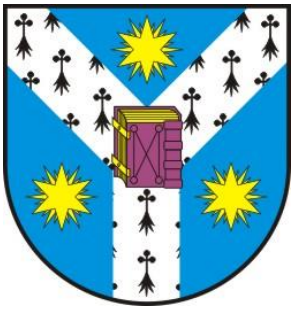
Universitatea „Alexandru Ioan Cuza” din Iași, Romania
Facultatea de Informatică



INSTRUMENTE PENTRU PROCESAREA LIMBII ROMÂNE VECHE
(Corpusuri de Antrenare, Modele, Statistici)

Rezumat

Doctorand
CĂTĂLINA MĂRĂNDUC

Coordonator Științific
DAN CRISTEA

Cuprins

Pentru o lungă perioadă de timp, lingvistica modernă a avut un punct de vedere destul de limitat asupra obiectului de studiu: ne interesează limba contemporană, mai ales pentru că ne este direct accesibilă, suntem în contact direct doar cu ea, iar orice altceva prezintă foarte puțin interes. Prin urmare, nu putem fi surprinși că acesta a fost și punctul de vedere al informaticienilor care s-au ocupat de prelucrarea limbajului natural, cel puțin până la sfârșitul secolului trecut. Mai mult, pe măsură ce domeniul lingvisticii computaționale s-a maturizat și a început să aibă o istorie, au apărut probleme. Unele sunt legate de stocarea informațiilor din cercetările anterioare. Deoarece programele care procesează datele lingvistice au evoluat atât de rapid, datele vechi nu mai puteau fi citite de noile programe. Mai mult chiar, stocarea datelor lingvistice din secolele trecute s-a dovedit a fi o sarcină dificilă, deoarece computerele trebuie să fie capabile să acceseze aceste informații și să permită căutări avansate asupra acestor date.

În primele faze de construire a analizoarelor morfologice și sintactice automate, scopul era de a putea procesa doar fraze simple, astfel încât construcțiile dificile au fost eliminate în mod deliberat din corpusurile de antrenament. Scopul final al întreprinderii noastre este, totuși, acela de a avea o tehnologie capabilă să analizeze orice frază din limba română, nu doar cele mai simple și nu doar cele din limba utilizată curent.

În această teză, nu suntem interesați să construim un corpus mare de limbă standard contemporană. Un corpus mare pentru limba română standard a fost deja

construit. Se numește CoRoLa[1] (Barbu-Mititelu *et al.*, 2017), și a fost lansat în noiembrie 2017, având în prima livrare aproape 400.000 de fișiere, aproximativ 1,26 miliarde de token-uri (cuvinte + punctuații) și aproximativ 900.000.000 de apariții de cuvinte. Acest corpus este în întregime adnotat morfologic și se află în curs de adnotare sintactică automată.

În consecință, scopul nostru în această teză este de a construi corpusuri pentru limba română nestandardizată: un corpus pentru chat, un corpus pentru limba română veche și un corpus pentru limba română regională, vorbită în Moldova pe ambele maluri ale Prutului. Apoi vom antrena diverse instrumente de prelucrare pe corpusurile create și le vom adapta astfel încât să obținem rezultate optime în analiza morfo-sintactică automată. Un corpus de limba română standard contemporană nu este reprezentativ pentru modul în care arată cu adevărat limba naturală. Textele regionale, textele poetice, limbajul vorbit, limbajul familiar, limbajul din social media, textele jurnalistice care pot utiliza imagini poetice sau mijloace specifice limbajului oral-familiar, toate aceste tipuri de limbaj conțin frecvent fenomene non-standard. De asemenea, ele trebuie să fie prelucrate automat.

Scopul cercetării noastre este de a crea un corpus amplu și echilibrat de variante nestandardizate ale limbii române, cu o atenție specială asupra limbii române vechi, și de a studia modul în care diverse instrumente de limbaj natural pot fi antrenate pe fiecare dintre variantele de limbă pentru care am reușit să avem un corpus de

[1] http://corola.racai.ro/

antrenament suficient. Vom vedea care sunt particularitățile fiecărei astfel de variante de limbă, ce erori apar în procesare, care sunt cauzele acestora și cum putem optimiza instrumentele de procesare și le putem crește performanța.

Structura acestei teze este următoarea. Prezentăm în capitolul II conținutul și amploarea actuală a corpusurilor de training UAIC sintactic, UAIC semantic, UD sintactic. De asemenea, vor fi descrise programele de transformare și acuratețea acestora: cele două variante ale Treeops și programul de transpunere din XML în formatul CONLLU a băncilor de arbori UD.

Apoi, în capitolul III, vom explica seturile de etichete morfologice pentru fiecare convenție și stratul morfologic de adnotare. De asemenea, descriem o tentativă de creare a unui POS-tagger pentru limba română veche (Mărănduc *et al.* 2017), bazat pe POS-tagger-ul hibrid UAIC. Acest instrument este încă în lucru, deoarece este nevoie de un corpus de instruire de cel puțin 60.000 de propoziții.

În capitolul IV prezentăm convențiile sintactice de adnotare a celor două corpusuri, diferențele dintre ele, instrumentele de transpunere a formatului de bază în cel internațional și analizatorii sintactici aplicați pe ambele convenții, precum și statisticile.

În capitolul V este prezentată convenția semantică și regulile de transpunere a formatului de bază în cel semantic.

În capitolul VI se descrie operațiunea de extragere a structurilor pentru elaborarea Dicționarului verbelor românești, acesta aflându-se într-o fază

incipientă de elaborare. Prezentăm tipurile de accesări de arbori, în căutare, în corpusurile elaborate.

În capitolul VII este prezentată o aplicație de aliniere a Noului Testament de la Alba Iulia (1648) cu versiunile greacă, latină și slavonă veche ale cărții sfinte. Aceste versiuni reprezintă pentru potențialii specialiști surse ale traducerii românești, iar studiul lor comparativ este foarte important pentru a clarifica originea în limba română veche a cuvintelor și structurilor.

În încheierea capitolului VIII se face un scurt rezumat al contribuțiilor originale ale tezei și se indică direcțiile în care trebuie dezvoltate în continuare resursele create și modalitățile de utilizare a acestora în cercetarea lingviștilor și a lingviștilor computaționali.

În cele ce urmează vom rezuma fiecare capitol în parte.

Primul capitol oferă o introducere a lucrării prezentate în teză și o descriere detaliată a pașilor făcuți în realizarea corpusului nostru.

Capitolul II prezintă corpusul de test și colecţiile noastre de arbori sintactici. Corpusul de antrenament UAIC-RoDia include acum 32.753 de propoziții și 671.235 token-uri. Dar un parser sintactic pentru limba română ar trebui să fie format separat pentru fiecare dialect. Conversația și limba veche sunt foarte diferite din punct de vedere sintactic de limba actuală. Ca atare, am alcătuit separat un parser de dependente pentru textele de pe rețelele de socializare. Parserul, instruit pe un sub-corpus de numai 2 579 de propoziții, a obținut rezultate mai bune după adăugarea tuturor textelor utilizate in ziua de azi (Perez *et al.*, 2016 a, b). Prin urmare, atât secțiunile contemporane, cât și cele

specializate (vechi, sociale etc.) ar trebui să fie îmbunătățite pentru a creste acuratețea unui proces complet automat de adnotare sintactică. Cu tehnologia actuală și respectând aceleași reguli, limita minimă pentru limbajul standard contemporan este de 10 000 de propoziții. Pentru o categorie de texte în care se aplică o varietate de reguli gramaticale în timp, numărul de propoziții ar trebui să fie mai mare din cauza evoluției limbii. Textele de pe rețelele de socializare și limba vorbită, fiind creative, își asumă foarte multă libertate în ceea ce privește regulile gramaticale. Prin urmare, avem nevoie de mai mult de 10 000 de propoziții pentru fiecare dintre aceste registre. De exemplu, pentru o pregătire optimă a post-tagger-ului, care este o etichetă specială atribuită fiecărui cuvânt într-un corpus text pentru a indica partea de vorbire și adesea și alte categorii gramaticale, estimăm dimensiunea minimă a corpusului de test la 70 000 de propoziții. Ceea ce am realizat până acum este un corpus considerabil în trei formate. În tabelul 8, întregul conținut al bazei noastre de date poate fi văzut.

Tabel 8. Conținutul tuturor formatelor Treebank-ului nostru

Nr. crt.	Format	Fraze	Token-uri	Media de token-uri/fraze
1	**UAIC sintactic XML**	**32 753**	**671 235**	**20,49**
	Din care, Old-Ro	19 254	126 564	21,47
1a	**Forma**	2 794	46 708	16,71

	cuvântului Cirilic			
2	**UD sintactic CoNLLU**	**16 936**	**348 562**	**20,58**
	Din care, Old-Ro	14 437	29 109	20,57
	Din care, Folclor	2 499	50 077	20,03
3	**UAIC semantic XML**	**5 566**	**99 341**	**17,84**
	Din care, Old-Ro	5 032	88 350	17,55
	Din care, folclor	230	4 157	18,07
Total	**corectate**	**55 255**	**1119138**	**20,25**

În **capitolul III** arătăm că scopul nostru nu este doar creșterea corpusului ci și instruirea acestor tipuri de corpusuri cu mai multe instrumente pentru prelucrarea limbii române arhaice. Orice aplicație viitoare bazată pe aceste corpusuri care va fi creată, în orice fel de convenție, nu trebuie să-și bazeze segmentarea textelor în cuvinte și adnotarea lor morfologică. Astfel, adnotarea morfologică automată este baza unei prelucrări mai avansate a limbajului natural și trebuie să începem cu aceasta. Pentru a construi un POS-tagger pentru limba română veche, având un instrument similar testat pe limba contemporană, este necesar, mai întâi, să se stabilească lista de etichete, apoi să se adauge la formele de cuvinte care corespund limbii vechi, și, în cele din

urmă, să aibă un corpus de antrenament cu un număr mare de propoziții adnotate în mod constant cu noul set de tag-uri.

Informațiile sintactice (**capitolul IV**) sunt esențiale pentru procesarea limbajului natural. Șirul de intrare este mai întâi descompus de programe cum ar fi splitter și tokenizer, apoi segmentele sunt analizate de POS-tagger și, în cele din urmă, informațiile sintactice sunt recompuse într-o structură. Fiecare simbol are un început, cu excepția rădăcinii propoziției. Relația prin care se leagă de început poate fi opțională sau obligatorie și este de mai multe tipuri. Pentru fiecare dintre aceste locuri determinate în structură, sunt alese elemente cu anumite caractere morfologice. Sintaxa este realizata atât morfologic cât și de semantic. În acest capitol, prezentăm diferite constrângeri întâlnite în procesul nostru atunci când adnotăm aranjamentul, substantivele predicative sau rolurile duble.

Scopul nostru este de a introduce în arborele sintactic UD-Româno-non-Standard alte texte vechi în limba română. Arborele sintactic al UAIC are 6 590 de propoziții, cu 162 231 de cuvinte și semne de punctuație în lucru. Pe lângă română arhaică și folclor, corpusul va fi inserat în arborele sintactic UD-Romanesc-Nonstandard. Vom continua să editǎm dicționarul de modele al verbelor românești. Va avea un site, unde vor fi marcate modelele limbii române arhaice și regionale. Aceste modele pot fi folosite pentru a crea un nou parser sintactic sau un model semantic-sintactic mixt bazat pe constrângeri, permisiuni și interdicții.

În **capitolul V** vă prezentăm detalii despre adnotarea semantică. Am făcut tabelul de transpunere

pentru transpunerea automată a convențiilor noastre în cele UD și o parte din UAIC-RoDiaTb a fost transpusă în UDV în 2016, de către grupul RACAI (Institutul de Cercetare a Inteligenței Artificiale). Există multe probleme teoretice care diferențiază convențiile UAIC de cele UD. De exemplu, abordarea cuvintelor relaționale. Categoriile sintactice sunt clasificate în conformitate cu convențiile UD în ceea ce privește clasele morfologice (adică adjectival, adverbial, modificator nominal), în plus, considerăm că informațiile sintactice ar trebui corelate cu cele semantice. Acest capitol propune un tip de adnotare semantică cu mai multe categorii, deoarece ne propunem să păstrăm toate informațiile care au fost adnotate în stratul sintactic UAIC. Aceste informații sunt importante deoarece pot fi exploatate de alte aplicații. Un alt scop a fost acela de a găsi un standard internațional de adnotare cu categorii similare, având în vedere o afiliere viitoare. Asemănările cu stratul telegramatic al PDT și cu categoriile logice AMR sunt evidente. Cu toate acestea, există diferențe, deoarece graficul rezultat al adnotării semantice AMR nu este un arbore de dependență, iar nodurile nu sunt cuvinte, ci concepte. Pentru a arăta izomorfismul dintre sintactică și structurile semantice, am ales să construim un corpus de arbori de dependență semantică, cu etichete similare cu stratul telegramatic al PDT.

Am discutat, de asemenea, procesul de transformare a adnotării sintactice a treebank-ului de dependență UAIC RoDia în adnotarea logico-semantică. Această transformare se face automat pentru relații sintactice non-ambigue și manual pentru relații ambigue. În viitor, ne propunem să transformăm adnotarea

sintactică și morfologică a celei de-a doua părți a noului Testament din 1648 (faptele Apostolilor și scrisorile Apostolilor) într-o adnotare semantică. Vom constitui un parser statistic pe acest corpusul, pentru ca parserul să învețe să transforme relațiile sintactice ambigue.

Capitolul VI prezintă disponibilitatea resurselor noastre. Corpusul sintactic UAIC RoDia dependență se află în formatul XML și include facilități pentru căutarea avansată in arbori. Acest arbore sintactic permite accesul oricui să contribuie la elaborare sau îmbunătățire. Site-ul va include link-uri către corpusul pentru limba română UD, în format CONLLLU, care sunt, de asemenea, open source. Pentru a permite utilizatorului să descarce date în formatul preferat, vor fi incluse convertoarele XML-CONLLU și invers.

Un site web este locația definitorie a unei resurse. Nu poate fi găsită și utilizată de persoanele interesate dacă nu are o adresă web. Fiecare limbă din familia arborilor sintactici Big UD la care au contribuit peste 150 de echipe de cercetare răspândite în întreaga lume, utilizează același format de reprezentare de bază, arbori sintactici specifici fiind fiecare situat undeva în spațiul virtual. Din convențiile specifice de adnotare ale limbilor contribuitoare, forma UD este abstractizată ca o convenție internațională comună și, pentru fiecare arbore sintactic care contribuie, această formă se obține fie printr-un proces de transformare semiautomat, fie complet automat. Toate resursele sunt legate, aliniate, comparate și integrate în marea familie UD. Autorii, conținutul și site-urile originale sunt difuzate prin articole descriptive, care prezintă și rezultatele activității creatorilor.

În **capitolul VII** prezentăm alinierea Noului Testament și convențiile sale de adnotare. Alinierea este utilă pentru traduceri, studiul etimologiei și stabilirea primelor atestări. De asemenea, orice alte adnotări sau informații care au fost adăugate la Noul Testament (ordinea pragmatică a cuvintelor, particule de discurs, referință pronominală sau evenimente de fundal) pot fi importate în Noul Testament românesc.

Textele sunt adnotate anterior în convențiile UAIC și un program a făcut transformarea automată (supravegheată) în convențiile UD. Deci, UD_Romanian-Nonstandard face parte din UAIC-RoDia Dependency Treebank (RoDia – pentru Diacronica românească), care este recunoscută în catalogul internațional de resurse, cu id-ul ISLRN 156-635-615-024-0. Momentan am inclus în UD_Romanian-Nonstandard treebank cele 11 documente în care Noul Testament (Alba Iulia, 1648) a fost prelucrat în XML în format UAIC. Aceste documente au fost validate și transformate în convențiile UD.

După cum se vede în acest capitol, procesul de potrivire automată ar trebui să fie supravegheat de specialiști lingvistici cunoscători. Dar efortul este justificat de marile beneficii ale studierii proceselor folosite de traducătorii români la scrierea cărților sfinte din secolul al XVII-lea.

În **ultimul capitol** prezentăm concluziile și lucrările ulterioare. Prelucrarea limbajului natural este o activitate esențială pentru lumea modernă, pentru accesul la patrimoniul cultural, în care putem efectua căutări de informații, traduceri automate, rezumate.

Caracterul specific al treebank-ului nostru este că el conține o varietate de stiluri de limbă română, vechi contemporan și regional, cu intenția de a fi un corp de instruire atât pentru româna standard, cât și pentru cea nonstandard.

Ne vom concentra în continuare asupra studiului românei vechi, care are patru secole de evoluție, iar ponderea acestora ar trebui echilibrată, precum și ponderea stilurilor; deocamdată predomină stilul bisericesc în defavoarea narațiunii sau a stilului legislativ.

Numărul total de propoziții ar trebui să ajungă la 70.000, pentru a permite antrenarea corectă a POS-tagger-ului pe propriul corp și pentru a extrage statistici adecvate din acesta. Apoi, antrenamentul analizatorului ar putea fi efectuat pe sub-corpuri, adică secole și stiluri, fiecare având aproximativ 10.000 de propoziții.

Precizia parserului sintactic pe formatul de bază trebuie crescută, prin creșterea corpusului de antrenament și a consistenței acestuia. După ce se ajunge la o acuratețe de peste 85% (=LAS), supravegherea rezultatelor va fi mai ușoară și prin metoda bootstrapping corpusul de antrenament va fi mărit și bine structurat. Ultimele teste au arătat o creștere a preciziei de LAS=87% pe anumite sub-corpi, ceea ce este încurajator.

Introducerea datelor în lexiconul POS-tagger ar trebui continuată permanent. Programul care extrage noi combinații cuvânt-lemă-MSD din treebank corectat încă lasă în urmă numeroase erori ce trebuie eliminate din treebank (nu din rezultatul programului; după eliminarea erorilor și re-antrenarea parserului, se speră ca aceleași

erori să nu mai apară). Prin aceste corecții asigurăm atât consistența corpului, cât și pregătirea noilor date care urmează să fie introduse în lexic (rămânând cele corecte).

O altă operațiune necesară ar fi separarea lanțului de procesare automată, astfel încât ieșirea POS-tagger-ului să poată fi corectată înainte ca analizatorul sintactic să fie aplicat. Aceste programe funcționează în prezent și sunt în curs de dezvoltare. Analizatorul sintactic trebuie să stabilească legături între analiza morfologică a cuvântului analizat și a predecesorului acestuia, pentru a propune o direcție de subordonare și un tip de relație, dar astfel de legături se realizează uneori la categoria morfologică greșită și, prin urmare, analizarea va fi greșită.

Analizatorul statistic și sintactic, Malt, trebuie antrenat separat pe cele două convenții sintactice de adnotare. Este întotdeauna antrenat pe un format CONLLU de propoziții, adică, UAIC RoDia Dependency Treebank trebuie mai întâi convertit în CONLLU, pentru antrenament, iar după antrenament, trebuie folosit pentru adnotarea automată a textelor noi. Rezultatul va fi și în CONLLU și un alt convertor va transpune noul text adnotat automat în formatul XML al băncii noastre de arbori.

Astfel de convertoare sunt necesare pentru tot felul de aplicații. Programele de căutare ale unor site-uri, precum CoRoLa, nu pot face căutări în alt format decât XML. Programul bazat pe regulile TREEOPS care poate transforma formatul de bază al băncii noastre de arbori în formatul UD are reguli pentru conversia XML.

În ceea ce privește convenția de adnotare semantică, după TREEOPS ar face transformarea parțială din convenția sintactică de bază, conform convenției semantice, am putea antrena un analizator statistic asupra rezultatului transformării, astfel încât versatilitatea relațiilor sintactice să fie rezolvată statistic. Întrebarea este câte propoziții ar trebui să aibă corpusul de aur, raportat la numărul total de 96 de relaţii de dependenţă semantică, dat fiind că numărul de 10.000 de propoziții este considerat potrivit pentru un set de aproximativ 40 de etichete.

Întrucât dimensiunile mari implică necesitatea unui număr tot mai mare de opțiuni posibile, corpusul de instruire va avea o creștere mai mult decât proporțională cu setul de tag-uri. Corpusul de formare deja transformat, apoi supravegheat şi completat cu informaţii semantice, de 5.566 de propoziții, este clar insuficient.
O altă modalitate de a rezolva adnotarea semantică în viitor ar fi folosirea Dicționarului de Modele al Verbelor Românești, în care cuvintele să fie corelate cu PWN (Princeton WordNet), astfel încât calculatorul să poată extrage valorile circumstanțiale ale determinanților nominali, de exemplu.

Adnotarea în format semantic al treebank va trebui urmată și diversificată în funcție de stilurile de limbaj, mergând de la bisericesc la narativ, legislativ și folcloric. Cu un număr mic de exemple, rezultatele nu pot fi concludente.

Când vor fi cel puțin 15.000 de propoziții, mecanismul de sortare a structurilor semantice similare ar putea funcționa și am putea vedea care este procentul

de apariții pentru fiecare tip de judecată este reprezentat în cadrul aparițiilor unui verb din corpus semantic.
Programul de sortare a modelelor semantice ar trebui să calculeze procentele în care este așteptată apariția lor în corpusul nostru semantic, deoarece calculează procentele în care sunt așteptate modele sintactice în corpus nostru sintactic nonstandard, iar aceste statistici ar trebui să servească pentru orientarea analizorului statistic.

Următorul pas ar fi să introduceți în treebank mai multe propoziții de tipul celei greșite adnotate automat, pentru a le analiza în mod consecvent, pentru a le introduce în corpusul de aur de antrenament, astfel încât analizatorul să poată extrage corect regula de adnotare sintactică.

Un atribut pe care nu trebuie să-l abandonăm sau să-l neglijăm este și îmbunătățirea instrumentelor pentru OCR-ul literelor chirilice românești vechi. Crearea unui program de recunoaștere optică a caracterelor capabil să recunoască și transformarea textului în litere Chirilice Vechi editabile este un mare pas în direcția accesului digital la patrimoniul cultural textual, care nu poate fi realizat doar prin scanarea cărților vechi.

Citirea manuscriselor ar fi un pas suplimentar în preluarea informațiilor din documente vechi. Este dificil pentru că manuscrisele au un aspect poliglot, conținând date în Română și Slavonă sau în Română și Greacă Medievală.

Această activitate se bazează pe compilarea unui mare inventar de litere din text, precum și asupra creșterii dimensiunilor lexicului inclus, care trebuie totuși pregătit pentru includerea în analizorul morfologic.

Stocarea textelor corectate de lingviști și introducerea lor într-o bază de date cu care este antrenat programul OCR este un alt pas important. Întrucât toate aceste programe, la fel ca multe altele, se bazează pe corpus, vom continua să ne implicăm în dezvoltarea lui și a lor.

Teza mai include o bibliografie a titlurilor menționate în text (cu 248 de titluri), o listă de publicații personale (cu 35 de titluri, dintre care 5 sunt de categoria B, și mai multe de categoria D, totalizând minimum 22 de puncte CNATDCU) si 3 anexe. Urmând lista publicațiilor personale, oferim doar titlurile la care se face referire în acest rezumat.

Contribuții originale

Această teză descrie munca noastră desfășurată pe parcursul celor 7 ani de cercetare doctorală (2015-2022), dar care are rădăcini pe contribuții și publicații dezvoltate și dincolo de perioada specificată. Nucleul acestei lucrări constă în dezvoltarea unei resurse lingvistice computerizate care este esențială pentru modernizarea mașinilor de prelucrare a limbii române la nivelul altor limbi importante din lume: un corpus adnotat de construcții sintactice. Ca atare, principalele noastre contribuții sunt următoarele:

Contribuții teoretice:

- analiză amănunțită a resurselor similare dezvoltate pentru limba română și pentru alte limbi;
- cristalizarea unei dovezi empirice bazate pe exemple că nivelul lexical al limbajului folosit în Republica Moldova și în România este același, cu

diferențe minore la nivel sintactic datorită și mai multor libertăți în ordinea cuvintelor;
- o identificare clară a diferențelor dintre convențiile UAIC de adnotare a băncilor de copaci (reprezentând cel mai bine idiosincraziile limbii române) și convențiile UD;
- o propunere de mărire a bancului de copaci românesc cu un strat semantic; Sunt propuse 14 valori circumstanțiale diferite, strâns legate de echivalentele lor sintactice;
- propuneri de metodologie de cercetare de urmat în vederea atingerii unui nivel matur de dezvoltare a resurselor de instruire și evaluare pentru limba română, acoperind întreaga diversitate de registre, stiluri, diacronicitate și sincronicitate.

Resurse lingvistice:
- Treebank-ul Diacronic de Dependență pentru limba română – UAIC-RoDia, 34.794 de propoziții și 714.377 de token-uri, disponibil în 3 formate;
- UD_Romanian-Nonstandard treebank, cu aproape 16.000 de propoziții dintr-o varietate de stiluri de română veche, contemporană și regională, un treebank aliniat cu alte 87 de treebank-uri din marea familie a Dependențelor Universale;

Instrumente pentru procesarea limbajului:
- convertorul de la formatul UAIC la UD;
- programul TREEOPS XML-transformer (în cooperare)
- un instrument care transformă relațiile sintactice semantic neambiguu în relații semantice;
- un program de statistică a rimelor (în cooperare);
- site-ul PDRoV– RoDia – un portal pentru editarea, căutarea și vizualizarea structurilor de dependență

sintactică și semantică, împreună cu depozitul care găzduiește structuri XML + CONLLU.

Referințe bibliografice

1. Barbu-Mititelu, V.: 2013 *Derivational Semantic Network for Romanian*, Bucharest. National Museum of Romanian Literature Press.
2. Bobicev, Victoria, T. Bumbu, V. Lazu, V. Maxim, D. Istrati 2016 Folk poetry for computers: Moldovan Codri's ballads parsing. Proceedings of the 12th International Conference "Linguistic Resources and Tools for Processing the Romanian Language, p. 39–50.
3. Colhon Mihaela and Radu Simionescu. 2012. Deriving a Statistical Parsing from a Treebank. In Proceedings of International Conference on Web Intelligence, Mining and Semantics WIMS (WIMS). ACM Publishing, Craiova Romania, 34:1–34:8.
4. Cristea, D.: Formalisms and Tools for Description and Natural Language Processing, Iasi, Alexandru Ioan Cuza University Press. (2002).
5. Curteanu Neculai and Alexandru Moruz. 2012. A Procedural DTD Project for Dictionary Entry Parsing Described with Parameterized Grammars. In Proceedings of COGALEX-III – The Third Workshop on Cognitive Aspects of the Lexicon.Bombay, India, 127–136.
6. Diac, Paul, Cătălina Mărănduc, and Mihaela Colhon, Relationships and Sentiment Analysis of

Fictional or Real Characters. 2018. In Proceedings of 19th International Conference on Computational Linguistics and Intelligent Text Processing, 18-24 March 2018, Hanoi, Vietnam.

7. Gîfu. D, The Analysis of Diachronic Variation in Romanian Print Press. Proceedings of the First PhD Symposium on Sustainable Ultrascale Computing Systems, NESSUS PhD (2016), pp. 49--53.
8. Hall, J., and, Nilsson, J., 2007. *CoNLL-X Shared Task: Multi-lingual Dependency Parsing*, MSI report 06060, Växjö University, School of Mathematics and Systems Engineering.
9. Ion, R., Ştefănescu, D.: Unsupervised Word Sense Disambiguation with Lexical Chains and Graph-Based Context Formalization, in Z. Vetulani, *LTC 2009*. pp. 435-443. Heidelberg, Springer, (2011)
10. Irimia, E,, and Barbu-Mititelu, V., *Two resources developed in the project Semantics Driven Syntactic Parser for Romanian,* in Proceedings of ConsILR, Mălini, 27-29 Oct. 2016, p. 69-78. (2016)
11. Mărănduc, Cătălina, Cenel-Augusto Perez, A Resource for the Written Romanian: the UAIC Dependency Treebank, in Proceedings of ConsILR, Mălini, 27-29 Oct. pp. 79-90. (2016)
12. Moruz, Al., Dezvoltarea unui adnotator FDG (Functional Dependency Grammar) pentru limba română, lucrare de disertaţie, Facultatea de Informatică, Universitatea „Alexandru Ioan Cuza" Iaşi. (2008)
13. Nivre. Joakim 2003. An efficient algorithm for projective dependency parsing. In Proceedings of

the 8th InternationalWorkshop on Parsing Technologies (IWPT). ACL Publishing, Nancy France, 149–160.

14. Trandabăţ, D.: Natural Language Processing Using Semantic Frames, PHD Thesis, Computer Science Faculty, Alexandru Ioan Cuza University of Iasi, Romania.
15. Tufiș Dan și Dragomirescu Liviu Tiered Tagging Revisited In Proc. of the Fourth Intl. Conf. on Language Resources and Evaluation, LREC'2004, ELRA, Paris, p. 39-42. (2004)