ACADEMIA ROMÂNĂ

SECŢIA DE ŞTIINŢA ŞI TEHNOLOGIA INFORMAŢIEI

INSTITUTUL DE CERCETĂRI PENTRU INTELIGENŢĂ ARTIFICIALĂ „MIHAI DRĂGĂNESCU”

Programul de cercetare: **Corpus computaţional de referinţă pentru limba română contemporană**

Subprogramul de cercetare Nr. Program prioritar

Faza I **Dezvoltarea cantitativă și funcțională a corpusului și mediului de exploatare a lui**

Coordonator subprogram: Acad. Dan Tufiș

Raport de Cercetare

BUCUREŞTI, iunie 2015

Colectivul de cercetare

Acad. Dan Tufiş – C.S.I, coordonator subprogram

C.S.III Dr. lingv. Verginica Mititelu

C.S.III Dr. inf. Elena Irimia

C.S. III Dr. ing. Ștefan Daniel Dumitrescu

C.S. III Dr. ing. Tiberiu Boroș

C.S. Dr. mat. Corina Forăscu (1/2 normă)

C.S. lingv. Cătălin Mihăilă (1/2 normă)

AsC Eric Marian Curea

Cuprins

Precizare

Dată fiind natura proiectului prioritar la care lucrăm, cu etape care se întind pe (aproape) toată durata sa, rapoartele noastre de cercetare vor reflecta această rutină în structurarea și conținutul lor. Astfel, pe structura raportului de cercetare anterior vom face și raportul de față (și probabil și altele viitoare), evidențiind clar activitatea din semestrul surprins de raport.

Introducere

Activitățile desfășurate în vederea creării corpusului computațional de limbă română contemporană sunt: colectarea corpusului, crearea metadatelor pentru textele colectate, alinierea textelor orale cu transcrierile lor, pregătirea textelor pentru prelucrarea automată, prelucrarea și adnotarea lor, corectura semi-automată a unui procent din textele adnotate, dezvoltarea de resurse și instrumente pentru niveluri de adnotare suplimentare. În acest semestru, activitatea noastră a vizat: colectarea de texte, crearea de metadate pentru textele colectate, dezvoltarea unei interfețe utile pentru transcrierea textelor orale primite fără transcriere, pregătirea textelor pentru prelucrarea automată, prelucrarea și adnotarea lor, dezvoltarea unui treebank adnotat cu relații din Universal Dependency Grammar, din care să se poată apoi antrena un parser sintactic pentru limba română.

Prezentăm în continuare activitățile și subactivitățile desfășurate, metodele abordate și instrumentele cu care am lucrat.

1. Colectarea corpusului

Această activitate este permanentă pe durata desfășurării proiectului și are ca scopuri:

- preluarea fișierelor cu texte de la furnizori;
- încărcarea acestora pe serverele noastre.

1.1. Activități

Următoarele activități se desfășoară în această etapă:

a) *contactarea posibililor furnizori* de texte și prezentarea proiectului nostru în vederea încheierii unui protocol de colaborare în termenii căruia să se desfășoare următoarele etape;

b) *selectarea textelor* care să intre în corpus de la furnizorii cu care am semnat protocol de colaborare;

c) *preluarea textelor* pe care furnizorul ni le poate pune la dispoziție; sub formă de fișiere de tip txt, doc, rtf sau pdf, în cazul textelor scrise, respectiv mp3 în cazul textelor orale;

d) încărcarea fișierelor pe serverele de lucru.

a) În semestrul actual, ICIA a încheiat două noi protocoale de colaborare[1] cu doi bloggeri: Irina Șubredu (http://irina.subredu.name) și Teodora Forăscu (https://travelearner.wordpress.com). Atfel, lista actuală a tuturor celor cu care avem stabilit un parteneriat pentru colaborare este următoarea:

Nr. crt.	**Partener**	**Data până la care este valabil contractul**
1	Humanitas	31 dec. 2015
2	Editura Academiei	31 dec. 2015
3	România literară	31 dec. 2015
4	Revista Colegiului Național Unirea din Focșani	31 dec. 2015
5	Uniunea Compozitorilor și Muzicologilor din România	31 dec. 2016
6	Editura Universității din București	31 dec. 2016
7	DCNEWS	31 dec. 2017
8	Editura Economică	31 dec. 2017
9	Polirom	31 dec. 2017
10	Rador – Radio România	31 aug. 2015
11	Editura Simetria	31 dec. 2017
12	Simona Tache	-
13	Dragoș Bucurenci	-
14	Irina Șubredu	-
15	Teodora Forăscu	-

Tabelul 1. Lista actuală a partenerilor.

b) *Selectarea textelor* care să ne fie puse la dispoziție s-a făcut în mod diferit.

De la Editurile Humanitas și Polirom am făcut noi selecția titlurilor de cărți de pe site-urile lor. Aceste selecții se regăsesc în Anexa 1, respectiv 2 la prezentul raport. Criteriul de alegere a titlurilor l-a constituit încercarea de completare a domeniilor și subdomeniilor deficitare la acest moment în corpusul nostru, conform statisticii din raportul anterior.

Reprezentanții Editurii Academiei Române au ales să facă ei înșiși selecția textelor.

De pe cele două bloguri (http://irina.subredu.name și https://travelearner.wordpress.com) au fost descărcate automat, la zi (mai 2015),

[1] Pentru protocoalele încheiate de IIT-Iași, a se vedea raportul coleglor din Iași la acest proiect

postările și s-au selectat, conform regulii stabilite în etapa anterioară, cele care au peste 500 de cuvinte.

c) *Preluarea textelor* de la furnizori. Editura Humanitas ne-a pus la dispoziție fișierele .doc ale cărților din Anexa 1. Acest format permite preluarea (automată) cu rezultate foarte bune a textului propriu-zis pentru a crea fișierul .txt corespunzător.

În cazul blogurilor, am ales postările de minimum 500 de cuvinte scrise de autorul blogului, nu preluate din cărți, din mailuri primite de acesta sau din comentarii ale cititorilor.

1.2. Date cantitative

Prezentăm aici contribuția partenerilor până în acest moment:

Partener	Număr de texte	Spațiu pe disc
Humanitas	72	1086 MB
România literară	15838	4.7 MB
Revista Colegiului Național Unirea din Focșani	78	1.06 GB
Uniunea Compozitorilor și Muzicologilor din România	10 + versiuni electronice disponibile pe site	272 MB +
DCNEWS	Tot ce era pe site în august 2014	n.a.
Editura Economică	57	136 MB
Polirom	27	178 MB
Rador – Radio România[2]	1 h de înregistrări/zi lucrătoare începând cu 1 septembrie 2014	9 GB
Blog Simona Tache	385	1 MB
Blog Dragoș Bucurenci	384	1 MB
Blog Irina Șubredu	43	102KB
Blog Teodora Forăscu	51	223 KB

Tabelul 2. Numărul de texte per partener.

2. Metadate

2.1. Instrumente de creare a metadatelor pentru texte scrise

Pentru textele primite până acum am creat metadate în două moduri:

- manual: cu ajutorul unor interfețe, utilizatorul a introdus datele despre fiecare text. Inițial, am folosit Arbil (https://tla.mpi.nl/tools/tla-tools/arbil/) pentru

[2] Întrucât Rador ne trimite zilnic înregistrări și transcrieri, datele se modifică de la o zi la alta. AM preferat să notăm contribuția zilnică a acestui furnizor de texte.

acest lucru, apoi am utilizat platforma creată de colegii de la Iași (Moruz și Scutelnicu, 2014) .

- automat: această modalitate de creare a metadatelor este folosită în cazul textelor descărcate automat de pe Internet, folosind informațiile puse la dispoziție de site-ul de pe care sunt preluate. Generatorul de metadate pentru textele descărcate automat a fost realizat la ICIA.

Indiferent de modul de creare, structura metadatelor este aceeași pentru toate textele introduse în corpus.

2.2. Date cantitative

Mod de creare a metadatelor	Număr de fișiere de metadate create astfel
Arbil	1310
Platforma dezvoltată la IIT	2782
automat	76962
TOTAL fișiere cu metadate	81054

Tabelul 3. Numărul de fișiere cu metadate.

2.3. Statistici despre tipurile de documente, despre stilurile funcționale, domeniile și subdomeniile reprezentate până acum în corpus

Odată create metadatele pentru texte, putem observa gradul de acoperire a fiecăruia dintre tipurile de texte pe care ne-am propus să le includem în corpus. Prezentăm mai jos date despre tipurile de texte pe care le-am inclus în corpus până acum.

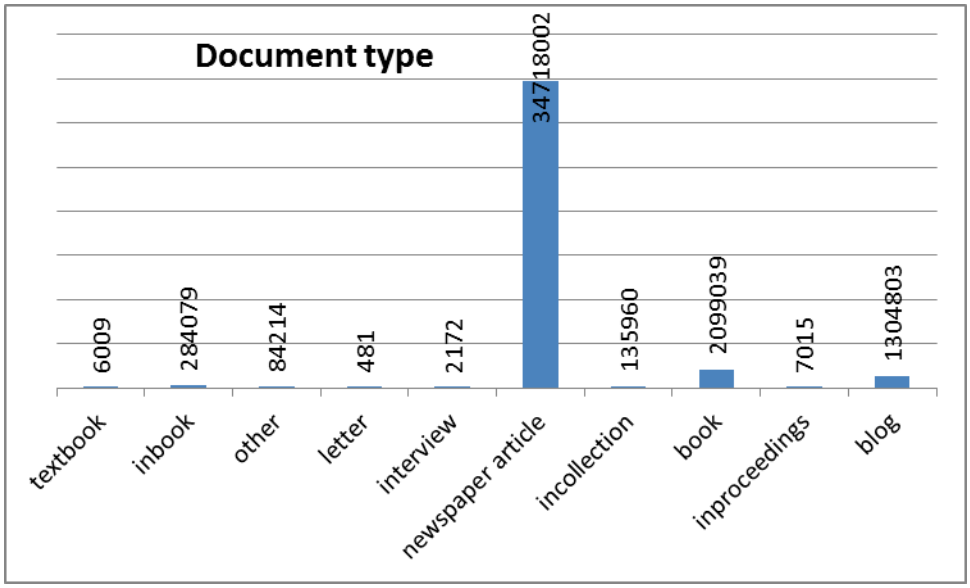


Figura 1. Date cantitative despre tipurile de documente.

Articolele de ziar continuă să fie tipul de text cel mai bine reprezentat până acum. Urmează, la o distanță considerabilă, cărțile. Având în vedere faptul că ne-am propus

ca 60% din totalul de texte să provină din cărți, va trebui ca în etapele următoare să ne concentrăm asupra colectării mai multora.

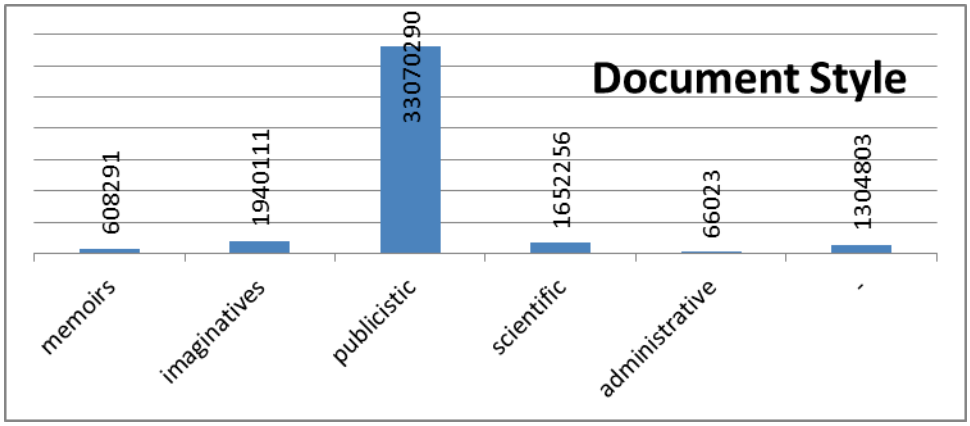


Figura 2. Date cantitative despre tipurile de stilurile funcționale.

Primul lucru ce trebuie remarcat din această figură este acoperirea tuturor stilurilor funcționale[3], desigur, în proporții diferite.

Corelat cu datele din figura 1, remarcăm preponderența stilului jurnalistic. Se pare că dintre cărți, aproximativ două treimi au fost științifice, iar restul beletristice.

Textele cărora nu le-a fost alocat un stil funcțional sunt postări de pe bloguri.

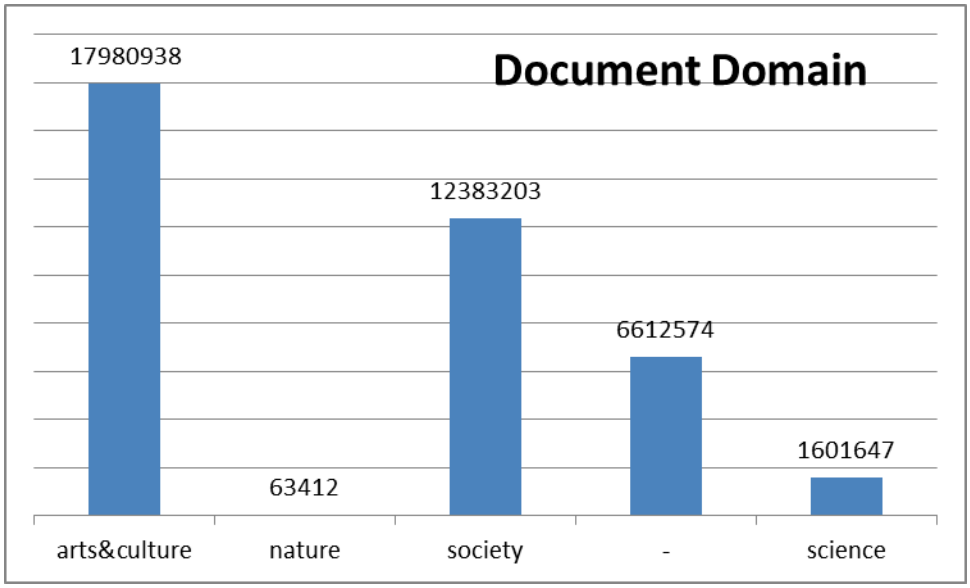


Figura 3. Date cantitative despre domeniile textelor.

Datele din figura de mai sus sunt pertinente textelor informative, nu celor imaginative. Observăm că toate domeniile sunt acoperite: *arts&culture* cu cel mai mare număr de cuvinte, iar *nature* cu cel mai mic. Liniuța a fost folosită pentru a marca textele cărora

[3] Stilul colocvial nu apare evidențiat pentru că am decis să nu includem texte aparținând exclusiv acestui stil. El nu va fi însă absent din corpus, pentru că se va regăsi în texte de ficțiune, adică în stilul numit aici *imaginative*.

nu li se poate atribui un domeniu: în principal texte imaginative și postări de pe bloguri.

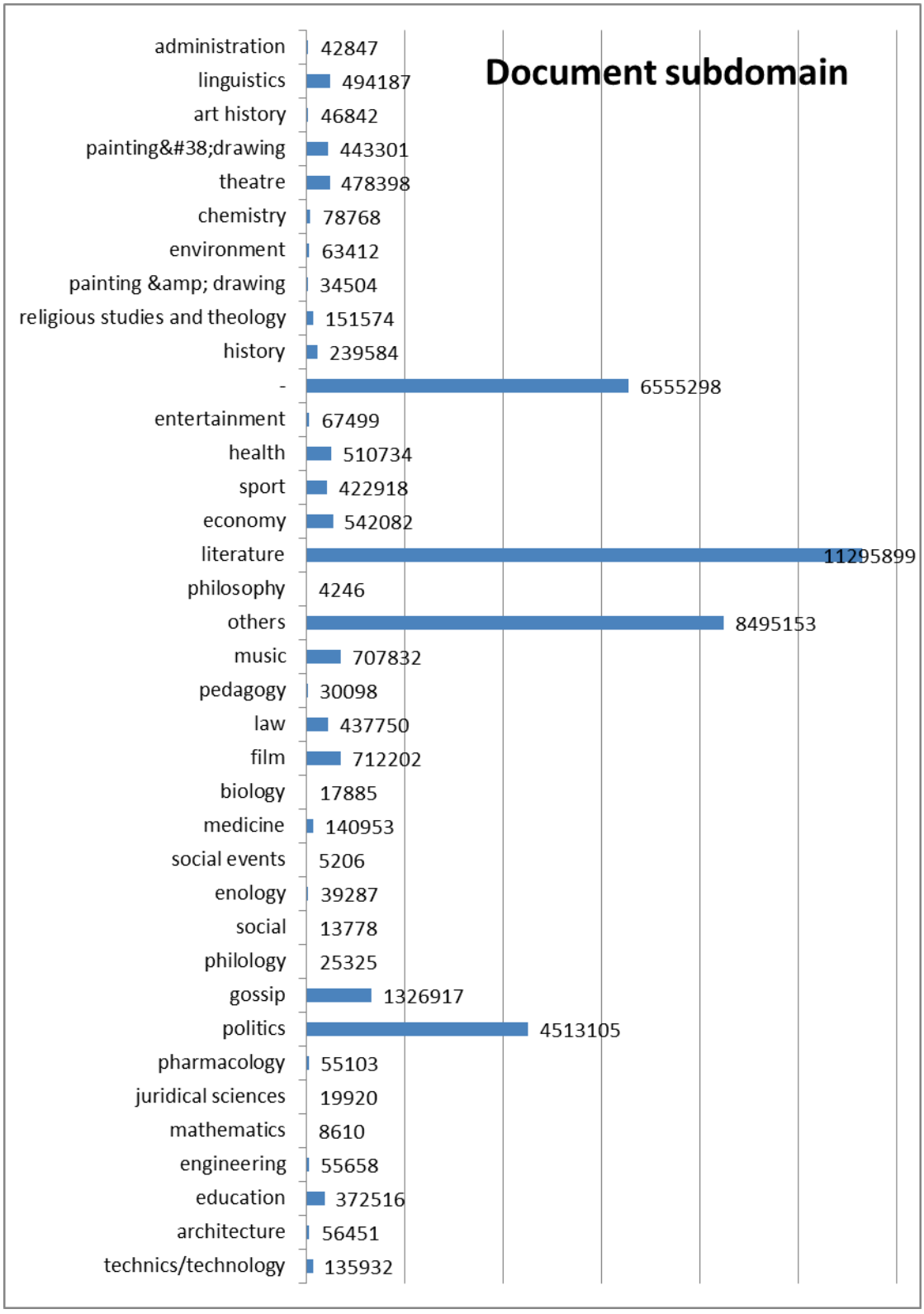


Figura 4. Date cantitative despre subdomeniile textelor.

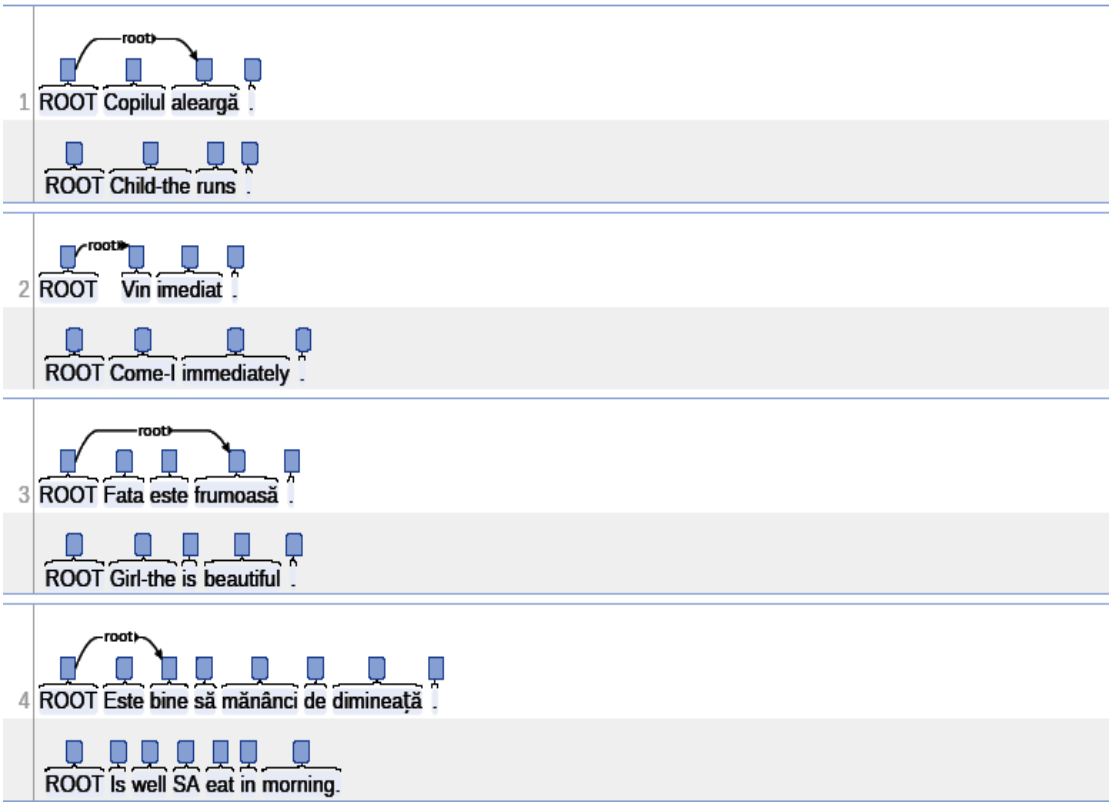
Acestea sunt subdomeniile pentru care avem texte în momentul de față. Multe texte pentru care nu s-a putut aloca un domeniu dintre cele propuse de noi au intrat în categoria *Others*.

3. Descrierea sintaxei limbii române în termenii Universal Dependency Grammar

Având intenția de a crea un treebank românesc și de a-l afilia proiectului Universal Dependency, am început descrierea fenomenelor sintactice românești în acest cadru. Aceasta este redată mai jos, în engleză, și poate fi regăsită și pornind de la adresa http://universaldependencies.github.io/docs/ro/dep/index.html.

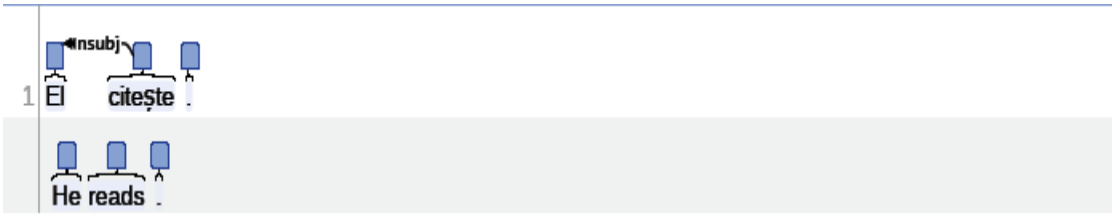
3.1. root: root

The `root` grammatical relation points to the root of the sentence. A fake node "ROOT" is used as the governor. The ROOT node is indexed with "0", since the indexation of real words in the sentence starts at 1.



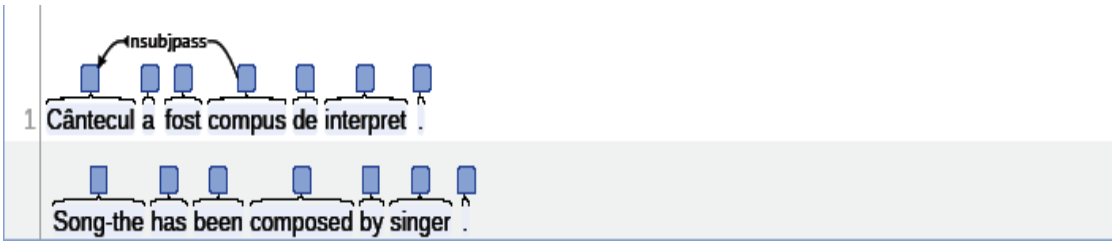
3.2. nsubj: nominal subject

A nominal subject is a nominal phrase which is the syntactic subject of a clause.



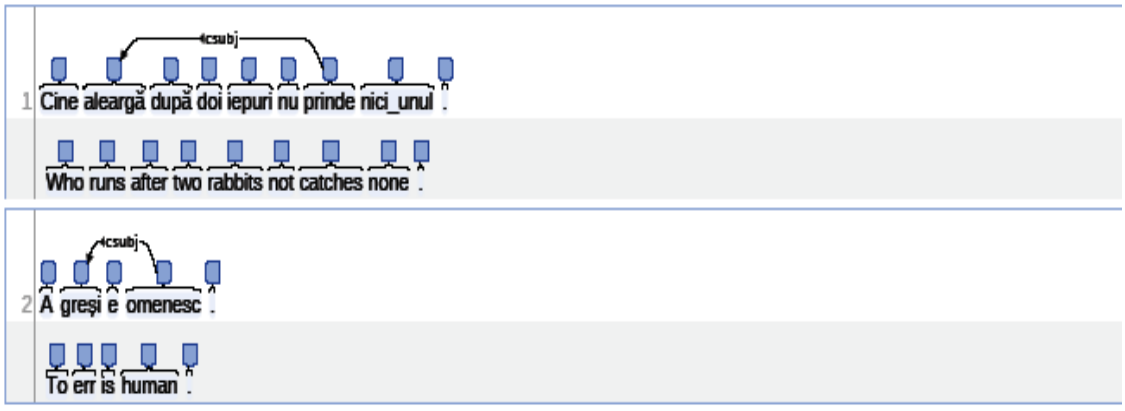
3.3. nsubjpass: passive nominal subject

A passive nominal subject is a noun phrase which is the syntactic subject of a passive clause.



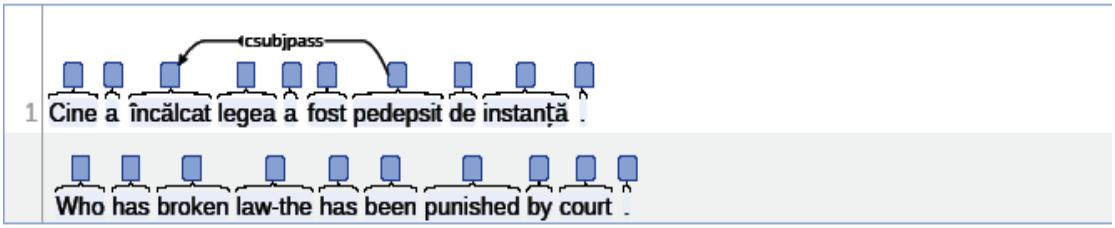
3.4. csubj: clausal subject

A clausal subject is a clausal syntactic subject of a clause, i.e., the subject is itself a clause.



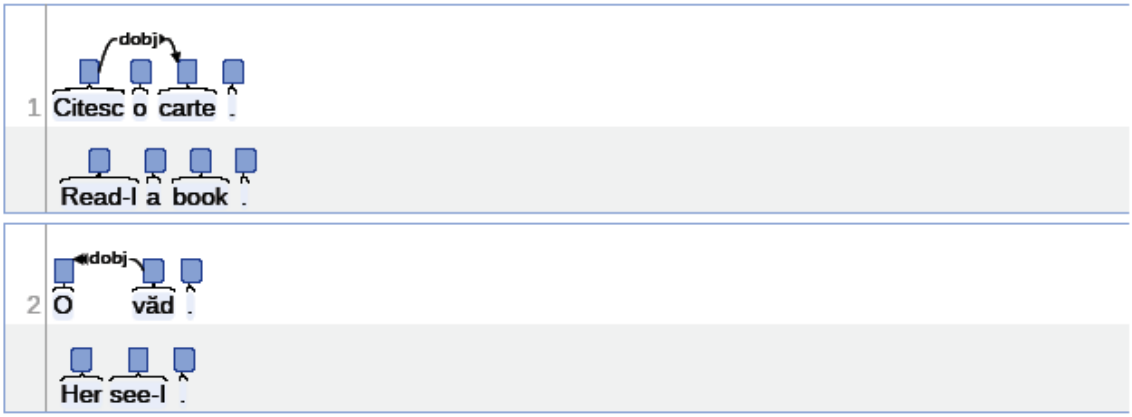
3.5. csubjpass: clausal passive subject

A clausal passive subject is a clausal syntactic subject of a passive clause:



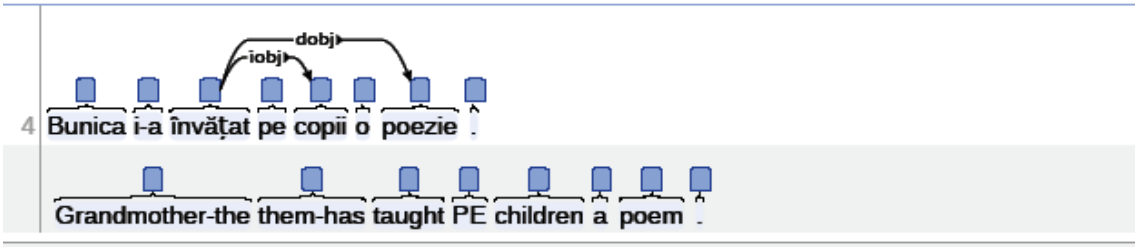
3.6. dobj: direct object

The direct object of a verb is the noun phrase that denotes the entity acted upon.



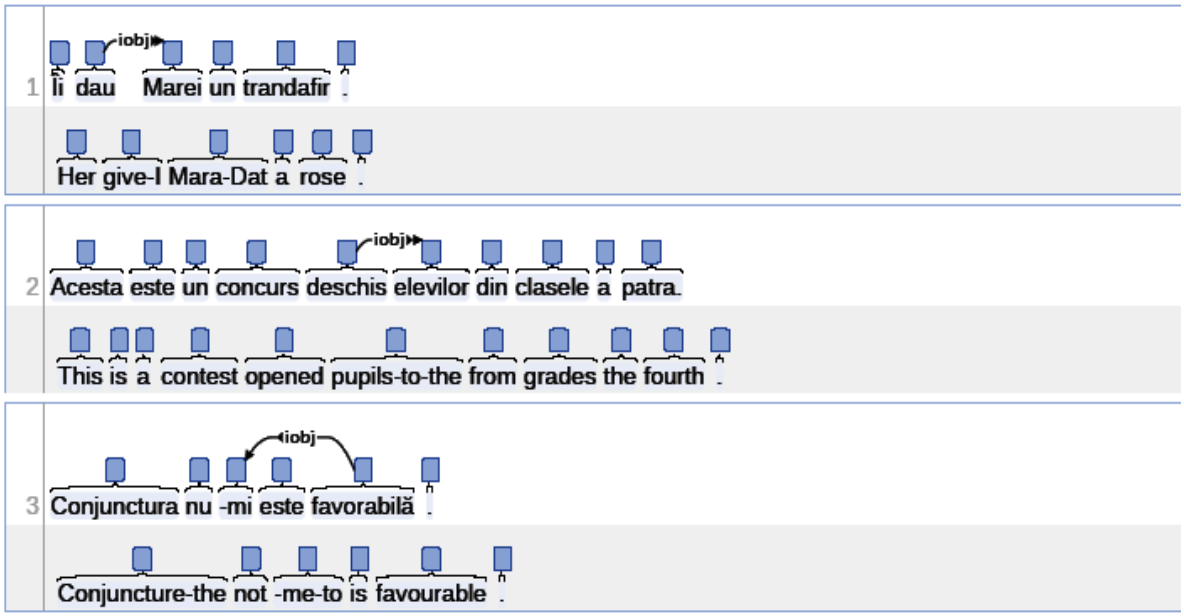
When the direct object is doubled by a pronoun, this is marked as `expl`.

Romanian allows for the occurrence of two Accusative objects with some (uses of certain) verbs: the [+Animate] object (the direct object in traditional grammar terms) is anlysed here as `iobj`, while the other Accusative object (the secondary object in traditional grammar terms) is `dobj`:

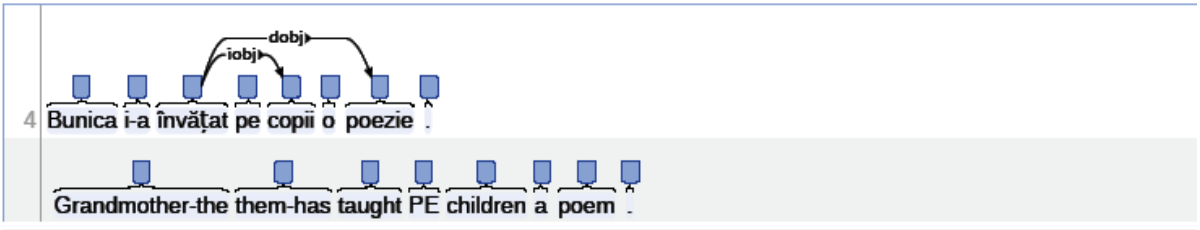


3.7. iobj: indirect object

The indirect object of a verb is any nominal phrase that is a core argument of the verb, usually expressing the recipient, the addressee or beneficiary of the predicate:

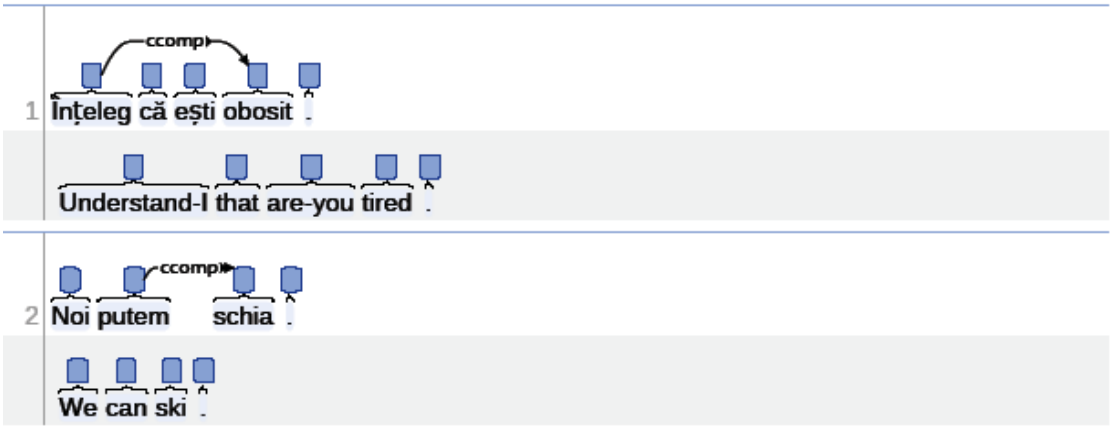


We also analyse as `iobj` the [+Animate] object (the direct object in traditional grammar terms) of verbs with two Accusative objects, whereas the other object (the secondary object in traditional grammar terms) is `dobj`:



3.8. ccomp: clausal complement

A clausal complement of a verb or adjective is a dependent clause which is a core argument. That is, it functions like an object of the verb, or adjective. Such clausal complements may be finite or nonfinite.

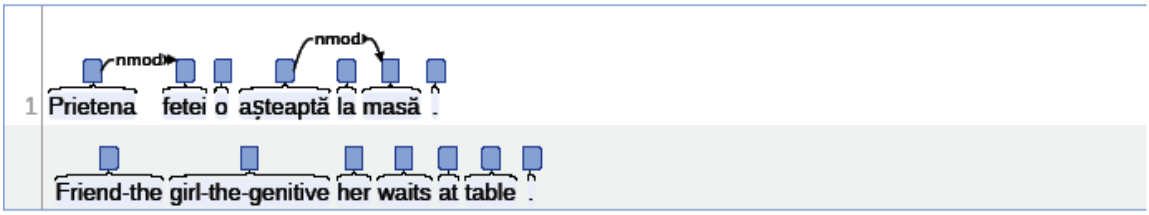


The clausal predicative of the copula verb `a fi` is also analysed as ccomp. NB: This is the only case when the copula verb `a fi` is treated as a head.

3.9. nmod: nominal modifier

The `nmod` relation is used for nominal modifiers. They depend either on another noun (group "noun dependents") or on a predicate (group "non-core dependents of clausal predicates").

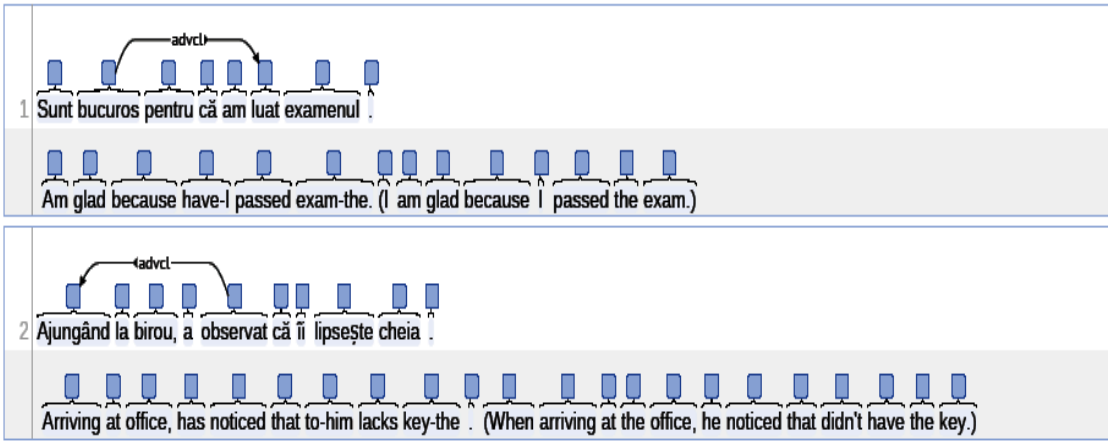
`nmod` is a noun (or noun phrase) functioning as a non-core (oblique) argument or adjunct. This means that it functionally corresponds to an adverbial when it attaches to a verb, adjective or other adverb. But when attaching to a noun, it corresponds to an attribute.



3.10. advcl: adverbial clause modifier

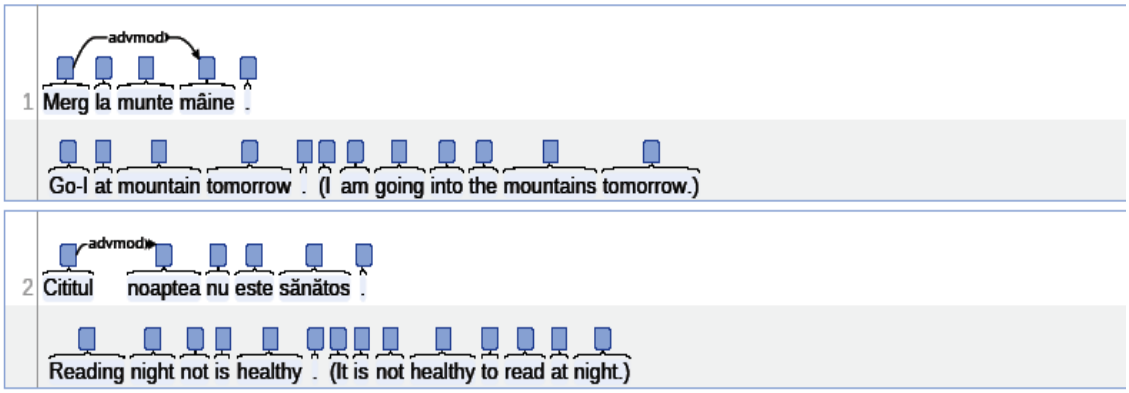
An adverbial clause modifier is a clause which modifies a verb or other predicate (adjective, etc.), as a modifier not as a core complement. This includes things such as

a temporal clause, consequence, conditional clause, purpose clause, etc. The dependent must be clausal (or else it is an advmod) and the dependent is the main predicate of the clause.



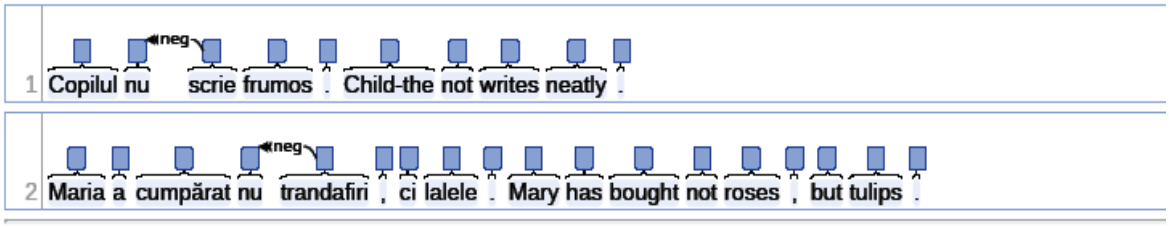
3.11. advmod: adverbial modifier

An adverbial modifier of a word is an adverb or adverbial phrase that serves to modify the meaning of the word.



3.12. neg: negation modifier

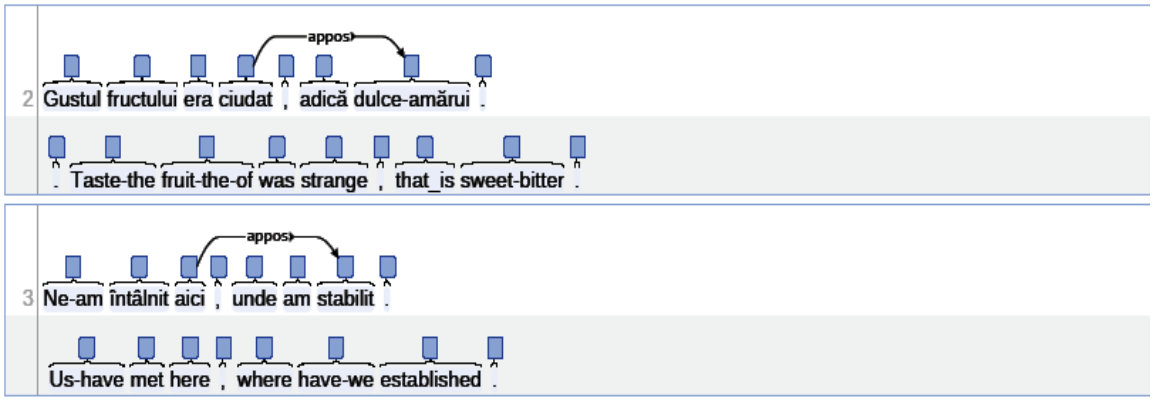
The negation modifier is the relation between a negation word and the word it modifies. Modifiers labeled neg depend either on a noun (group “noun dependents”) or on a predicate (group “non-core dependents of clausal predicates”).



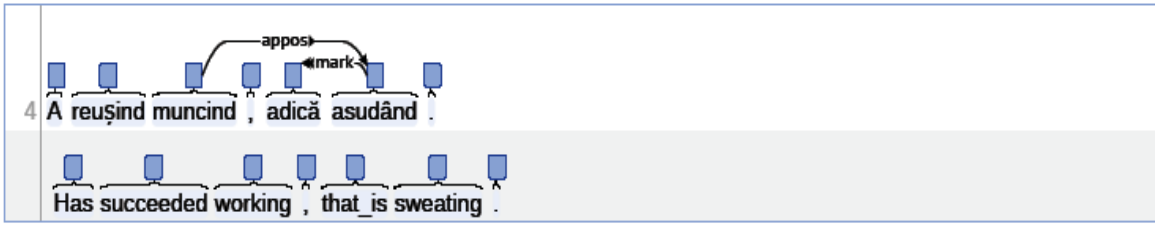
3.13. appos: appositional modifier

An appositional modifier serves to identify its head in a different way. This relation is usually established between noun phrases.

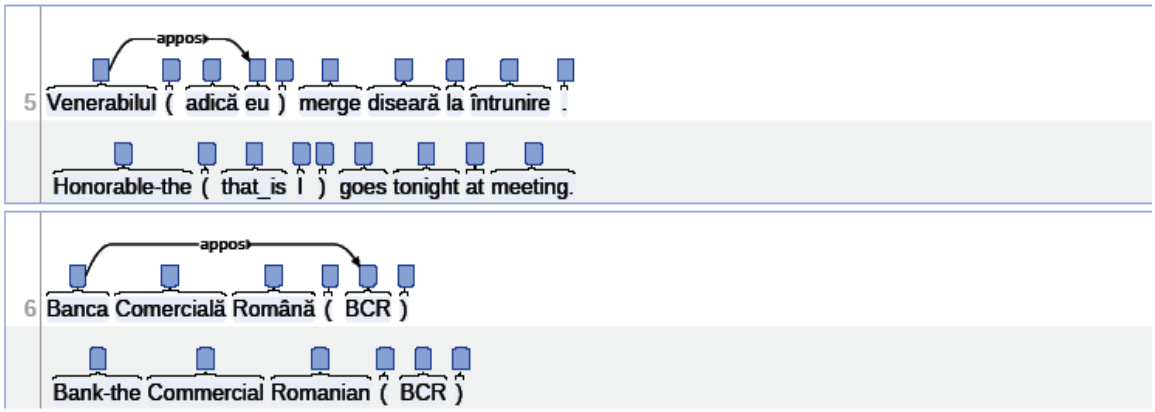
However, other parts of speech and even clauses can also be involved in the relation:



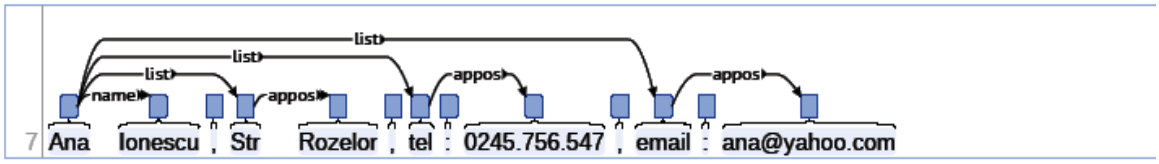
The apposition can be introduced by an adverb (e.g. ‘adică’, ‘anume’, ‘respectiv’, ‘alias’, etc.), which is analysed as a ‘mark’ for the apposition:



It includes parenthesized examples, as well as defining abbreviations in one of these structures.

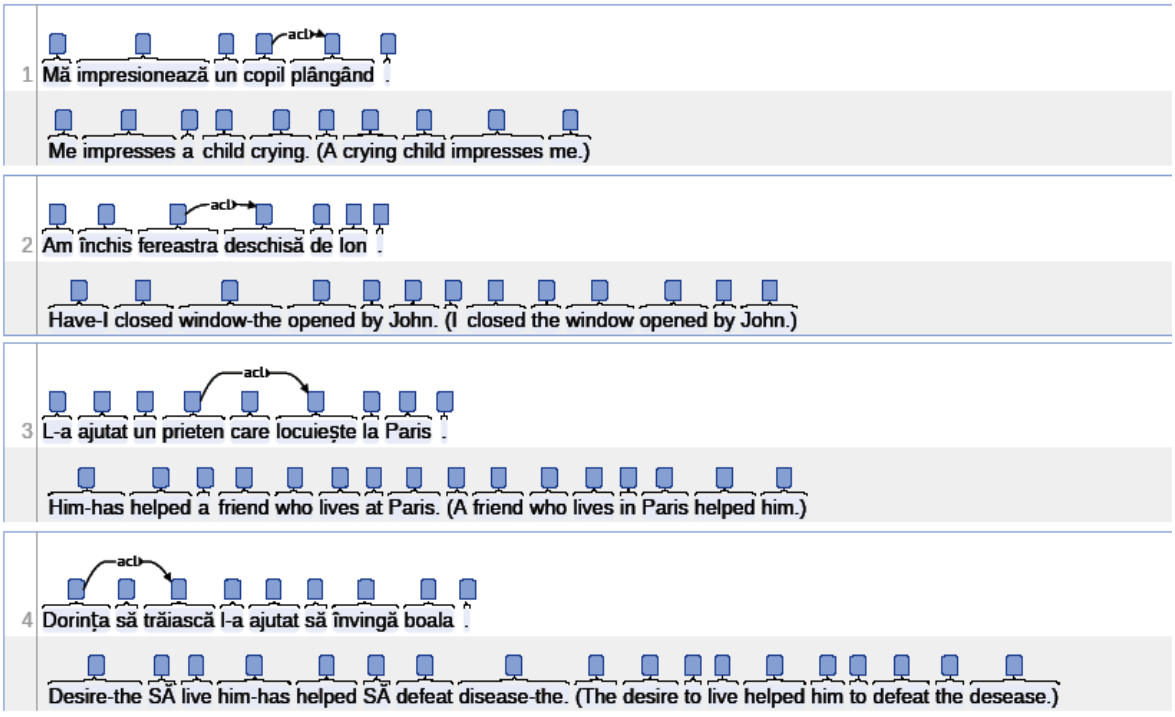


'appos' is also used to link key-value pairs in addresses, signatures, etc.:



3.14. acl: clausal modifier of noun

`acl` stands for finite and non-finite clauses that modify a nominal.

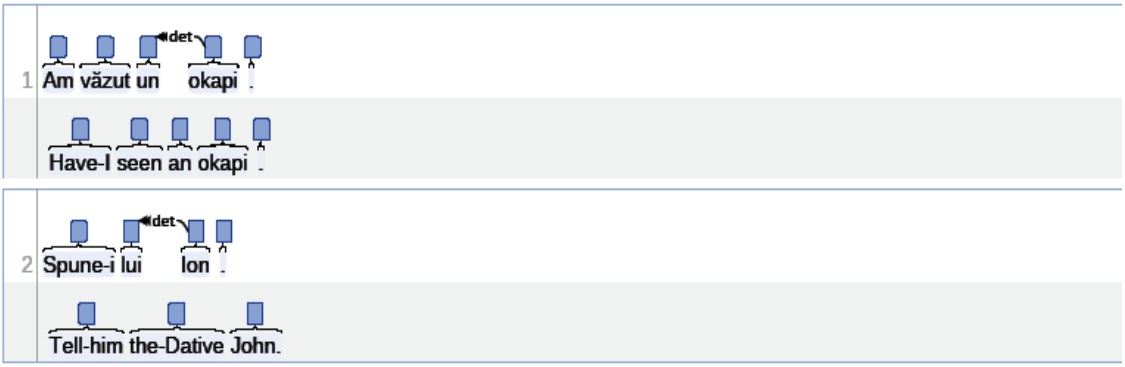


3.15. amod: adjectival modifier

An adjectival modifier of a noun is any adjectival phrase that serves to modify the meaning of the noun.

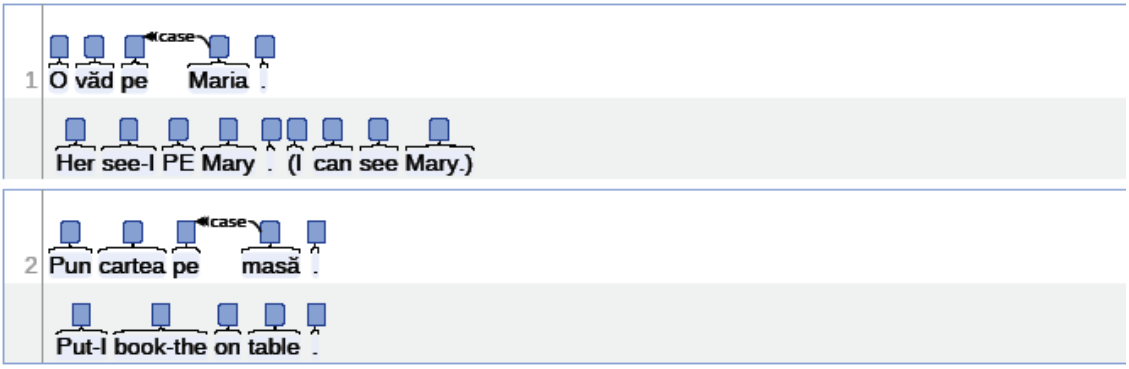
3.16. det: determiner

The relation determiner (`det`) holds between a nominal head and its determiner:



3.17. case: case marking

The `case` relation is used for linking prepositions to their heads:



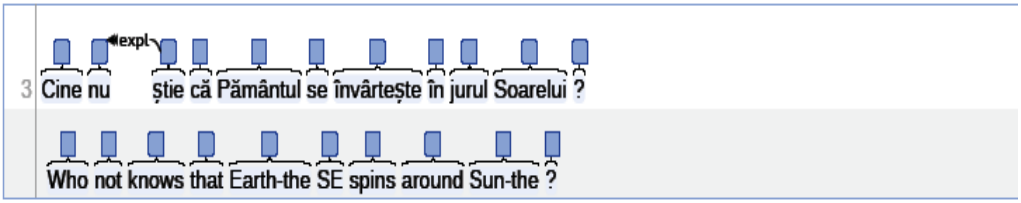
3.18. expl: expletive

Romanian does not have expletives of the English sort. However, we use the `expl` label for the following situations:

- clitic doubling:

- non-referential use of pronouns:

- expletive negation:



3.19. aux: auxiliary

An auxiliary of a clause is a non-main verb of the clause.

Exception: The auxiliary verb used to construct the passive voice is not labeled ‘aux’, but ‘auxpass’.

3.20. auxpass: passive auxiliary

A passive auxiliary of a clause is a non-main verb of the clause which contains the passive information.

3.21. cop: copula

A copula is the relation between the complement of a copular verb and the copular verb `a fi` (only). (We normally take a copula as a dependent of its complement.)

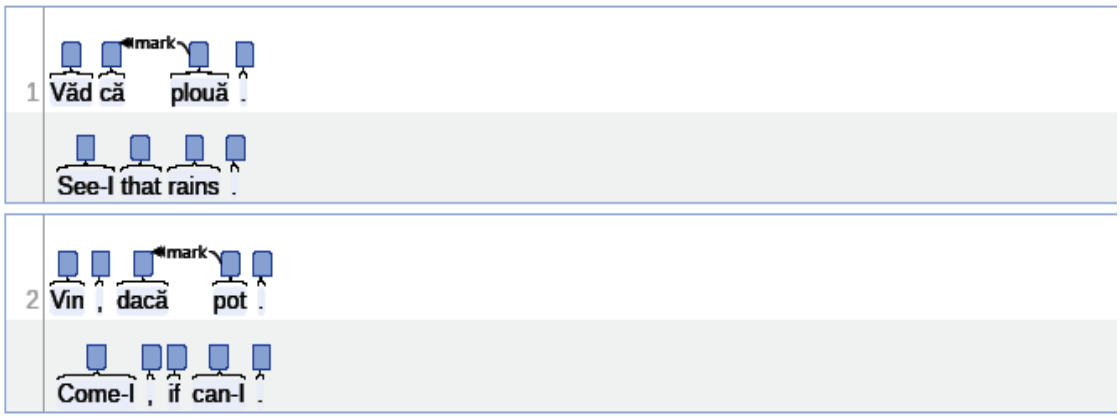
All other copula verbs are heads of clauses and their complements are in `xcomp` relation to them:

When the copula verb has auxiliaries, they are also dependents of the lexical predicate:

When the complement of the copula verb `a fi` is a clause, the copula is the head, and the subordinate clause is in `ccomp` relation with it:

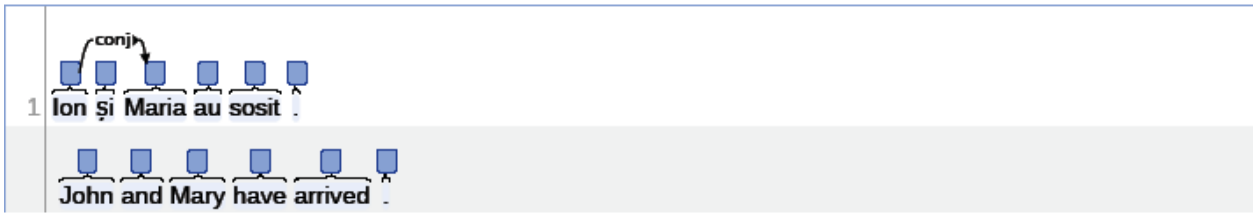
3.22. mark: marker

A marker is the word introducing a finite clause subordinate to another clause:



3.23. conj: conjunct

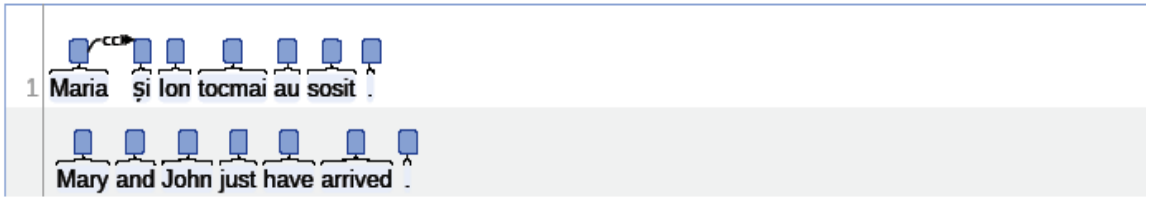
A conjunct is the relation between two elements connected by a coordinating conjunction, such as and, or, etc. We treat conjunctions asymmetrically: the head of the relation is the first conjunct and all the other conjuncts depend on it via the conj relation.



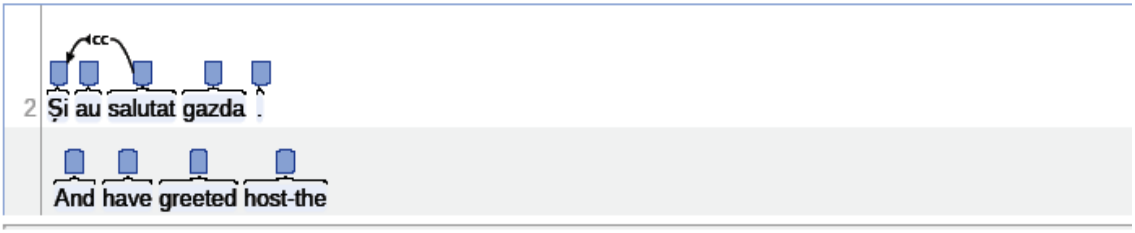
Coordinate clauses are treated the same way as coordination of other constituent types:

3.24. cc: coordinating conjunction

A `cc` is the relation between the first conjunct and the coordinating conjunction delimiting another conjunct:



A coordinating conjunction may also appear at the beginning of a sentence. This is also called a `cc`, and it depends on the root predicate of the sentence. (In fact there is a coordination that spans multiple sentences. We cannot attach a word to the first conjunct because it is in another sentence. Thus we attach it to the first conjunct available in the current sentence: its main predicate.)



3.25. compound: compound

`compound` is used for linking compound words of any part of speech:

4. Aplicație de segmentare a fișierelor audio

Aplicatia de segmentare este folosita pentru a usura adnotarea si segmentarea fisierelor audio, in acelasi timp corectand fisierele text sursa. Astfel ca intrare, aceasta aplicatie foloseste un fisier audio integral (nu este nevoie de fisiere mai mici) si un fisier text care reprezinta transcrierea audio. De preferat, fisierul text este impartit in propozitii inainte de utilizare. Mai specific, este dorita adnotarea fiecarei propozitii cu timp start si timp stop pentru a putea apoi taia automat din fisierul audio care contine tot semnalul audio exact propozitia care se doreste.

Aplicatia are forma unei platfome web, fiind disponibila online. Este scrisa in PHP + Javascript. Baza de date folosita este de tip SQLite.

Utilizarea aplicatiei:

1. **Pregatirea inainte de utilizare**. Administratorul de sistem pregateste o serie de fisiere audio si texte, corespondenta 1-la-1. Este rulat un script care citeste directorul cu fisierele audio si cu textele; fiecare text si fisier audio corespondent este inregistrat in baza de date principala („master.db”), dandu-i-se un cod unic; acest cod unic reprezinta un nume de director in care se va muta automat fisierul audio; fisierul text este importat intr-o baza de date locala (tot SQLite) care are forma unui fisier de tip „.db”, plasat in acelasi director cu fisierul audio. In aceasta baza de date („sentences.db”) sunt salvate fiecare dintre propozitiile gasite in text. Numarul de randuri adaugate in baza de date este chiar numarul de propozitii distincte afisate pe ecran la rulare, din acest motiv fiind indicat sa se ruleze un tool de segmentare text in propozitii individuale. In acest moment aplicatia poate fi folosita de catre utilizatorii platformei.

2. **Segmentarea propozitiilor de catre utilizatori.** Utilizatorii platformei se autentifica (nume de utilizator/parola), apoi apasa pe „Document nou”. Daca au deja un document in lucru, acest document va fi automat afisat pe ecran.
   Documentul va fi afisat atat sub forma audio (waveform) in partea de sus a ecranului, cat si sub forma text in partea de jos. Partea de afisare audio ca waveform ajuta utilizatorul sa intuiasca inceputul si finalul fiecarei propozitii (acolo unde este liniste in audio banda rosie are amplitudine mica – este ingusta). In timpul ascultarii fisierului audio exista o linie distincta care se deplaseaza deasupra semnalului indicand pozitia curenta in audio.
   In partea de jos a ecranului se afla propozitiile citite din baza de date.
   Utilizatorul apasa pe tasta spatiu pentru a incepe si a termina de segmentat o propozitie. Astfel, el apasa pe spatiu, moment in care in difuzoare incepe sa se auda fisierul audio (in acelasi timp banda albastra se misca corespunzator cu sunetul). Tot in acelasi moment este marcata pentru propozitia curenta un timp de start. Propozitia curenta este evidentiata printr-o icoana speciala in dreptul ei. Celelalte icoane posibile sunt fie verde (propozitia a fost adnotata – are timp start si stop in dreptul ei), fie albastra (propozitia nu a fost adnotata, nu are niciun timp). Propozitia curenta, odata cu apasarea tastei spatiu, va avea timpul start automat pus in dreptul ei. La urmatoarea apasare a tastei spatiu va fi marcat timpul de final al propozitiei curente, si va fi marcat cu acelasi timp timpul de start al propozitiei urmatoare. Sunetul se va opri.
   In acest moment utilizatorul poate efectua operatiunea de editare a textului propozitiei curente (el poate edita practic orice propozitie in orice moment in care nu este redat continut audio). Editarea propozitiei curente este necesara deoarece exista cazuri in care in audio sunt cuvinte in plus sau minus, cuvinte schimbate, etc. De asemenea poate este necesara o normalizare de text: numerele vor fi trecute din cifre in litere (de exemplu „33” ar trebui transcris ca „treizeci si trei”). Totodata, in acest moment este posibila stergerea sau adaugarea de propozitii noi. De exemplu, daca in „propozitia” curenta avem defapt doua sau mai multe propozitii (acest text a fost astfel prezentat utilizatorului din varii motive legate de import, segmentare in propozitii cu erori, etc.), utilizatorul ar trebui sa stearga din propozitia curenta urmatoarele propozitii si sa adauge randuri noi in care sa copieze propozitiile taiate. De asemenea, daca este nevoie, utilizatorul poate sa stearga propozitii (daca este

nevoie, poate prelua textul din propozitiile urmatoare si sa il concateneze la propozitia curenta – acest lucru este necesar daca intre propozitii nu exista o pauza bine definita in audio, moment in care este de preferat ca propozitiile sa fie unite, astfel ca la segmentare sa obtinem propozitii bine definite).

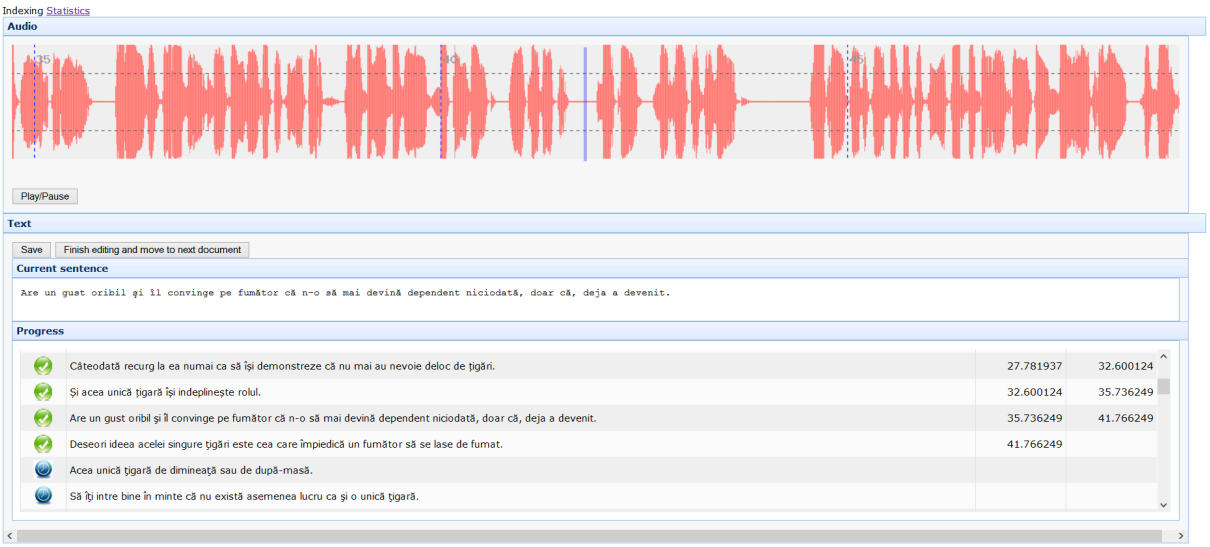
Toate aceste operatii sunt facilitate prin existenta de scurtaturi (short-cut-uri):

- pentru editarea automata a unei propozitii se apasa Ctrl+E;
- pentru reascultarea unei propozitii si refacerea marcajului de inceput/final se apara Ctrl+R;
- pentru stergerea unei propozitii se apasa Ctrl+D cu ea selectata;
- pentru inserarea unei propozitii noi in lista de propozitii se apasa Ctrl+I;

Procedeul de lucru este urmatorul: Utilizatorul apasa spatiu cand considera ca fiecare propozitie se termina, si spatiu pentru a incepe redarea propozitiei urmatoare. In lista de propozitii timpul start/stop se completeaza automat. Acolo unde este nevoie, utilizatorul editeaza, adauga sau sterge propozitii, astfel incat sa fie o corespondenta 1-la-1 intre ceea ce se aude cu propozitiile transcrise.

Odata cu terminarea propozitiilor din cadrul unui document se apasa pe „finalizare editare si deschidere document nou”. Acest buton confirma salvarea timpilor si modificarilor propozitiilor in baza de date si va trece la urmatorul fisier audio si document (text) neadnotat.

De asemena, la apasarea butonului „Salveaza” toate modificarile curente vor fi incarcate in baza de date corespunzatoare documentului curent. Astfel, in orice moment un utilizator poate salva datele, inchide si relua lucrul la un timp ulterior fara a pierde nimic.



3. **Segmentarea automata a fisierelor audio**. Odata incheiat procedeul de adnotare al fiecare propozitii cu timpul start/stop pentru toate documentele pentru care se doreste acest lucru, administratorul de sistem va rula un script numit „dump_corpus.php”. Acest script efectueaza urmatorii pasi:
    a. Creaza o lista cu toate documentele care au fost finalizate.
    b. Pentru fiecare document in parte, citeste din baza de date „sentences.db” timpii de start si stop ai fiecarei propozitii.
    c. Pentru fiecare propozitie, scriptul va crea 3 fisiere:
        i. Fisierul wav: acest fisier este decupat automat din fisierul principal, incepand de la timpul de start pana la timpul de final al propozitiei curente. Fisierul va fi denumit automat cu un indicator de nume unic si un numar de propozitie.
        ii. Fisierul txt: fisierul va avea acelasi nume ca fisierul wav, insa cu terminatia .txt. Acest fisier contine pe o singura linie propozitia nemodificata. Singura corectura care i se poate aduce este modificarea diacriticelor ş si ţ din varianta veche cu sedila in varianta noua cu virgula.
        iii. Fisierul lab: similar cu fisierul text, insa cu extensia „.lab” contine propozitia pe un singur rand, insa modificata. Toate literele sunt trecute in litere mici; Toate semnele de punctuatie sunt eliminate; Toate liniutele sunt eliminate si cuvintele care erau unite prin aceste linuite sunt unite la propriu (ex: „Ducandu-i” devine „ducandui”, sau „şi-o” devine „şio”).

Aceste fisiere sunt necesare mai departe pentru antrenarea modelelor fonetice.

Aplicatia de segmentare este utila pentru a reduce volumul de munca necesar pentru a efectua operatiuni de taiere audio si corectare text. Utilizand aceasta platforma nu mai este necesara folosirea unei intregi suite de aplicatii neconectate intre ele (ex: editare audio cu Audacity, apoi editare text in notepad, apoi taiere audio manuala cu Audacity sau sox, etc).

5. Aplicație de aliniere fonetică

Aceasta aplicatie de aliniere fonetica este utilizata pentru a obtine o aliniere intre sunet si fonemele individuale ale fiecarui cuvant. Aplicatia are forma unui set de scripturi si necesita o serie de utilitare aditionale instalate pe masina gazda.

Astfel, alinierile la nivel de foneme sunt obtinute folosind toolkit-ul HMM (HTK) (Young et al., 1993), folosind urmatoarea procedura:

1. Se genereaza dictionarul de transcriere fonetica (folosind HDMan) pentru toate cuvintele din corpus folosind o versiune imbogatita a dictionarului RSS (Stan et al., 2011). Cuvintele OOV (Out-Of-Vocabulary Words) gasite in corpus sunt automat transcrise folosind trei algoritmi diferiti de grafem-fonem (G2P): algoritmul bazat pe MIRA, dezvoltat de ICIA, un clasificator MaxEnt precum si un algoritm propriu numit DLOPS (Boroș et al., 2012). Toate variantele generate au fost adaugate in lexicon ca ulterior folosind HVite sa fie aleasa cea mai probabila transcriere;
2. Se genereaza o transcriere fonetica initiala (prin HLEd) folosind prima pronuntie disponibila din dictionar pentru fiecare cuvant din corpus;
3. Utilizand corpusul transcris fonetic din pasul anterior, se genereaza un model initial HMM tri-stare, stanga→dreapta (model HMM monofon in terminologia HTK) pentru fiecare fonem necesar (fara a include pauzele scurte – ‚sp') folosind HERest, reestimand modelul initial de 4 ori. Limita de pruning (folosind optiunea –t din HERest) a fost setata la '250.0 150.0 1000.0';
4. Se adauga modelul HMM pentru pauze scurte (‚sp') initial copiat din modelul de „liniste" (‚sil') gasit la inceputul si finalul propozitiilor, si se reestimeaza toate modelele HMM de inca 4 ori folosind aceiasi parametrii HERest;
5. Se regenereaza transcrierea fonetica (inclusiv generarea de pauze scurte) pentru corpusul audio (folosind HVite) folosind cel mai bun model HMM

obtinut in pasul precedent pentru a obtine pronuntia care se potriveste cel mai bine pe datele acustice (in cazul in care un cuvant are mai multe moduri in care poate fi pronuntat in dictionar)

6. In final, se reestimeaza (tot de 4 ori) modelul monofon HMM incluzand pauzele scurte cu noua transcriere a corpusului si se genereaza alinierile folosind HVite cu optiunea de a exporta timpul start/stop pentru fiecare monofona.

Astfel, se obtine o baza de date care contine fiecare fonem corect aliniat corpusului audio, baza de date necesara pentru a putea ulterior antrena modele de sinteza de voce.

6. Diseminare

Cele două colective (ICIA şi IIT) au trimis articolul din Anexa 3 la workshop-ul Challenges in the management of large corpora (http://corpora.ids-mannheim.de/cmlc.html). Acesta a fost acceptat spre prezentare (Lancaster, iulie 2015).

7. Atragerea de voluntari

Dată fiind amploarea proiectului nostru, încercăm atragerea de voluntari dintre studenții facultăților filologice din diverse centre din țară. Pentru aceasta, am avut o prezentare la Facultatea de Limbi Străine din cadrul Universității București, pe data de 10 martie 2015, organizată cu sprijinul Asociației Studenților din Facultatea respectivă. Prezentarea se găsește în Anexa 4 a raportului de față.

De asemenea, am încheiat protocoale de practică cu Universitatea din București, în baza cărora studenții să efectueze la noi ore de practică, în care să lucreze la proiectul CoRoLa.

Anexa 1. Selecția titlurilor solicitate la Editura Humanitas (februarie 2015).

Autor	**Titlu**
Marius Chivu	Trei săptămâni în Himalaya
Cristian Pătrăşconiu	Noua şcoală de gândire a dreptei
Ionuţ Sterpan, Dragoş Paul Aligică	Dreapta intelectuală
Neagu Djuvara	Cum s-a născut poporul român?
Sanda Marin	Carte de bucate
Vasile Răducă	Ghidul creştinului ortodox de azi
Mircea Cinteză	Ce-i cu inima mea, doctore?
Irina Nicolau	Ghidul sărbătorilor româneşti
Bebe Mihăescu	A face dragoste aproape perfect
Aurelia Marinescu	Codul bunelor maniere astăzi
Monica Lovinescu	O istorie a literaturii române pe unde scurte
Ana Blandiana	Fals tratat de manipulare
Ioana Pârvulescu	În intimitatea secolului 19
Sega	Namaste. Un roman de aventuri spirituale în India
Andrei Pippidi	Case şi oameni din Bucureşti (vol. I)
Constantin Eretescu	A doua naştere
Ioan Popa	Robi pe Uranus
T. O. Bobe	Contorsionista
T. O. Bobe	Cum mi-am petrecut vacanţa de vară
Ana Blandiana	Întoarcerea lui Arpagic
Mirela Stănciulescu	Emoţia
Lidia Stăniloae	Raiul inocenţilor
Radu Paraschivescu	Fluturele negru
Ruxandra Cesereanu	Angelus
Adrian Oprescu	Tomi
Alexandru Graur	Dicţionar al greşelilor de limbă
Alexandru Graur	„Capcanele“ limbii române
Alex Ştefănescu	Cum te poţi rata ca scriitor
Dan C. Mihăilescu	Despre omul din scrisori. Mihai Eminescu
Mihaela Nicola	Cu mănuşi
Septimiu Chelcea	Ruşinea şi vinovăţia în spaţiul public
Ana Blandiana	Intoarcerea lui Arpagic
Marius Daniel Popescu	Simfonia lupului
Valeria Gutu Romalo	Corectitudine si greseala. Limba romana de azi
Teodor Baconsky	Rasul patriarhilor
Marius Oprea	Adevarata calatorie a lui Zahei
Adriana Bittel, Adriana Bittel, Ana Blandiana, Gabriel Liiceanu, Nicolae Manolescu, Ioana Pârvulescu	Povesti de dragoste la prima vedere
Eugen Munteanu	Lexicologie biblica romaneasca
Vladimir Tismăneanu	Raport final
Tatiana Niculescu Bran	Spovedanie la Tanacu

Nicolae Steinhardt	Primejdia mărturisirii
Doina Cornea	Puterea fragilității
Horia-Roman Patapievici	Omul recent
Andrei Oisteanu	Imaginea evreului in cultura romana
Mioara Avram	Gramatica pentru toti
Silviu Angelescu	Calpuzanii
Dumitru Nicodim	Casa lui David
Stefan Agopian	Manualul intamplarilor
Gabriel Badea-Păun	Carmen Sylva
Ovid S. Crohmălniceanu	Amintiri deghizate
Oana Pellea	Jurnal 2003 - 2009
Mircea I. Manolescu	Arta avocatului
Ioan Gliga	Drept financiar
Stelian Tanase	Luxul melancoliei
Maria Ellis	Carte de colorat pentru orbi
Gellu Naum	Gellu Naum
Cristian Presură	Fizica povestită
Carmen Matei Musat	Romanul romanesc interbelic. Dezbateri teoretice, polemici, opinii critice
Gabriela Duda	Literatura romaneasca de avangarda
Rodica Zafiu	Poezia simbolista romaneasca
Liviu Papadima	Comediile lui I.L. Caragiale
Dan Horia Mazilu	Cronicarii moldoveni
Ion Manolescu	Literatura memorialistica
Maria Cvasnii Catanescu	Limba romana. Origini si dezvoltare

Anexa 2. Selecția titlurilor solicitate la Editura Polirom (aprilie 2015):

Razboaiele mele	Adelin Petrisor
Saman	Adina Dabija
Instruire asistata de calculator. Didactica informatica	Adrian Adascalitei
1989	Adrian Buz
Pirati si corabii. Incursiune intr-un posibil imaginar al marii	Adrian G. Romila
Politica si cultura	Adrian Marino
Pentru Europa	Adrian Marino
Introducere in filosofia politica	Adrian Miroiu
Actiune colectiva si bunuri comune in societatea romaneasca	Adrian Miroiu (coord.), Iris-Patricia Golopenta (coord.)
Competitia politica in Romania	Adrian Miroiu (coord.), Serban Cerkez (coordonator)
Psihologia servitutii voluntare	Adrian Neculau
Inconstientul cognitiv: modele teoretice, suport experimental si aplicatii	Adrian Opre
Psihopedagogie speciala. Modele de evaluare si interventie	Adrian Rosan (coord.)
Incursiune in fitoterapie. Plantele medicinale in dermatocosmetica	Adrian Vasilca-Mozaceni
Ghid pentru cercetarea educatiei	Adrian Vicentiu Labar, Nicoleta Laura Popa, Liviu Antonesei (coord.)
Asistenta in familie a persoanei cu deficienta functionala. Tehnici de ingrijire si manevrare a bolnavului	Adriana Albu, Constantin Albu, Ioan Petcu
Amazoanele. O poveste	Adriana Babeti
Managementul de succes	Adriana Prodan
Critica in transee. De la realismul socialist la autonomia esteticului	Alex Goldis
4 decenii, 3 ani si 2 luni cu filmul romanesc	Alex. Leo Serban
In dialog cu anticii	Alexandra Ciocarlie
Dupa Sodoma	Alexandru Ecovoiu
Munca obligatorie a evreilor din Romania (1940-1944). Documente	Alexandru Florian (editor), Ana Barbulescu (editor), Alexandru Climescu, Laura Degeratu
Despre lucrurile cu adevarat importante	Alexandru Paleologu
Jurnal (2 volume)	Alice Voinescu
Raport de cornere. Cit se intinde plapuma sportului?	Alin Buzarin

Psihologie interculturala	Alin Gavreliuc
De ce nu iau romanii premiul Nobel	Alina Mungiu-Pippidi
Tranzitia. Primii 25 de ani	Alina Mungiu-Pippidi, Vartan Arachelian
Povesti cu scriitoare si copii	Alina Purcaru (coord.)
Psihopedagogia persoanelor cu cerinte speciale. Strategii de educatie integrata	Alois Ghergut
Elaborarea si managementul proiectelor in serviciile educationale. Ghid practic	Alois Ghergut, Ciprian Ceobanu
Miini cuminti. Copilul meu autist	Ana Dragu
Adoptia si atasamentul copiilor separati de parintii biologici	Ana Muntean
Transmisiunea in direct	Anamaria Neagu
Stilul religiei in modernitatea tirzie	Anca Manolescu
Violenta, trauma, rezilienta	Anca Munteanu, Ana Muntean
Fundamentele educatiei interculturale. Diversitate, minoritati, echitate	Anca Nedelcu
Cum gindesc si cum vorbesc ceilalti. Prin labirintul culturilor	Andra Serbanescu
Intrebarea. Teorie si practica	Andra Serbanescu
Forta politica a femeilor	Andreea Paul (Vass) (coord.)
Bunul, raul si uritul in cinema	Andrei Gorzo
Reconstructia pragmatica a filosofiei (vol.I)	Andrei Marga
Manual de Relatii Internationale	Andrei Miroiu, Radu-Sebastian Ungureanu
Cutia cu batrini	Andrei Oisteanu
Ordine si Haos. Mit si magie in cultura traditionala romaneasca	Andrei Oisteanu
Imaginea evreulu in cultura romana	Andrei Oisteanu
Zaraza	Andrei Ruse
Caragiale dupa Caragiale. Arcanele interpretarii: exagerari, deformari, excese	Angelo Mitchievici
Tranzactii comerciale internationale	Aurel Burciu (coord.), Rozalia Iuliana Kicsi, Irina Stefana Cibotariu, Marcela Cristina Hurjui

Anexa 3. Articolul acceptat la workshop-ul Challenges in the management of large corpora

CoRoLa Starts Blooming – An update on the Reference Corpus of Contemporary Romanian Language

Dan Tufiș, Verginica Barbu Mititelu, Elena Irimia, Ștefan Daniel Dumitrescu, Tiberiu Boroș
Research Institute for Artificial Intelligence "Mihai Drăgănescu"
13 Calea 13 Septembrie, 050711, Bucharest, Romania
{tufis, vergi, elena, sdumitrescu, tibi}@racai.ro

Horia Nicolai Teodorescu, Dan Cristea, Andrei Scutelnicu, Cecilia Bolea, Alex Moruz, Laura Pistol
Institute for Computer Science, Iași
2 T. Codrescu St, 700481, Iași, Romania
hteodor@etti.tuiasi.ro, dcristea@info.uaic.ro, andreiscutelnicu@gmail.com, cecilia.bolea@iit.academiaromana-is.ro, mmoruz@info.uaic.ro laura.pistol@iit.academiaromana-is.ro

Abstract

This article reports on the on-going CoRoLa project, aiming at creating a reference corpus of contemporary Romanian (from 1945 onwards), opened for on-line free exploitation by researchers in linguistics and language processing, teachers of Romanian, students. We invest serious efforts in persuading large publishing houses and other owners of IPR on relevant language data to join us and contribute the project with selections of their text and speech repositories. The CoRoLa project is coordinated by two Computer Science institutes of the Romanian Academy, but enjoys cooperation of and consulting from professional linguists from other institutes of the Romanian Academy. We foresee a written component of the corpus of more than 500 million word forms, and a speech component of about 300 hours of recordings. The entire collection of texts (covering all functional styles of the language) will be pre-processed and annotated at several levels, and also documented with standardized metadata. The pre-processing includes cleaning the data and harmonising the diacritics, sentence splitting and tokenization. Annotation will include morpho-lexical tagging and lemmatization in the first stage, followed by syntactic, semantic and discourse annotation in a later stage.

1 Introduction

In 2012 the Romanian Academy Research Institute for Artificial Intelligence "Mihai Drăgănescu" from Bucharest (RACAI) finalized the Romanian Balanced Corpus (ROMBAC[4]) (Ion et. al, 2012) containing 44,117,360 tokens covering four domains (News, Medical, Legal, Biographic and Fiction). The nucleus of ROMBAC was represented by the RoCo_News corpus (Tufiș and Irimia, 2006), a hand validated corpus of almost 7 million tokens from the weekly magazine Agenda (2003-2006).

[4] http://www.meta-net.eu/meta-share

Since 2014 the concern for creating a bigger corpus has been joined by the Institute for Computer Science in Iasi, in a larger priority project of the Romanian Academy: The Reference Corpus of Contemporary Romanian Language.

The time span covered by the project is 1945-present, with two subperiods (1945-1990, 1990-present), with clear differences, mainly at the lexical level. From this perspective, a big challenge for us is the collection of electronic texts to cover the whole period. For the last couple of decades there is an important amount of such texts available. However, in the case of the texts from previous decades considerable effort needs to be done for finding the owners of the texts IPR, for scanning, OCRizing and correcting the texts. This could imply raising the awareness of main libraries about the cultural responsibility of digitizing even contemporary books, not only the old ones.

2 Objectives

When finished, CoRoLa will be a medium to large corpus (more than 500 million word forms), IPR cleared, in which all functional styles will be represented: scientific, official, publicistic and imaginative. Although the colloquial style is not a major concern for us, it will definitely be included, due to its use in imaginative writing. The provisional structure of the corpus is described in some details in Barbu Mititelu and Irimia (2014). Unlike its predecessor, CoRoLa will include a syntactically annotated sub-corpus (treebank) and an oral component. All textual data will be morpho-lexically processed (tokenized, POS-tagged and lemmatized). The treebank (we target 10,000 hand validated sentences) and the oral component (targeted: 300 hours of transcribed recorded speech) have additional annotations (dependency links, respectively speech segmentation at sentence level, pauses, non-lexical sounds and partial explicit marking of the accent).

Particular attention is paid to data documentation, i.e. associating it with standardized metadata. We adopted the CMDI (Component MetaData Infrastructure)[5] approach for the creation of our metadata.

3 Data Collection and Cleaning

The resource we are building will have two important attributes: it will be representative for the language stage, thus covering all language registers and styles; it will be IPR cleared, which is a challenging task, triggered by the need to observe the intellectual property law. The categories of content excepted by this law are: political, legislative, administrative and judicial. Therefore, without the written accept from IPR owners, from the other kinds of texts only tiny fragments of no more than 10,000 characters can be used. We must also consider only texts written with correct diacritics (otherwise, the linguistic annotation will be highly incorrect).

To ensure the volume and quality of the texts in the corpus, as well as copyright agreements on these texts, our endeavour was to establish collaborations with publishing houses and editorial offices. So far (March 2015), we have signed agreements with the following publishing houses: Humanitas, Polirom, Romanian Academy Publishing House, Bucharest University Press, "Editura Economică", ADENIUM Publishing House, DOXOLOGIA Publishing House, the European Institute Publishing House, GAMA Publishing House, PIM Publishing House. Some magazines and newspapers have also agreed to help our project by providing access to their articles: România literară, Muzica, Actualitatea muzicală, Destine literare, DCNEWS, PRESSONLINE.RO, the school magazine of Unirea National College from Focşani, SC INFOIASI SRL, Candela de Montreal. Until now four bloggers have also agreed to allow us to include some of their posts in the corpus: Simona Tache[6], Dragoș Bucurenci[7], Irina Șubredu[8] and Teodora Forăscu[9]. Also, we have

[5] http://www.clarin.eu/content/component-metadata
[6] http://www.simonatache.ro
[7] http://bucurenci.ro
[8] http://irina.subredu.name

signed agreements with the writers Corneliu Leu and Liviu Petcu.Oral texts (read news, live transmissions and live interviews) (one hour per working day) are provided by Rador (the press agency of Radio Romania) and Radio Iaşi – a local broadcasting agency. All data providers readiness to get involved was a very pleasant surprise for us and we express here, again, our gratitude.

Another challenge in corpus creation is to have texts in a clean format, easy to process and annotate. Once our collaborators dispatch a textual resource (usually in unprotected pdf files, rarely in doc files), the first step is to convert it into an adequate format for our pre-processing tools[10].

Given the large amount of texts, we automated a part of the process (Moruz and Scutelnicu, 2014): the text is automatically retrieved from the pdf files, paragraph limits are recuperated, column marking newlines are erased as well as hyphens at the end of the lines. However, a lot of manual work remains to be done: separating articles from periodicals in different files, removal of headers, footers, page numbers, figures, tables, dealing with foot- or end-notes, with text fragments in foreign languages, with excerpts from other authors, etc. When copied from their original sources, the content is converted into the UTF-8 encoding and saved as plain text documents.

CoRoLa is developed and refined in successive steps and the automatic processing chain of the texts to be included has to conform to the format requested by the indexing and searching platform, IMS Open Corpus Workbench (CWB, http://cwb.sourceforge.net/), an open source medium that allows complex searching with multiple criteria and support for regular expressions. It allows to choose the (sub)corpus/(sub)corpora with which to work (choose from among the domains and subdomains, but also from the available authors), to find out words frequencies in a (specified) (sub)corpus, to search for a word or a word form, to search for more words (either consequent or permitting intervening words), to find words collocations and co-occurrences (within a window of a pre-established size), to find lexicalization of specified morphological or/and syntactic structures, n-gram models, etc. The platform has already been installed and tested on the ROMBAC corpus and coupled with our processing chain which produces the adequate annotated format for morphological and shallow syntactic searches. For the near future, we plan to switch to the more powerful corpus management platform KorAP (Bański et al., 2014).

The TTL (Ion, 2007) processing chain ensures, at the time of this writing, the following specific functionalities: sentence splitting, tokenisation, tiered-tagging (Tufiş, 1999), lemmatising and chunking. Future services regarding processing and query facilities for discourse (Cristea & Pistol, 2012) will be provided. CoRoLa will be automatically annotated, but a fragment of it (~2%) will be manually validated.

4 Current Statistics

4.1 Textual Data

At the moment, the corpus contains the data presented in Table 1, where one can notice the domain distribution of the texts, as well as quantitative data related to each domain: tokens (word forms and punctuation).

A finer classification of the documents, according to their sub-domains, outlines the following categories: literature, politics, gossip co-lumns, film, music, economy, health, linguistics, theatre, painting/drawing, law, sport, education, history, religious studies and theology, medicine, technology, chemistry, entertainment, environment, architecture, engineering, pharmacology, art history, administration, oenology, pedagogy, philology, juridical sciences, biology, social, mathematics, social events, philosophy, other.

In parallel with the CoRoLa corpus, at ICIA and UAIC a Romanian treebank is under development (Irimia and Barbu Mititelu, 2015), (Perez, 2014), (Mărănduc and Perez, 2015). Currently each of the two sections of the treebank contains almost 5,000 sentences, which are

[9] https://travelearner.wordpress.com
[10] http://www.racai.ro/en/tools/

in the process of being mapped into the UD project specifications[11]. The final version of the CoRoLa corpus will include the Romanian treebank as well.

DOMAIN		STYLE	
arts&culture	32838881	journalistic	44248356
society	33582123	science	26990172
others	9990383	imaginative	11945283
science	19923533	others	1777475
nature	106196	memoirs	1511676
		administrative	865660
		law	9102494
TOTAL[12]	96441116	TOTAL	96441116

Table 1. Domain and style distribution of textual data.

4.2 Speech data

Speech data collected so far is accompanied by transcriptions (observing the current orthography). Partially (about 10%), it was automatically pre-processed and the transcriptions were XML encoded with mark-up for lemma, part-of-speech and syllabification. Additionally to the XML annotations we provide 3 files which contain the original sentences (".txt" extension) the stripped version (which is obtained by removing all punctuation from the original sentences – useful in training systems such as Sphinx or HTK (Hidden Markov Model Toolkit) – ".lab" extension) and time aligned phonemes (tab separated values which contain each phoneme in the text with its associated start and stop frame – ".phs" extension).

- **RASC** (Romanian Anonymous Speech Corpus) is a crowd-sourcing initiative to record a sample of sentences randomly extracted from Ro-Wikipedia (Tufiș et al., 2014). The corpus is automatically aligned at phoneme/word level.
- **RSS-ToBI** (Romanian Speech Synthesis Corpus) is a collection of high quality recordings compiled by (Stan et al., 2011) and designed for speech synthesis. It was enhanced with a prosodic ToBI-like (Tone and Break Indices) annotation (reference to be added). It is automatically aligned at phoneme/word level.
- **RADOR (**Radio Romania) and **Radio Iaşi** is a collection of radio news and interviews, provided daily by the Romanian Society for Broadcasting and the main Iaşi radio channel. At the time of this writing, the transcriptions are under pre-processing. They are not yet aligned at phoneme/word level.

Corpus	Type	Source	Time length (h:m:s)
RASC	many speakers	RoWikipedia	04:22:02
RSS-ToBI	single speaker	news&fairy tales	03:44:00
RADOR	many speakers	news& interviews	106:52:33
Radio Iaşi	many speakers	interviews	07:00:00 under development
			>121:58:35

[11] https://code.google.com/p/uni-dep-tb/

[12] Currently more textual data, not included into CoRoLa, has been collected, which may be used for improving models of our statistical processing tools. Among them are Wiki-Ro, the Romanian part of a big collection of sentences extracted from Wikipedia within the ACCURAT European project (http://www.accurat-project.eu/) and the Romanian part of the Acquis-Communautaire (Steinberger et al. 2006). They are already pre-processed and contain more than 50 million words. Similarly, we acquired some audio-books (not IPR clarified and thus, not included into CoRoLa) used only for evaluation of our tools.

Table 2. Speech corpora.

Besides these speech corpora, we contracted professional recordings (about 10 hours) of senten-ces selected by us from Romanian Wikipedia. These recordings will enlarge the RASC corpus.

Further information on the already processed speech data are given in the table below.

Corpus	sentences	words	phonemes
RASC	2866	39489	270591
RSS-ToBI	3500	39041	235150
	6266	78530	505741

Table 3. Currently pre-processed speech corpora

A special mention deserves the site "Sounds of the Romanian Language" (Feraru et al., 2010), which is a systematically built, explanatory small collection of annotated and documented recordings of phonemes, words, and sentences in Romanian, pronounced repeatedly by several speakers; the corpus also includes as annex materials numerous papers on the topic and several instruments for speech analysis. Sections of the corpus are devoted to emotional speech, to specific processes as the double subject, and to phonetic pathologies. The corpus is maintained by the Institute for Computer Science of the Romanian Academy[13].

5 Metadata Creation

The challenge in CoRoLa is to create a corpus from which more than only concordances to be extracted, i.e. giving the user the possibility to construct his/her own subcorpus to work with, depending on the domain/style/period/author/etc. The only way to obtain this is to document each file with metadata. For documents sent by publishing houses, etc., we created the metadata files manually. For text files crawled from the web (articles, blogs), we automatically created metadata, with a preliminary phase of mapping the existent classifications of texts on those sites onto our classification of texts.

6 Annotation of the data

As mentioned before, a processing chain[14] has been established, consistent with the tabular encoding specific to the CWB platform and comprising more program modules that execute particular functions. The web-service chain provides:

- sentence splitting: it uses regular expressions for the identification of a sentence end;
- tokenization: the words are separated from the adjacent punctuation marks, the compound words are recognized as a single lexical atom and the cliticized words are split as distinct lexical entities;
- POS tiered-tagging with the large MULTEXT-East tag set; its accuracy is above 98%;
- lemmatization: based on the tagged form of the word, it recovers its corresponding lemma from a large (over 1,200,000 entries) human-validated Romanian word-form lexicon; the precision of the algorithm measured on running texts is almost 99%; for the unknown words (which are not tagged as proper names), the lemma is provided by a five-gram letter Markov Model-based guesser, trained on lexicon lemmas with the same POS tag as the token being lemmatized. The accuracy of the lemma guesser is about 83%. A better lemma-guessing (about 93%) is ensured by a new neural network based-tagger (Boroș et al., 2013), not yet integrated in the processing chain for CWB.

[13] http://www.etc.tuiasi.ro/sibm/romanian_spoken_language/en/

[14] http://ws.racai.ro/ttlws.wsdl

- chunking: for each lexical unit previously tagged and lemmatized, the algorithm assigns a syntactic phrase, guided by a set of regular expression rules, defined over the morpho-syntactic descriptions.

For the further stages in the corpus development, we envisage adding other types of annotations: syntactic parsing, semantic annotation and discourse analysis.

The annotation of the speech data includes, additionally, the syllabation and accent mark-up plus the grapheme to phoneme alignment.

7 Annotation correction

In our previous experiments (Tufiș and Irimia, 2006) with the task of collecting corpora and ensuring a satisfying quality of the resources, we implemented a coherent methodology for the automatic identification of annotation errors.

Most of the errors identified in this manner can also be automatically corrected. This validation procedure was used in the past to correct tagging and lemmatization errors for the journalistic corpus RoCo_News and for ROMBAC and reduced the estimated error rates to around 2%.

The TTL processing workflow explicitly marks the out-of-dictionary words (ODW), excepting proper nouns, abbreviations and named entities. The ODW can be extracted, sorted and counted, then divided into frequency classes. In the past, we concentrated our analysis on the words with at least two occurrences in the corpus (assuming that the others are typographic errors or foreign words) and structured them into error classes, thus being able to split them into errors that need human correction and errors that can be dealt with by implementing automatic correction strategies.

Besides using the mentioned methodology to improve the quality of the entire corpus, we intend to manually validate a limited part of it (2%, i.e. 10 million words). As the process of collecting and managing such an important resource is a life-time task, our attention on assuring its quality will continuously accompany this enterprise.

8 Conclusions

In the international context of growing interest for creating large language resources, we presented here the current phase in the creation of a reference corpus of contemporary Romanian. It is a joined effort of two academic institutes, greatly helped by publishing houses and editorial offices, which kindly accepted the inclusion of their texts at no costs. The corpus will be available for search for all those interested in the study or processing of the Romanian language.

We emphasize the idea that, although large amount of texts are out there on the web, creating an IPR clear reference corpus is quite a challenge, not only due to vast efforts invested in persuading IPR holders to contribute to a cultural action, but also to achieve agreements on what texts and how much of them to include in the corpus. In spite of the decided CoRoLa structure (text types and quantities) of the linguistic data the supplementary data we manage to collect (mainly from the web) is not discarded, but stored for training specialized statistical models to be used in different data-driven applications (CLIR, Q&A, SMT, ASR, TTS).

Acknowledgements

We express here our gratitude to all CoRoLa volunteers, undergraduate, graduate and Ph.D. students, as well as researchers and university staff in computer science and linguistics, who, noble-minded and aware of the tremendous importance that such a corpus will have for the Romanian culture, have generously agreed to help in the process of filling in metadata and cleaning the collection of texts.

References

P. Bański, N. Diewald, M. Hanl, M. Kupietz, A. Witt. 2014. Access Control by Query Rewriting. The Case of KorAP. *Proceedings of the Ninth Conference on International Language Resources and Evaluation (LREC'14)*: 3817-3822.

V. Barbu Mititelu, E. Irimia. 2014. The Provisional Structure of the Reference Corpus of the Contemporary Romanian Language (CoRoLa). In M.Colhon, A. Iftene, V. Barbu Mititelu, D. Tufiş (eds.) *Proceedings of the 10th Intl. Conference "Linguistic Resources and Tools for Processing Romanian Language"*: 57-66.

T. Boroş, R. Ion, D. Tufiş. 2013. Large tagset labeling using Feed Forward Neural Networks. Case study on Romanian Language. *Proceedings of ACL 2013*: 692-700.

T. Boroş, A. Stan, O. Watts, S.D. Dumitrescu. 2014. RSS-TOBI - A Prosodically Enhanced Romanian Speech Corpus. *Proceedings of 9th LREC 2014*: 316-320.

D. Cristea, I.C. Pistol. 2012. Multilingual Linguistic Workflows. In Cristina Vertan and Walther v. Hahn (Eds.) *Multilingual Processing in Eastern and Southern EU Languages. Low-resourced Technologies and Translation*, Cambridge Scholars Publishing, UK: 228-246.

S.D. Dumitrescu, T. Boroş, R. Ion. 2014. Crowd-Sourced, Automatic Speech-Corpora Collection–Building the Romanian Anonymous Speech Corpus. *CCURL 2014: Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era*: 90-94.

S.M. Feraru, H.N. Teodorescu, M.D. Zbancioc. 2010. SRoL - Web-based Resources for Languages and Language Technology e-Learning. *International Journal of Computers Communications & Control*, Vol. 5, Issue 3: 301-313.

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

R. Ion, E. Irimia, D. Ştefănescu, D. Tufiş. 2012. ROMBAC: The Romanian Balanced Annotated Corpus. In Nicoletta Calzolari et al. (Eds.) *Proceedings of the 8th LREC*: 339-344.

E. Irimia, V. Barbu Mititelu. 2015. Building a Romanian Dependency Treebank, Proceedings of *Corpus Linguistics 2015*.

A. Hardie. 2012. CQPweb – combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics* 17 (3): 380–409.

C. Mărănduc, A.C. Perez. 2015. A Romanian dependency treebank. *Proceedings of CICLing 2015*.

A. Moruz, A. Scutelnicu. 2014. An Automatic System for Improving Boilerplate Removal for Romanian Texts. In M. Colhon, A. Iftene, V. Barbu Mititelu, D. Cristea, D. Tufiş, *Proceedings of the 10th International Conference "Linguistic resources and Tools for Processing the Romanian Language"*: 163-170.

A. Perez. 2014. Resurse lingvistice pentru prelucrarea limbajului natural (Lingvistic Resources For Natural Language Processing). Ph.D. thesis, „Alexandru Ioan Cuza" University of Iaşi.

J. Sinclair. 1996. *EAGLES – Preliminary recommendations on Corpus Typology* EAG--TCWC--CTYP/P

A. Stan, J. Yamagishi, S. King, M. Aylett. 2011. The Romanian Speech Synthesis (RSS) corpus: building a high quality HMM-based speech synthesis system using a high sampling rate. *Speech Communication*, *53*(3): 442-450.

R. Steinberger, B. Pouliquen, A. Widiger, C.Ignat, T. Erjavec, D. Tufiş, D. Varga. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In Proceedings of the

5th LREC Conference, Genoa, Italy, 22-28 May, 2006, pp.2142-2147, ISBN 2-9517408-2-4.

D. Tufiş. 1999. Tiered Tagging and Combined Classifiers. In F. Jelinek, E. Nöth (eds) *Text, Speech and Dialogue*, Lecture Notes in Artificial Intelligence 1692, Springer: 28-33.

D. Tufiş, E. Irimia. 2006. RoCo_News - A Hand Validated Journalistic Corpus of Romanian. In *Proceedings of the 5th LREC*: 869-872.

D. Tufiş, R. Ion, A. Ceauşu, D. Ştefănescu. 2008. RACAI's Linguistic Web Services. In Nicoletta Calzolari et al. (Eds.) *Proceedings of the 6th LREC*: 327-333.

D. Tufiş, R. Ion, Ş. D. Dumitrescu, D.Ştefănescu. 2014. Large SMT data-sets extracted from Wikipedia. In *Language Resources and Evaluation Conference (LREC 14)*.Reykjavik, Iceland, May 2014

Anexa 4. Prezentarea proiectului la Facultatea de Limbi Străine

Slide 1

COROLA
Corpus of **RO**manian **LA**nguage

Program Prioritar al Academiei Române
2014-2017
ICIA și IIT

Slide 2

Conținutul prezentării

- Ce este lingvistica corpusului? Ce este un corpus și cum se construiește?
- Ce putem face cu un corpus?
- Ce aflăm din corpus?
- Corpusuri naționale
- Corpusuri românești existente
- Cum poate fi susținut acest proiect?

Slide 3

Ce este lingvistica corpusului, un corpus, și cum se construiește?

"Corpus linguistics, broadly, is a collection of methods *for studying language. It begins with collecting a* large *set of language data – a corpus - which is made usable by computers. Corpora are usually so large that it would be impossible to analyse them by hand, so software packages (often called concordancers) are used in order to study them."*

"A corpus is a body of language representative of a particular variety of language or genre which is collected and stored in electronic form for analysis using specialized software."

"Corpus construction: designing the structure of a corpus, collecting texts according to the decided structure, encoding the corpus, assembling and storing the metadata, marking up the texts where necessary and possibly adding linguistic annotation."

Tony McEnery – Director al Centrului de Cercetări Bazate pe Corpus al Universității Lancaster (UK)

Slide 4

Ce putem face cu un corpus?

- Nimic!!
- Dar... dacă avem un software pentru a accesa informațiile din corpus... putem face o mulțime de tipuri de căutări, pe ale căror rezultate ne putem sprijini cercetările lingvistice.

Slide 5

Ce aflăm din corpus? DATE & CANTITĂȚI

- Colocații: *interested + in*
- Coligații (bazate pe grupuri sintactice): *I don't know + WH word*
- Modele lexico-gramaticale: *BE + difficult + for + NOUN GROUP + TO-INF*
- Conotații: *utterly against, utterly destroying, utterly ridiculous, utterly unsympathetic, utterly stupid, utterly unreasonable*
- Sensurile cuvintelor
- Diferențele dintre cuvinte asemănătoare
- Cum se folosește un cuvânt în diferite texte

Slide 6

Exemple extrase cu un concordanțier

58	a B10 35 classic	example	of good analysis followed by an unworthy
59	it P03 93 is an	example	of the kind of style you should use when
60	ne saw this as an	example	of Morningside's antipathy H26 48 to th
61	ould have been an	example	of that story-telling trick of not L21 131
62	E28 67 follow the	example	of some restaurants and keep a supply o
63	You are an E07 45	example	of what heads of dynasties, ordinary or fa
64	The most notable	example	of this is human language. It J23 122 h
65	re he invokes the	example	of Abraham Ibn Ezra's D07 119 contenti
66	e has given us an	example	of how we should live, but he doesn't D
67	ness, teacher and	example	of divine love, D09 143 and, as their tho
68	t, in support for	example	of land J59 201 reform and free trade, a
69	ot E08 24 find an	example	of the Morris Minor, perhaps the most el
70	t but odd E35 194	example	of the posters of Paris at the turn of the

Slide 7

La ce putem folosi informațiile dintr-un corpus?

- Învățarea unei limbi străine
- Perfecționarea cunoștințelor de limbă
- Cercetări lingvistice
- Aplicații:
  - Indexare inteligentă a conținutului lingvistic (text & voce)
  - Regăsirea după conținut (text & voce) a informației
  - Clasificarea automată după conținut (text & voce) a documentelor
  - Extragerea inteligentă de informații din documente (text mining, text analytics)
  - Rezumarea automată a documentelor
  - Furnizarea de răspunsuri la întrebări în limbaj natural
  - Sisteme de interacțiune bidirecțională prin voce
  - Recunoașterea și sinteza automată a vorbirii
  - Traducere automată (text și voce)

Slide 8

Corpusuri naționale (în Europa)

Corpusul național britanic

BNC (*British National Corpus*) pentru limba engleză
- adresa: http://www.natcorp.ox.ac.uk/)
- volum: 100 de milioane de cuvinte;
- autori: *Oxford University Press* (coordonator) al consorțiului format din:
  - Editorii de dicționare Longman (acum Pearson Education) și Larousse Kingfisher Chambers;
  - centrele de cercetare academică *Oxford University Computing Services* (OUCS),
  - *University Centre for Computer Corpus Research on Language* (UCREL) la Universitatea Lancaster
  - *Research and Innovation Centre* al Bibliotecii Britanice.

Slide 9

Corpusuri naționale (în Europa)

Corpusul național ceh

- Adresa: http://ucnk.ff.cuni.cz/english/struktura.php
- Volum: 300 milioane de cuvinte
- Autori: Institutul Corpusului Național Ceh (ICNC), Facultatea de Arte și Centrul de Lingvistică Matematică de la Universitatea Charles din Praga.
- Corpusul conține 40% ficțiune, 27% literatură tehnică și 33% jurnalism. Acoperă trei perioade de timp succesive dar recente (este un corpus sincronic).

Slide 10

Corpusuri naționale (în Europa)

Corpusul național polonez

- Adresa: http://nkjp.pl/
- Volum: 1,8 miliarde de segmente lexicale
-Autori (2007-2011):
  Institutul de Informatică,Varşovia
  Institutul Limbii Poloneze, Cracovia
  Universitatea din Lodz
  Editura Tehnică Poloneză (*Polish Scientific Publishers*)

După 2011, întreținerea corpusului e asigurată de
  Institutul de Informatică,Varşovia
  Institutul Limbii Poloneze, Cracovia

Slide 11

Corpusuri naționale (în Europa)

Corpusul național croat (HNK)

- Adresă: http://www.hnk.ffzg.hr.
- Volum: 101,3 milioane de unități lexicale
-Autori:
  - Facultatea de Filozofie, Universitatea din Zagreb (coordonator)
  - Facultatea de Informatică, Universitatea din Zagreb
  - Societatea pentru tehnologia limbii croate

HNK este o colecție sistematizată de texte scrise selectate pentru a acoperi diferite medii, genuri literare, stiluri, domenii și subiecte. Corpusul este însoțit de informație lingvistică și nelingvistică. Este public, disponibil pentru cercetare, educație și alte scopuri non-comerciale

Slide 12

Corpusuri naționale (în Europa)

Corpusul de referință german

Adresă: http://www.ids-mannheim.de/

Volum: 25 miliarde de cuvinte (15 sept 2014) (cel mai mare corpus național cu acces public)

Autori: Institutul pentru Limba Germană (IDS)
CRG (Deutsches Referenzkorpus, DeReKo) este o arhivă electronică de corpusuri textuale în limba germană contemporană scrisă. Alcătuirea lui a început în 1964 și este găzduit de IDS, unde arhiva este actualizată și extinsă continuu.

Slide 13

Alte corpusuri naționale (în Europa)

Corpusul național rus

Pagina web: http://ruscorpora.ru

Cuprinde peste 500 de milioane de cuvinte. Realizat de o echipă de 105 lingviști, informaticieni, programatori de la 18 universități și companii de software.

Corpusul național bulgar

Pagina web: http://ibl.bas.bg/en/BGNC_en.htm

Corpusul conține aproximativ 450 de milioane de cuvinte.

Realizat de Centrul de Lingvistica computațională al Academiei Bulgare de științe (parțial adnotat, bazat aproape exclusiv pe internet).

Corpusul național turc

Pagina web: http://www.tnc.org.tr

Este proiectat să reflecte limba contemporană (1990-2009). Deocamdată, conține 50 de milioane de cuvinte.

Realizat de o echipă de 10 cercetători de la Universitatea Mersin ajutați de 153 de voluntari (studenți informaticieni și lingviști).

Slide 14

Alte corpusuri naționale (în Europa)

Corpusul de referință al limbii portugheze contemporane

Pagina web: http://www.clul.ul.pt/en/resources/183-reference-corpus-of-contemporary-portuguese-crpc

Acesta este cel mai mare corpus de limbă portugheză existent. Conține 311,4 milioane de cuvinte distribuite în texte scrise (309 milioane de cuvinte).

Corpusul a fost realizat începând cu anul 1988 de o echipă formată din 14 cercetători și cadre didactice.

Corpusul (diacronic) de limbă franceză

Pagina web: http://www.frantext.fr

Corpusul conține 271,599,218 cuvinte în 4515 texte din secolul al XII lea până în secolul al XXI lea.

A fost construit începând din anii 1970 de INaLF (ulterior ATILF-CNRS) – *Trésor de la Langue Française*.

Slide 15

Corpusuri românești existente

- ROMBAC: http://www.meta-net.eu/meta-share
- **RASC** (Romanian Anonymous Speech Corpus): http://rasc.racai.ro
- **RSS** (Romanian Speech Synthesis Corpus)

Slide 16

Exemplu de propoziție adnotată

Realizarea corpusului de limbă română contemporană este o obligație culturală. (78 caractere)

```
<seg lang="ro">
<s id="id_temp.1">
<w lemma="realizare" ana="Ncfsry" chunk="Np#1">Realizarea</w>
<w lemma="corpus" ana="Ncmsoy" chunk="Np#1">corpusului</w>
<w lemma="de" ana="Spsa" chunk="Pp#1">de</w>
<w lemma="limbă" ana="Ncfsrn" chunk="Pp#1,Np#2">limbă</w>
<w lemma="român" ana="Afpfsrn" chunk="Pp#1,Np#2,Ap#1">română</w>
<w lemma="contemporan" ana="Afpfsrn"
                        chunk="Pp#1,Np#2,Ap#1">contemporană</w>
<w lemma="fi" ana="Vmip3s" chunk="Vp#1">este</w>
<w lemma="un" ana="Tifsr" chunk="Np#3">o</w>
<w lemma="obligație" ana="Ncfsrn" chunk="Np#3">obligație</w>
<w lemma="cultural" ana="Afpfsrn" chunk="Np#3,Ap#2">culturală</w>
<c>.</c>
</s></seg>
```

(693 caractere)

Slide 17

Ce promitem în cadrul acestui proiect?

- Crearea unui portal cu acces liber, necomercial, oferind facilități de căutare și prelucrare în corpus.
  - Utilizatorii vor avea acces la date statistice, respectiv fragmente de text, ce respectă criteriile de căutare.
- Pe Portalul Corpusului toți colaboratorii vor fi menționați.
- Crearea de instrumente pentru diversificarea corpusurilor de limbă română (de pildă, din perspectivă diacronică).
- Întreținerea corpusului (actualizare, dezvoltare, corectare, etc.) pe o perioadă nedefinită.

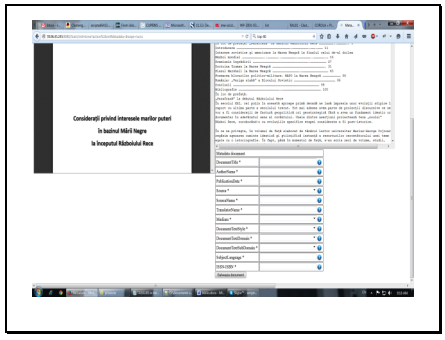
Slide 18

Cum poate fi susținut acest proiect?

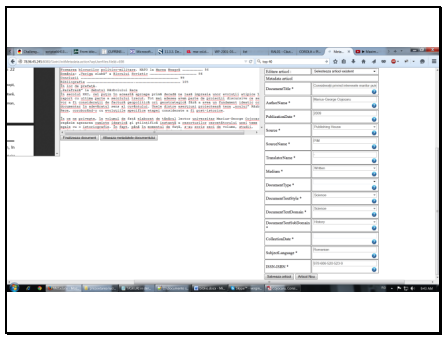
- Furnizarea textelor
- Practică
- Voluntariat
- Promovare

- Manual de lucru
- Lucru de acasă sau de la sediul ICIA

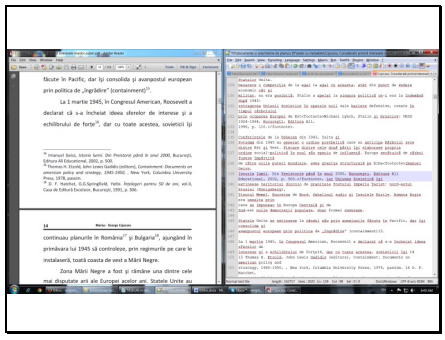
Slide 19



Slide 20



Slide 21



Slide 22

Furnizori de texte

- **Humanitas**
- **Polirom**
- **Editura Academiei Române**
- **Editura Economică**
- **România literară**
- **Muzica**
- **Actualitatea muzicală**
- **DCNEWS**
- **Destine literare**
- **Editura PIM**
- **Editura ADENIUM**
- **Editura DOXOLOGIA**
- **Editura Institutul European Iași**
- **EDITURA GAMA**
- **Simona Tache (blogger)**
- **Dragoș Bucurenci (blogger)**

Slide 23

Voluntari

- http://78.96.45.245/?page_id=671
- Studenți și masteranzi de la Facultatea de Electronică (UPB), de la Facultatea de Litere (UAIC)

Slide 24

Mulțumesc!