UN CORPUS DE TEXTE PENTRU LIMBA ROMÂNĂ ADNOTAT SINTACTIC SUB FORMĂ DE ARBORI

CENEL-AUGUSTO PEREZ
Universitatea „Alexandru Ioan Cuza” din Iaşi

În lucrarea de față vom inventaria rezultatele achiziționării corpusului de sintaxă a limbii române, punând accent atât pe problemele întâmpinate pe parcursul achiziționării acestui corpus (se vor menționa şi câțiva paşi ai creării corpusului), cât şi pe soluționările acestor probleme.

1. Introducere

Între resursele utilizate pentru studierea sintaxei limbilor naturale, o componentă importantă sunt treebank-urile, termen mai recent în NLP (Natural Language Processing – Prelucrarea Limbajului Natural), un subtip de corpus adnotat.

Un treebank este un corpus de texte în care fiecare propoziție are asociată o structură sintactică arborescentă (ceea ce explică denumirea de „treebank”). Structurile sintactice constau în unități lexicale legate prin relații binare de dependență, asimetrice, între un regent și un dependent. Treebank-urile sunt de cele mai multe ori create pe baza unor corpusuri care au fost deja adnotate prin POS-tagging (părți de vorbire și caracteristici flexionare). Ulterior, treebank-urilor li se pot atribui alte informații lingvistice, semantice, pragmatice.

Treebank-uri există pentru limbi precum: chineza, ceha, engleza, franceza, germana, italiana, japoneza, poloneza, portugheza, spaniola, turca etc. Cele mai reprezentative treebank-uri actuale sunt pentru limba engleză (Penn Treebank – construit la Universitatea din Pennsylvania, Philadelphia[1]) și pentru limba cehă (Prague Dependency[2] Treebank – construit la Universitatea Charles din Praga). O astfel de resursă lipseşte deocamdată pentru limba română, de aceea se impune crearea unui astfel de corpus.

În treebank-uri se notează fenomene lingvistice specifice limbilor, pe seturi de exemple suficient de numeroase pentru ca fiecare fenomen să se repete atât de mult, încât parserul să poată extrage automat o regularitate. Aceste corpusuri adnotate sunt utilizate în două moduri. În primul rând, din ele pot fi extrase reguli, ce se pot transpune în programe, cu care să se realizeze o analiză sintactică automată. În al doilea rând, treebank-urile sunt folosite pentru verificarea parserelor astfel construite.

Treebank-urile au multe alte aplicații, de la testarea teoriilor lingvistice până la construirea automată de gramatici. Cu ajutorul treebank-urilor, lingviştii pot căuta exemple sau contraexemple pentru teoria sau ipoteza pe care o susțin.

[1] http://www.cis.upenn.edu/~treebank/
[2] http://ufal.mff.cuni.cz/pdt/

Amintim câteva aplicații care fac uz de corpusuri ca cele prezentate în lucrarea noastră: clasificarea textelor, dezambiguizarea sensurilor cuvintelor, alinierea textelor multilingve, sisteme de întrebare-răspuns, rezumarea automată a textelor, sisteme de recunoaştere a inferenţelor textuale etc.

Pentru construirea corpusului treebank se vor avea în vedere următorii paşi: 1. achiziționarea surselor lexicale (colecții mari de texte); 2. dezvoltarea metodologiei de adnotare a corpusului (instrucțiuni care să permită o adnotare consistentă cu o teorie lingvistică); 3. marcarea automată a informațiilor de natură morfologică / lingvistică asupra corpusului (pos-tagging, tokenizare, segmentare etc.); 4. utilizarea unui instrument interactiv cu care se va realiza adnotarea propriu-zisă la sintaxă; 5. augmentarea acestei etichetări cu alte notații pentru descrierea unor fenomene de interes în traducerea automată.

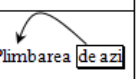
2. Procesul de adnotare

2.1. *Principii de analiză a structurilor sintactice*

Pe măsură ce am adnotat propoziții şi fraze, am elaborat o listă cu posibilii regenți, posibilele cuvinte subordonate, relații etc. Această listă devinte tot mai lungă şi mai cuprinzătoare pe măsură ce noi exemple sunt descoperite pe parcursul procesului de adnotare sintactică.

În figura de mai jos (primul rând) avem un substantiv ca regent. Există o propoziție cu un grup nominal în care substantivul este precedat de o prepoziție urmată de un adverb. În arborele de dependență, pe săgeata dintre substantiv (*plimbarea*) şi prepoziție (*de*), vom avea notată, aşadar, relația de atribut adverbial (*a.adv.*), ca în exemplul: „Plimbarea de azi", deoarece „de azi" este atribut adverbial pentru substantivul „plimbarea".

Pentru continuarea analizei acestui grup nominal, trebuie să căutăm în listă situația în care regentul este o prepoziție (*de*) şi cuvântul subordonat este adverb (*azi*) şi să aplicăm aceeaşi metodologie pentru a completa arborele.

REGENT	CUVÂNT SUBORDONAT	URMAT DE	RELAȚIE	ABREVIERE	EXEMPLU
Substantiv	prepoziție	adverb	atribut adverbial	a. adv.	
Substantiv	prepoziție	verb supin	atribut verbal	a. vb.	Maşina de spălat
Substantiv	prepoziție	substantiv	atribut substantival	a. subst.	Praf de puşcă

Tabelul 1. Fragment din lista cu relații de dependență

2.2. *Dificultăți de interpretare a structurii lingvistice*

Fiind vorba de eşantioane de limbaj natural, textele adnotate pot să conțină diverse situații incerte, neclare, care se pretează la mai multe interpretări lingvistice. În aceste cazuri, din punct de vedere lingvistic este indiferent pentru care dintre soluții optăm, deci o vom alege pe cea care se pretează mai bine formalismului gramaticilor de dependență care stă la baza treebank-ului nostru.

Nerecunoaşterea unor locuțiuni verbale, adverbiale, conjuncționale etc., după procesul de pos-tagging, îngreunează procesul de adnotare. Celelalte grupuri de cuvinte, care sunt separabile şi pot să apară în vorbire şi cu alt sens, numite expresii sau sintagme, se pretează unei analize interne. Tot în acest exemplu putem observa şi un caz în care am lăsat cuvintele

din expresia *umăr la umăr* în forma recunoscută de procesarea texului RACAI, şi anume trei cuvinte separate. Analiza lor ca un întreg ar fi preferabilă. Dar analiza relațiilor între componentele expresiei nu duce la interpretări incoerente nici logic, nici față de sistemul nostru de convenții de adnotare.

Expresia *umăr la umăr* este un complement circumstanțial de mod, prin urmare atribuim această etichetă relației dintre verb şi expresie. Apoi, în interiorul acestei expresii putem considera că primul cuvânt este regent, având drept centru verbul *mergă. La umăr* se comportă ca un atribut substantival (cu sens local) față de primul *umăr*. Prin urmare vom nota relația de atribut substantival între *umăr* şi *la umăr*. Între *la* şi al doilea *umăr* avem relația prepozițională, deoarece *la* este o propoziție care impune forma celui de-al doilea substantiv.

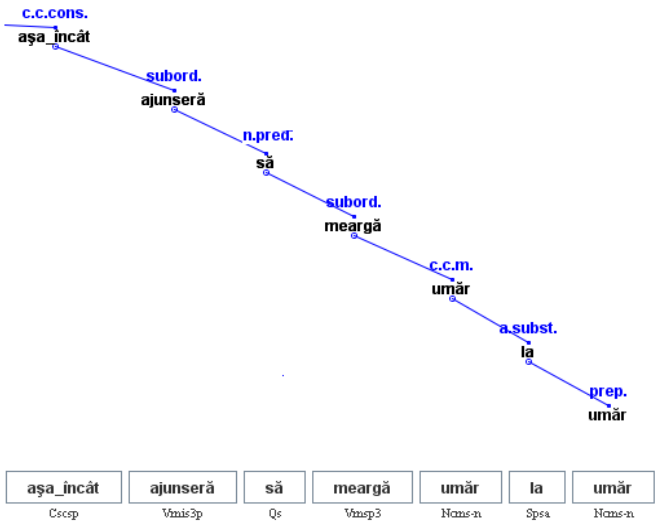


Figura 1. Exemplu de locuțiune conjuncțională şi adverbială

Prepozițiile suportă un regim special, deoarece ele devin, în arborii de dependență, regenți pentru cuvintele pe care le subordonează. După Gramatica Academiei, ediția a treia, prepoziția, în calitate de conector, se încadrează obligatoriu într-o structură ternară, prezența acesteia fiind condiționată de coocurența cu doi termeni lexicali autonomi, ce se află într-o relație de dependență (*zi* de *iarnă, fuge* la *mama*). În această structură, prepoziția devine *centru* pentru termenul pe care îl precedă, impunându-i anumite restricții gramaticale. Această relație strânsă este marcată şi de topica fixă a celor două componente: prepoziția se plasează pe primul loc (*tablă* de *şah*, nu *tablă şah* de).

În adnotare, s-a procedat conform acestei caracteristici şi s-a situat prepoziția între cele două elemente (determinantul şi cuvântul determinat). Ea este deci regent pentru cuvântul pe care îl introduce, căruia îi impune anumite restricții; aici, impune cazul acuzativ şi interzice ocurența cu articol hotărât sau nehotărât, permisă în alte limbi la substantivele precedate de prepoziție.

Dar în limbajul natural există cazuri în care prepoziția nu se conformează teoriilor şi convențiilor noastre şi nu se situează între regent şi cuvântul subordonat.

În propoziția: *Orice referire cât de cât clară la el ar fi fost fatal de primejdioasă.* cuvântul *primejdioasă* stabileşte relația *n.pred.*, fiind subordonat de verbul copulativ (*ar fi*) *fost*, însă problema este cu construcția prepozițională *fatal de*. Topica prin care prepoziția nu

este situată între regent şi determinant, ci urmează după determinantul exprimat prin substantiv, adverb sau numeral este destul de frecventă în corpusul adnotat şi a fost rezolvată în mod unitar prin stabilirea unei convenţii foarte economice, deşi puţin conformă cu topica firească a limbii. *Primejdioasă* trebuie să fie regentul grupului adjectival *fatal de primejdioasă*. Am considerat că peste tot prepoziţia trebuie să fie regentul determinantului, aflându-se într-o structură ternară (GALR I, 2008: 607), deci am legat de *primejdioasă* mai întâi prepoziţia *de*, iar de aceasta am legat adverbul *fatal*. Între *primejdioasă* şi *de* avem relaţia superlativă, deoarece *fatal de* ajută la formarea gradului superlativ / absolut al adjectivului *primejdioasă*, iar între *de* şi *fatal* vom nota relaţia de prepoziţie.

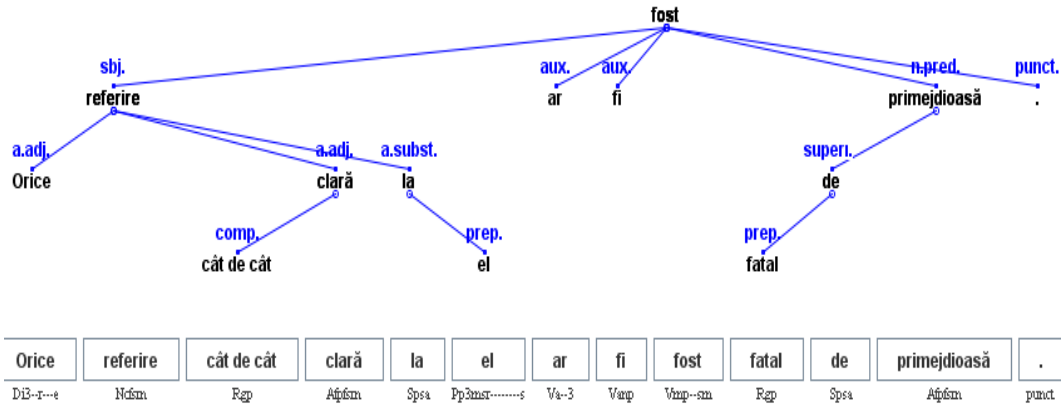


Figura 2. Exemplu de adnotare a relaţiei superlative marcate prin adverb cu prepoziţie

Tot în aceeaşi situaţie se află şi exemplul următor unde avem un numeral cu o prepoziţie. În figura 3 avem un caz simplu de dependenţă între un numeral şi un substantiv, însă în figura 4, din cauza prepoziţiei, a trebuit să luam aceeaşi măsură ca în exemplul anterior, deoarece peste tot prepoziţia a fost interpretată ca regentul determinantului.

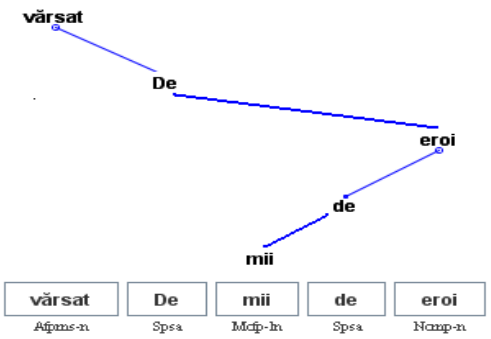


Figura 3. Grup nominal cu numeral şi prepoziţie

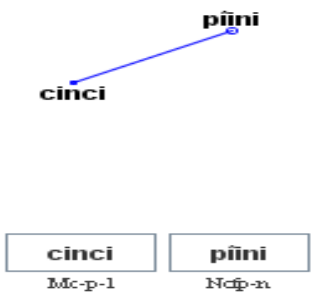


Figura 4. Substantiv şi numeral

În exemplul de mai jos, *dreptunghiulară* este un adjectiv, aşa cum ne arată flexiunea sa, el se acordă cu substantivul *zarea* şi nu poate stabili relaţia de *c.c.m.*, dar nici pe cea de *a.adj.*, nefiind în vecinătatea sintactică a substantivului pe care îl determină.

Potrivit gramaticilor transformaţionale, structura de suprafaţă este aici o transformare care abreviază structura profundă *Zarea este dreptunghiulară şi zarea se albeşte*. eliminând verbul copulativ lipsit de sens lexical, care este, cum arată şi numele „copulativ", doar un

mijloc de a lega între ele două noțiuni. Întrucât gramaticile de dependență nu acceptă reguli de transformare, vom releva doar faptul că *n.pred.* și *el.pred.* sunt modificatori cu sens modal ai verbului de care depind și, în același timp, modificatori ai unui nominal din text, căruia îi atribuie o caracteristică, la fel ca un atribut.

Deoarece formalismul gramaticilor de dependență nu ne permite subordonarea unui cuvânt de două noduri regente, am optat pentru subordonarea el.pred. de verb, așa cum și n.pred. este subordonat verbului copulativ.

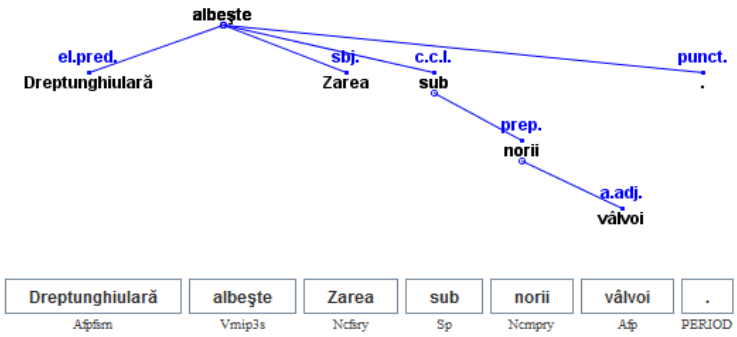


Figura 5. Adnotarea elemenului predicativ suplimentar

În limbajul natural există numeroase cazuri în care nu se poate stabili o coordonare între două blocuri ale frazei, fiindcă unul dintre ele este o structură exogenă importată din enunțarea altui emitent. Din această cauză am introdus conceptul de relație de incidență, pe care îl vom adnota ca și pe relația de coordonare, dar vom utiliza o altă etichetă, în conformitate cu această realitate discursivă.

Apariția propoziției în relație de incidență este semnalată de existența în structură a unui verb dicendi. Există cazuri în care acest verb aparține corpului principal al comunicării și introduce comunicarea altui emitent, ca în exemplul ce urmează.

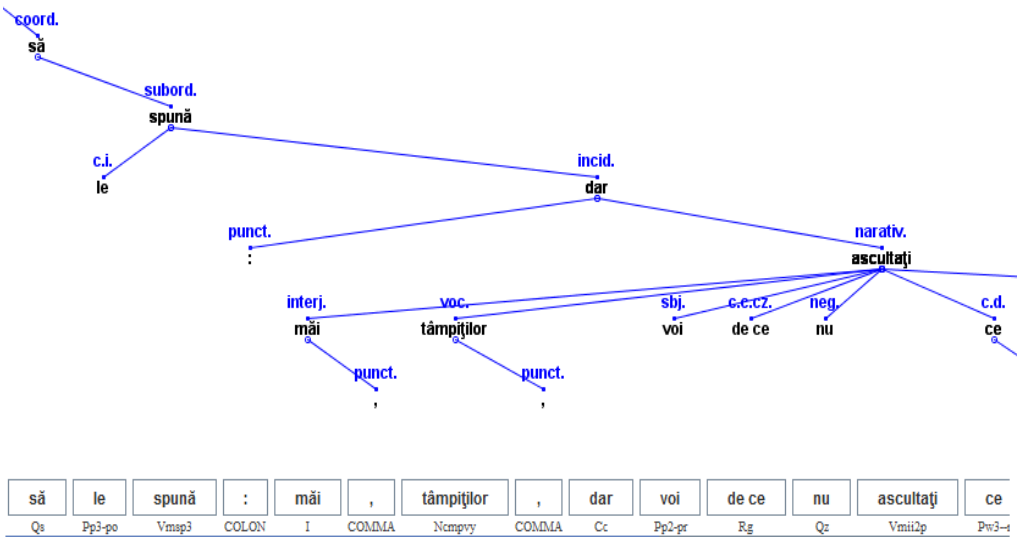


Figura 6. Secțiune dintr-un arbore cu structuri exogene

Faptul că anumite secvențe în limbajul natural nu se subordonează structurii reprezintă o abatere de la formalismul gramaticilor de dependență. În exemplul de mai sus, apare atât un bloc propozițional incident, cât şi substantivul în cazul vocativ şi interjecția, considerate de gramatica clasică drept lipsite de funcție sintactică. Ele sunt izolate prin virgule de restul structurii. Prin relațiile *incid.*, *voc.*, *interj.* stabilim o compatibilitate între această realitate lingvistică şi formalismul adoptat de noi.

De fapt, în aceste cazuri lărgim sensul relației de dependență, ea nu mai reperezintă subordonare, ci o relație de succesiune în care consecventul răspunde expectației pe care o generează antecedentul.

În alte cazuri, cum este acest exemplu, verbul dicendi nu introduce structura exogenă, ci face parte din ea, fiind chiar centrul ei. Esențială este însă existența, în ambele cazuri, a unor emitenți diferiți pentru două blocuri ale structurii sintactice.

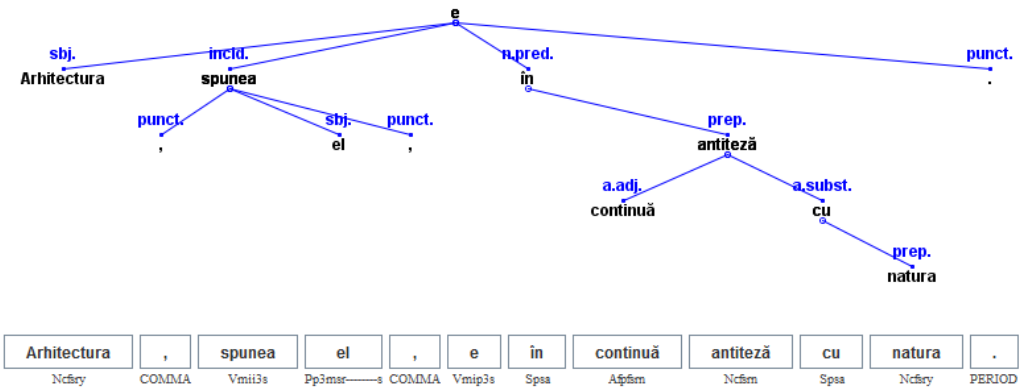


Figura 7. Verb dicendi ca head al construcției incidente

În exemplul de mai jos, *dar* este conjuncție coordonatoare adversativă. Astfel de conjuncții coordonatoare care se situează la începutul unei fraze sau propoziții sunt supuse unui tratament special în structura arborilor noştri şi, pentru că ele coordonează două elemente de discurs ce se constituie în fraze diferite, am hotărât să fie considerate conectori pragmatici cu rol narativ şi le-am atribuit relația narativă. Un astfel de conector, fiindcă funcția lui este de a stabili relații între fraze în cadrul textului, este considerat regent al verbului principal, se situează mai sus decât el în arbore şi deci preia funcția de rădăcină a arborelui, potrivit primei axiome a gramaticilor de dependență, tocmai pentru că funcția de coordonare a acestei conjuncții se situează la un nivel de analiză superior frazei adnotate aici.

Pentru uşurarea şi coerența interpretării sintactice, am mai introdus un simbol, *corel.*, care marchează faptul că în structură există simetrii între blocuri sintactice. Această etichetă este atribuită unor negații, adverbe sau alte cuvinte-unelte gramaticale, care se subordonează centrului celor două blocuri sintactice între care pun în evidență o simetrie.

Una dintre primele operații pe care am stabilit să le efectuăm, în analiza sintactică, este aceea de găsire a verbelor predicative care reprezintă head-uri pentru diferite ramuri principale ale arborelui. În limbajul curent însă, există numeroase cazuri când nu verbul este predicativ, ci adverbul sau interjecția (Ex.: *Hai să mergem*). O mențiune specială merită adverbele predicative impersonale. În construcția *Sigur că ai dreptate*, observăm conjuncția *că*, marcator al subordonării, prin urmare elementul regent este o propoziție principală, iar aceasta este compusă doar din adverbul *sigur*. Cum într-o construcție neabreviată, el ar trebui

să fie nume predicativ (*Este sigur*), rezultă că propoziția subordonată de acesta stabileşte cu el relația *sbj*.

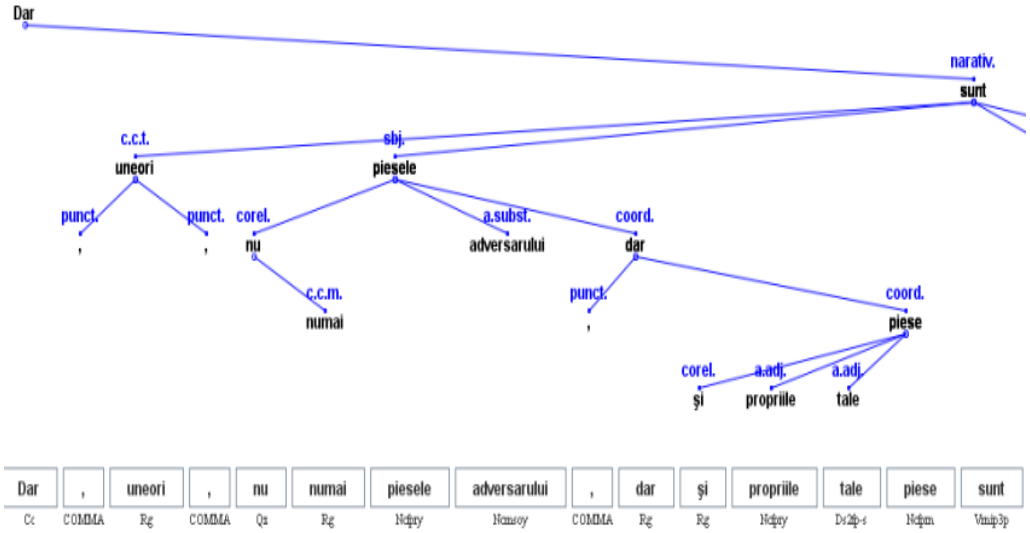


Figura 8. Exemplu de adnotare a relației narative şi relației de corelație

Dacă aici pare să fie vorba de un verb copulativ NUL, în alte cazuri verbul copulativ şi numele predicativ se sudează, devenind neanalizabile, sub forma unor adverbe predicative ca *fireşte, desigur*. Și acestea vor avea un determinant în relație de *sbj*. față de ele. În aceeaşi situație se află locuțiunea adverbială predicativă din exemplul de mai jos.

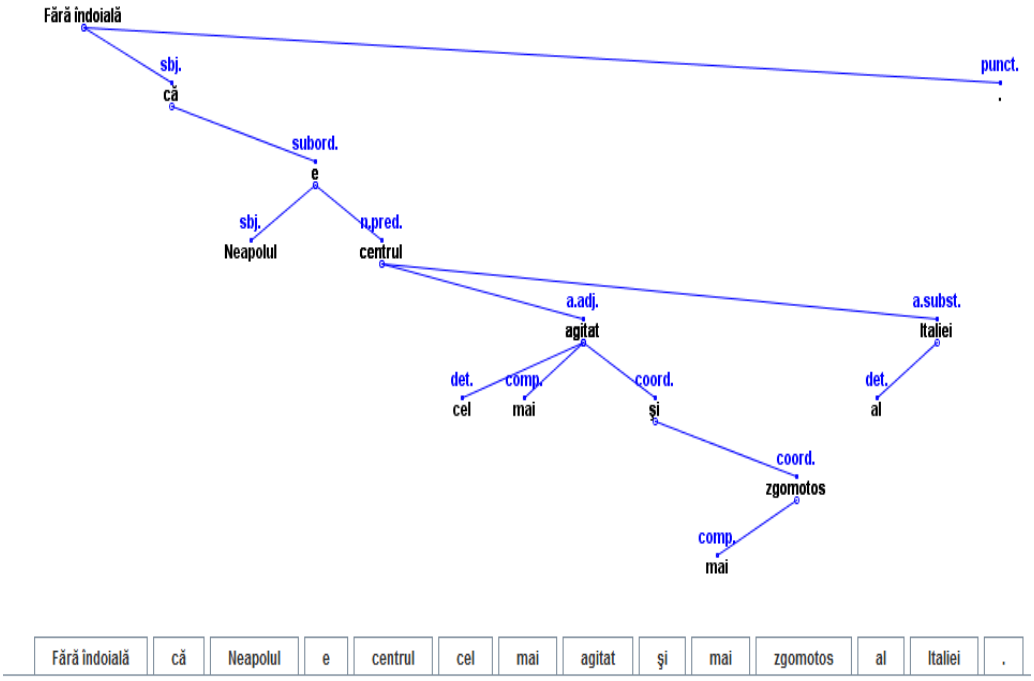


Figura 9. D-arbore având drept rădăcină locuțiunea adverbială predicativă

O altă situație problematică este poziționarea adverbului *cam*, care se poate subordona unui adjectiv (*cam veche*) dar şi unui substantiv (*cam la sfârşitul lui august*). Avem, de exemplu, structura „… *ajunsese cam în acelaşi loc*”. De cele mai multe ori, adverbul *cam* determină un adjectiv, prin urmare, în ciuda topicii propoziției, ar fi trebuit să legăm adverbul de adjectivul demonstrativ *acelaşi*, însă, în cazul nostru, *cam* nuanțează întreg complementul circumstanțial de loc (*în acelaşi loc*), *cam* fiind un nuanțator care poate determina orice tip de construcție, inclusiv o propoziție („am cam terminat”).

Vom decide, prin convenție, pentru a adnota unitar toate situațiile, ca adverbul nuanțator *cam* să fie subordonat întotdeauna centrului construcției pe care o nuanțează, chiar dacă acesta este o prepoziție.

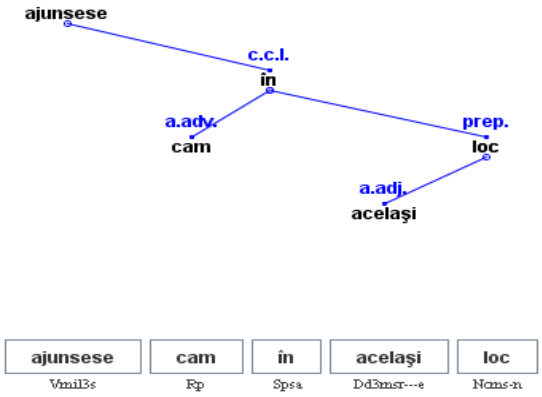


Figura 10. Exemplu de arbore cu adverbul *cam*.

3. Concluzii

Convențiile adoptate pot fi modificate în funcție de noi situații ivite sau pentru a ne pune de acord cu alte treebank-uri spre a putea efectua studii comparative între limbi.

Internaționalizarea infomației a făcut necesară găsirea unor modalități de a stabili automat corespondențe între limbile naturale. Corespondențele nu se pot stabili decât între limbi cu un grad înalt de informatizare; o limbă are în față două alternative: să se informatizeze sau să dispară ca limbă de știință și cultură.

Informatizarea limbii presupune existența unor dicționare, gramatici, corpusuri de texte diverse, adnotate astfel încât calculatorul să le poată stoca, aliniate cu cele ale limbilor de circulație, baze de date între care să se poată stabili (automat) corespondențe. Limba română nu dispune de corpusuri adnotate de mari dimensiuni, care sunt necesare pentru creşterea gradului ei de informatizare.

În cadrul cercetării doctorale, am pus bazele unui corpus de texte adnotat sintactic, actualmente format din 4467 de propoziții și fraze, cu ajutorul căruia am antrenat un parser sintactic pentru limba română. Primele 2000 de fraze au fost adnotate manual, celelalte au fost adnotate automat şi corectate manual. După corectare, textele noi sunt introduse în corpusul de antrenare a parserului, astfel încât acesta face tot mai puține greşeli de adnotare și viteza de corectare a frazelor adnotate automat crește. Activitatea de mărire a corpusului va fi continuată.

Ca și corpusul limbii engleze, corpusul nostru are la bază modelul gramaticilor de dependență. Trebuie făcută o distincție între modelele sintactice teoretice elaborate de lingviști pornind de la modelele lui N. Chomsky și modelele de gramatici elaborate de

informaticieni pornind de la același cercetător. Dacă primele au ca scop descrierea limbajului natural și ilustrarea unor teorii de filosofie a limbajului, celelalte au ca obiectiv elaborarea unor modele pentru computer care să se apropie cât mai bine de specificul limbajului natural și să-l poată procesa.

Am pornit de la cele două axiome ale gramaticilor de dependență, și anume că nu există decât o unică rădăcină pentru un arbore și că toate celelalte elemente au câte un singur determinant. Teoria lui L. Tesnière, pe care am luat-o ca punct de plecare, are numeroase elemente comune cu gramatica lui J. Fillmore. A trebuit să stabilim numeroase convenții de adnotare pentru a adapta modelul la specificul limbajului natural și la specificul limbii române.

Corpusul poate fi utilizat pentru efectuarea unor statistici: cea mai mare frecvență în corpus o are relația prepozițională (14,48%, dintr-un număr de peste 50 000 de relații găsite în cele 4 467 de texte), urmată de relația de atribut adjectival (9,32%), relația de punctuație (10,71%), de atribut substantival (7,93%) și de relația de complement direct (7,04%).

Se pot compara texte aparținând unor diferite stiluri ale limbii sau se poate efectua o comparație între corpusul românesc și corpusul altor limbi. De exemlu, față de limba engleză, putem demonstra statistic o prezență mai scăzută a subiectului exprimat.

Corpusurile dezvoltate în cadrul acestei teme vor fi utilizate pentru perfecționarea, prin antrenare automată, şi testarea câtorva tipuri de instrumente de prelucrare a limbajului natural, cum ar fi analizoare sintactice (parsere) şi traducătoare automate. Ele constituie, totodată, surse importante de date pentru testarea teoriilor şi ipotezelor lingvistice.

În zona psiholingvisticii, corpusurile pot fi utilizate în evaluarea predicțiilor asupra frecvenței anumitor tipuri de construcții sintactice.

Odată create, treebank-urile pot sta la baza dezvoltării altor tipuri de adnotări, de exemplu, la nivel de discurs, semantic ori pragmatic.

BIBLIOGRAFIE

Călăcean, M. & Nivre, J., 2008, *Data-driven Dependency Parsing for Romanian*, Uppsala University.

Chomsky, Noam, 1975, *Aspects of the Theory of Syntax*, The M.I.T. Press, Massachusetts Institute of Technology Cambridge, Massachusetts, Tenth printing, May, 1975.

Cornilescu, Alexandra, 1995, *Concept of Modern Grammar. A Generative Grammar Perspective*, București, Editura Universității din București.

Cristea, Dan, 2002, *Formalisme şi instrumente de descriere şi prelucrare ale limbajului natural*, Iaşi, Editura Universității „Alexandru Ioan Cuza”.

Fillmore, Charles J., 1968, *The Case for Case,* in *Universals in Linguistic Theory*, edited by E. Bach & R, Harms.

Hajic, Jan, 1998, *Building a Syntactically Annotated Corpus: The Prague Dependency Treebank*, In *Issues of Valency and Meaning*, p. 106-132.

Hristea, Florentina & Popescu, Marius, 2003, „A Dependency Grammar Approach to Syntactic Analysis with Special Reference to Romanian”, în, Hristea, F. & Popescu, M. (eds.) *Building Awareness in Language Technology*, București, Editura Universității din București, p. 9-34.

Gramatica limbii române, 2005, 2008, coord. Valeria Guțu-Romalo, Vol. I: *Cuvântul.* vol. II: *Enunțul.* București, Editura Academiei Române.

Ionescu, Emil, 2007, *Gramatici formale cu referire specială la HPSG*, București, Editura Universității din București.

Pană Dindelegan, Gabriela, 1974, *Sintaxa transformațională a grupului verbal în limba română*, București.
Penn Treebank http://www.cis.upenn.edu/~treebank/
Prague Dependency Treebank http://ufal.mff.cuni.cz/pdt2.0/
Robinson, J. J., „Dependency Structures and Transformational Rules", *Language* 46, 1970, p. 259-285.
Seretan, V., Wehrli, E., Nerima, L. & Soare, G., 2010, *Fips Romanian: Towards a Romanian Version of the Fips Syntactic Parser*, in *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*, European Language Resources Association (ELRA).
Tesnière, L., 1959, *Elements of structural syntax*, Paris, Klincksieck.
Vasiliu, Emanuel, Golopenția-Eretescu, Sanda, 1972, *The Transformational Syntax of Romanian*, Haga, Mouton.

A SYNTACTIC ANNOTATED CORPUS – ROMANIAN TREEBANK

(*Abstract*)

In the present paper we will inventory the outcomes of the achievement of the Romanian language syntactic corpus, stressing on the problems met during the acquisition of this corpus (we will also mention some steps in the creation of the corpus) and on the solutions of these problems.