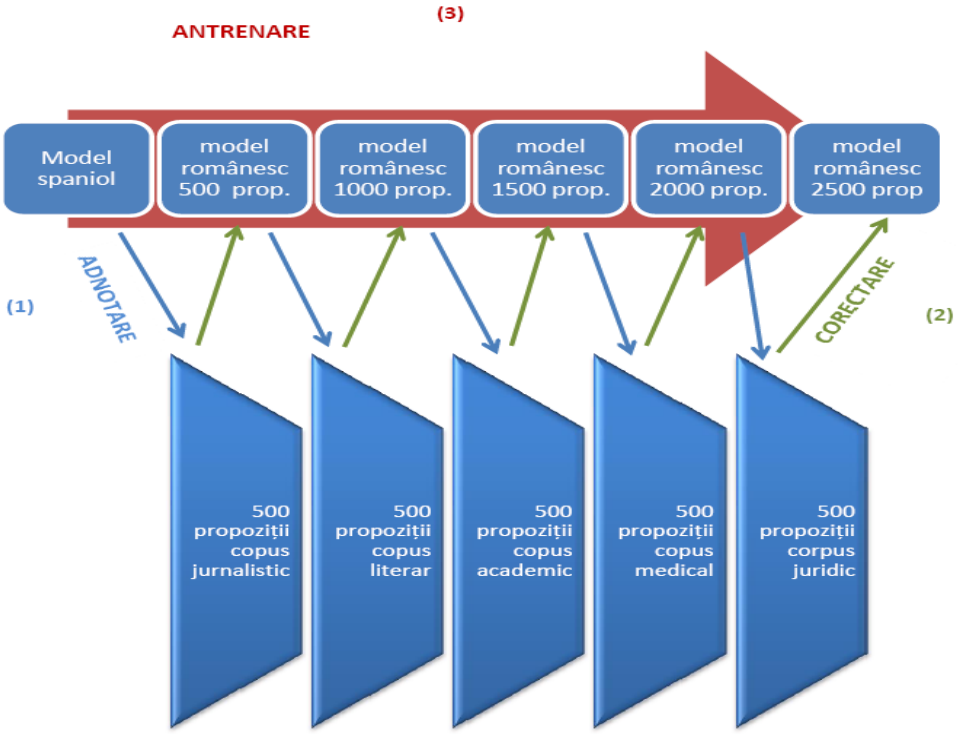

Figura 3. Ciclul de adnotare/corectare/re-antrenare pentru prima jumătate a corpusului. În acest moment dispunem de un model statistic lexicalizat pe limba română, antrenat pe 2500 de propoziții, cu care vom continua munca de corectare.

Folosind instrumentul MaltEval introdus în secțiunea 3, am putut evalua scorul LAS pentru fiecare nou model statistic antrenat luând ca referință (gold standard) varianta corectată a fiecărui set de propoziții și ca mulțime de test varianta ne-corectată. Așa cum se poate observa în Tabelul III, creșterea performanței la trecerea de la modelul spaniol la cel românesc a fost substanțială: de la 0,243 la 0,580. Experiența noastră în munca de corectare manuală confirmă această creștere. Ulterior, scorul LAS a continuat să crească odată cu creșterea dimensiunilor modelului statistic. Excepție face ultimul set de propoziții corectat, extras din sub-corpusul juridic, deoarece acest tip de text are o structură aparte: propozițiile (articole de lege) încep adeseori cu un identificator specific (ex. „Alin. 1", „Art. 2"), pe care am decis să-l adnotăm ca legat de centrul propoziției (verbul) prin relația *parataxis*[17]. Cum acest gen de structură sintactică nu se întâlnește în nici unul dintre sub-corpusurile deja incluse în modelul statistic, parser-ul nu a putut analiza corect aceste structuri, uneori eroarea propagându-se și la alte elemente ale propoziției.

Corpus folosit pentru antrenarea modelului statistic	Set de propoziții adnotate	LAS
Iula Spanih LSP	Jurnalistic	0,243
Jurnalistic	Proză	0,580
Jurnalistic+Ficțiune	Academic	0,738
Jurnalistic+Ficțiune +Academic	Medical	0,773
Jurnalistic+Ficțiune +Academic+Medical	Juridic	0,710

6. Concluzii

Metodologia aleasă pentru dezvoltarea treebank-ului de limbă română s-a dovedit inspirată, reușind ca în timp scurt (aproximativ șase luni) să obținem 2500 de propoziții corect adnotate și un model statistic de calitate satisfăcătoare, care să garanteze că timpul necesar pentru adnotarea celor 2500 de propoziții rămase se va reduce substanțial. De asemenea, ne așteptăm ca scorul LAS să continue să crească în etapele de re-antrenare succesive viitoare, chiar dacă într-un ritm tot mai lent: performanțele oricărui instrument statistic sunt tot mai greu de îmbunătățit când valorile măsurilor de evaluare se apropie de 1. Reamintim că propozițiile din a doua etapă de corectare vor fi propoziții de lungime mai mare (între 30 și 40 de cuvinte), aspect care va influența de asemenea performanța adnotării automate, introducând mai multă complexitate sintactică.

După finalizarea sa, corpusul va fi integrat în CoRoLa și folosit în continuare ca model statistic pentru adnotarea de noi texte. De asemenea, vom produce o variantă a sa complet compatibilă standardelor *UD* și vom distribui resursa și în cadrul acestui proiect.

* * *

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

BIBLIOGRAFIE

1. **TRANDABĂȚ, D.; IRIMIA, E.; BARBU MITITELU, V.; CRISTEA, D.; TUFIȘ, D.:** The Romanian Language in the Digital Age. Limba română în era digitală. In White Papers Series (Rehm, Georg and Uszkoreit, Hans). Springer-Verlag, Berlin, Heidelberg, 2012.
2. **TUFIȘ, D.; CRISTEA, D.:** Methodological issues in building the Romanian Wordnet and consistency checks in BalkaNet. In Proceedings of LREC 2002 Workshop on Wordnet

[17] În *UD*, relația *parataxis* se formează între elemente alăturate fără coordonare sau subordonare explicită. Dependentul în relația *parataxis* nu este nici argument al cuvântului centru și apare adeseori urmat de semne de punctuație precum ":" sau ";".

Structures and Standardisation (Christodoulakis, Dimitris, N. and Kunze, Claudia and Lemnitzer, Lothar). Las Palmas, Spain, may 2002 pp. 35-41.

3. **BARBU MITITELU, V.; DUMITRESCU, Ş. D.; TUFIŞ, D.:** News about the Romanian Wordnet. In Proceedings of the 7th International Global WordNet Conference. Tartu, Estonia, 2014.
4. **BARBU MITITELU, V.; IRIMIA, E.:** 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, Mihaela and Iftene, Adrian and Barbu Mititelu, Verginica and Cristea, Dan and Tufiș, Dan). Editura Universității „Alexandru Ioan Cuza", Iaşi, September 2014, pp. 57–66.
5. **TUFIŞ, D.; ION, R.; DUMITRESCU, Ş. D.:** Wikipedia as an SMT Training Corpus. In Proceedings of the International Conference on Recent Advances on Language Technology (RANLP 2013). Hissar, Bulgaria, September 2013.
6. **IRIMIA, E.:** EBMT experiments for the English-Romanian Language Pair. In Recent Advances in Intelligent Information Systems (Klopotek et al.). Springer, Warsaw, 2009, pp. 91-102.
7. **TUFIŞ, D.; BOROŞ, T.; DUMITRESCU, Ş. D.:** The RACAI Speech Translation System. In Proceedings of the 7th International Conference on Speech Technology and Human-Computer Dialogue (SPED 2013). Cluj-Napoca, October 2013.
8. **OCH, F.-J.; TILLMANN, CH.; NEY, H.:** 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, 1999, pp. 20–28.
9. **MARCU, D.; WONG, W.:** A Phrased-Based, Joint Probability Model for Statistical Machine Translation. Proceedings of the Conference on Empirical Methods in Natural Language Processing, Philadelphia, PA, July, 2002, pp. 133-139.
10. **YAMADA, K.; KNIGHT, K.:** A Decoder for Syntax-based Statistical MT. Proceedings of the 40th Annual Conf. of the Association for Computational Linguistics, Philadelphia, PA, July, 2002, pp. 303-310.
11. **COLHON, M.:** Syntactic Translation Patterns from a Parallel Treebank. Workshop on Computational Linguistics and Natural Language Processing of Balkan Languages, Balkan Conference in Informatics, 2012, pp. 85-88.
12. **GARSIDE, R.; LEECH, G.; VARADI, T.:** Manual of Information for the Lancaster Parsed Corpus. Lancaster University, 1992.
13. **TAYLOR, A.; MITCHELL, M.; SANTORINI, B.:** The PENN Treebank: An Overview. In ABEILLE, A (ed.). Treebanks. Building and Using Parsed Corpora. Kluwer Academic Publishers, 2003, pp. 6-22.
14. **SKUT, W.; KRENN, B.; BRANTS, TH.; USZKOREIT, H.:** An Annotation Scheme for Free Word Order Languages. Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP-97). Washington, DC, USA.
15. **BRANTS, S.; DIPPER, S.; EISENBERG, P.; HANSEN, S.; KONIG, E.; LEZIUS, W.; ROHRER, C.; SMITH, G.; USZKOREIT H.:** TIGER: Linguistic Interpretation of a German Corpus. Journal of Language and Computation, 2004 (2), pp. 597-620.
16. **HAJIC, J.; HAJICOVA, E.; PAJAS, P.; PANEVOVA, J.; SGALL, P.; VIDOVA HLADKA, B.:** Prague Dependency Treebank 1.0 (Final Production Label). CD-ROM, CAT: LDC2001T10, ISBN 1-58563-212-0, Linguistic Data Consortium.
17. **HRISTEA, F.; POPESCU, M.:** A Dependency Grammar Approach to Syntactic Analysis with Special Reference to Romanian. F. Hristea şi M. Popescu (coord.), Building Awareness in Language Technology, Bucureşti, Editura Universităţii din Bucureşti, 2003, pp. 9-16.

18. **BICK, E.; GREAVU, A.:** A Grammatically Annotated Corpus of Romanian Business Texts. Proceedings of Multilinguality and Interoperability in Language Processing with Emphasis on Romanian, Editura Academiei Române, 2010, pp. 169-183.

19. **PEREZ, A.-C.:** Resurse lingvistice pentru prelucrarea limbajului natural. PhD thesis, "Al. I Cuza" University, Iaşi, 2014.

20. **MĂRĂNDUC, C.; PEREZ. A.-C.:** A Romanian dependency treebank. CICLing *2015*, Cairo, 14-20 Aprilie.

21. **PUNYAKANOK, V.; ROTH, D.; YIH, W.-T.:** The Importance of Syntactic Parsing and Inference in Semantic Role Labeling. Computational Linguistics, 34(2), 2008, pp. 257-287.

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

23. **WANG, Q. I.; SHUURMANS, S.; LIN, D.:** Strictly Lexical Dependency Parsing. In H. Bunt, P. Merlo, J. Nivre (eds.), Trends in Parsing Technology, Springer, 2010, pp. 105-120.

24. **COLLINS, M.:** A new statistical parser based on bigram lexical dependencies, 1996.

25. **COLLINS, M.:** Head-driven statistical models for natural language parsing. Ph.D. thesis, Computer Science Department, University of Pennsylvania, 1999.

26. **KLEIN, D.; MANNING, C. D.:** 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, 2003, pp. 3-10.

27. **CHEN, D.; MANNING, C. D.:** A Fast and Accurate Dependency Parser using Neural Networks. Proceedings of EMNLP 2014.

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

29. **CĂLĂCEAN, M.; NIVRE, J.:** A Data-Driven Dependency Parser for Romanian. Proceedings the Seventh International Workshop on Treebanks and Linguistic Theories, 2009, pp. 65-76.

30. **SERETAN, V.; WEHRLI, E.; NERIMA, L.; SOARE, G.:** FipsRomanian: Towards a Romanian Version of the Fips Syntactic Parser. Proceedings of the Seventh International Conference on Language Resources and Evaluation, Valletta, Malta, 2010.

31. **ION, R.; IRIMIA, E.; ŞTEFĂNESCU, D.; TUFIŞ, D.:** ROMBAC: The Romanian Balanced Annotated Corpus. Procedings of LREC 2012, Istanbul, Turkey.

32. **ARIAS, B.; BEL, N.; FOMICHEVA, M.; LARREA, I.; LORENTE, M.; MARIMON, M.; MILA, A.; VIVALDI, J.; PADRO, M.:** Boosting the creation of a treebank. Proceedings of LREC 2014, Reykjavik, Iceland.

33. **FLOREA, I. M.; REBEDEA, T.; CHIRU, C.G.:** Parser de dependențe pentru limba română realizat pe baza parserelor pentru alte limbi romanice. Revista Română de Interacțiune Om-Calculator 7(1), 2014, pp. 1-20.

34. **MARIMON, M.; BEL, N.:** Dependency structure annotation in the IULA Spanish LSP Treebank. Language Resources and Evaluation. Amsterdam: Springer Netherlands, 2014.

35. **NILSSON, J.; NIVRE, J.:** MaltEval: An Evaluation and Visualization Tool for Dependency Parsing. Proceedings of LREC 2008, Marrakesch, Morocco.

36. **TESNIERE, L.:** Éléments de syntaxe structurale. Paris, Klincksieck, 1959.

37. **MELCUK, I. A.:** Dependency syntax : theory and practice. Albany, State University Press of New York, 1987.