https://doi.org/10.52505/filomod.2023.17.23
CZU:81’322

TEHNOLOGII INFORMATICE ABORDÂND LIMBAJUL

Dan CRISTEA

ORCID: https://orcid.org/0009-0009-3362-3304
Institutul de Informatică Teoretică, Filiala Iași a Academiei Române Școala doctorală a Facultății de Informatică, Universitatea „Alexandru Ioan Cuza” din Iași

Rezumat. Încerc în această lucrare o trecere în revistă a domeniului tehnologiilor limbajului, aducând de asemenea în discuție și resursele ce alimentează cercetarea din acest domeniu, cu o aplecare specială asupra limbii române. Am evitat să intru în detalii tehnice sau de natură informatică, preferând o prezentare intuitivă, pentru lingviști și filologi în general. Scopul declarat al articolului este strângerea legăturilor dintre lingviști și informaticieni de o parte și de cealaltă a Prutului, în vederea antamării unor colaborări în beneficiul limbii române.

Cuvinte-cheie: lingvistica computațională, prelucrarea limbajului uman, tehnologiile limbajului, resurse lingvistice, corpusuri, wordnet, dicționare electronice.

Computer Science Technologies Approaching Language

Abstract. In this paper, I try to review the field of language technologies, bringing also into discussion the resources that feed the research in this field, with a special focus on the Romanian language. I have avoided going into technical or computer-related details, preferring an intuitive presentation that would focus a reader with a background in linguistics or philology, in general. The declared aim of the work is to strengthen the ties between linguists and computer scientists on both sides of the river Prut, that would trigger common approaches and efforts to increase the Romanian language.

Keywords: computational linguistics, human language processing, language technologies, linguistic resources, corpora wordnet, electronic dictionaries.

Introducere. Limbajul uman invită informaticienii la formalizări îndrăznețe, care merg de la imaginarea unor structuri de date capabile să suporte tipurile de acces necesare interpretărilor lingvistice până la generarea de modele de calcul capabile să reproducă pe calculator prelucrări care mimează procesele ce se petrec în mintea oamenilor când utilizează limbajul.

Din capul locului aș face o paralelă, de altfel ușor observabilă și care, cu siguranță, a mai fost remarcată și de alții. Dintre toate formele de manifestare a artelor, literatura se aseamănă cu muzica, prin esența lor secvențială. Ca și în muzică, unde portativul afișează o secvență de note, o carte afișează o secvență de litere. Dar, la fel cum doar interpretarea unui muzician transformă notele în muzică, acestea influențându-ne starea de spirit, la fel, doar lectura, care poate fi realizată de o minte de om, dă viață textului, transformându-l în reprezentări cognitive. Citirea unui portativ și a unei pagini de carte nu poate fi făcută decât în secvența notațiilor, note muzicale sau litere. Pe de altă parte, un tablou, o sculptură, o creație culinară, un parfum etc., adică producții ale altor genuri de artă, pot fi „citite" într-o ordine aleatorie, senzorii noștri vizuali, olfactivi ș.a.m.d. fiind impresionabili într-o ordine care nu urmează un traiect bine-stabilit. Și totuși, deși secvențialitatea reprezintă o evidentă îngrădire a libertăților de expresie ale celor două arte, la care se adaugă multe alte genuri de constrângeri, care țin de armonie, ritm etc. într-un caz, și cel puțin de sintaxă, convenții de stil și de conținut – în celălalt, creațiile acestor două arte nu au limite în diversitatea lor.

Invitat în această conferință a literaților, eu fiind un informatician nedeprins decât ca autodidact, deci mai mult ca probabil – imperfect, cu subtilitățile limbajului, voi risca să prezint ochilor dumneavoastră o poveste a meseriei noastre din punctul de vedere al cercetătorului și uneori chiar al creatorului. Examinând cercetările mele, am dat de o lucrare mai veche, din 2005 (Cristea, 2005). Răsfoind-o, mi-am dat seama că partea ei de prezentare generală a domeniului este încă actuală și oglindește destul de bine ideile mele de acum asupra domeniului ce ar putea să îi intereseze pe filologi, de aceea, în acord cu organizatorii, cărora le-am cerut voie să procedez în această manieră, am hotărât să-i folosesc structura și să o actualizez în privința noutăților.

Așadar, voi încerca în această lucrare să prezint sorgintea preocupărilor de lingvistică computațională și prelucrare a limbajul uman, menționând în relativă grabă, din cauza limitărilor de spațiu, o seamă de realizări pe care le-am considerat cele mai interesante pentru segmentul de cititori căruia i se adresează textul meu.

* * *

Lingvistica computațională și tehnologiile limbajului uman. *Lingvistica computațională* (LC), care definește osatura teoretică a domeniului, și sora ei cu predispoziție pentru experiment și aplicații, *Tehnologia limbajului uman* (TLU) sau *al limbajului natural* (TLN), se preocupă de studiul limbajului uman din două perspective: unul teoretic, pentru a adăuga metodelor de

investigare a limbajului, utilizate de lingvistica clasică, aparatul teoretic și de modelare al informaticii și celălalt experimental, pentru realizarea dezideratului ca omul să poată comunica cu mașina în limbaj natural, modul cel mai natural pentru el.

Aplecându-se asupra limbajului uman, cu o rigoare a metodelor moștenită din originea matematică a domeniului, informatica nu-i uzurpă misterul, ambiguitatea ori dimensiunea lui poetică. Ea caută să identifice sursele ambiguului, fără a brusca o claritate artificială atunci când limbajul este inerent ambiguu, și poate descoperi structură și regularitate în zonele unde un poet deslușește doar un desfrâu al imaginației transpus în cuvinte. Infinita libertate a exprimărilor pe care o oferă limbajul, bogăția inepuizabilă a alăturărilor de cuvinte are aparența de a-l face greu formalizabil. El se relevă însă ochiului atent al cercetătorului ca ascunzând coloane arhitecturale de o bizară și minunată regularitate, dezvăluirea lor cu instrumente informatice constituind o provocare pentru mulți dintre noi.

Un prim gen de structură în alambicul divers și aparent dezordonat al șirului de cuvinte e relevat de teoriile sintaxei. Apoi, cum se compune oare sensul exprimării prin îmbinarea sensurilor elementare ale cuvintelor, fiecare în parte atât de labil în semnificații când e rupt de context? Iată o problemă pentru semanticieni. În fine, cum se adună sensurile propozițiilor în povestea de ansamblu comunicată de text, care sunt legile ce fundamentează discursul? Acestea sunt întrebări cărora le caută răspuns teoriile computaționale ale discursului. Investigația modernă asamblează, la fiecare nivel, ingredientele teoretice cu cele experimentale, facilitate de mijloacele informatice de prelucrare. În plus, cercetarea contemporană în domeniu căpătă o tot mai însemnată pondere aplicativă.

Istoric și prezent. Istoria domeniului se întinde pe aproximativ 70 de ani, începutul lui putându-se identifica încercărilor de modelări ale limbajului din *lingvistica matematică* (Marcus, 1966). Pentru unii însă, începutul domeniului se identifică cu publicarea de către Warren Weaver în 1949 (Locke and Booth, 1955) a unui memorandum adresat oamenilor de știință americani, în care prezicea că traducerea automată este posibilă și care a produs imediat o emulație în cercetarea americană și mai apoi în cea sovietică și vest-europeană pentru abordarea de probleme dintre cele mai diverse, precum: cercetări lexicografice asistate de calculator, modele de „traducere directă”, analiza sintactică, analize statistice asupra limbajelor naturale, propuneri de reprezentări interlinguale, gramatici de dependență etc. După aproximativ 15 ani de studii și experimente în traducerea automată, rezultatele au fost totuși

puțin satisfăcătoare, *raportul ALPAC*[1], apărut în Statele Unite, descurajând cercetările în traducere automată pentru cel puțin un deceniu. Deși ajuns într-un impas, domeniul a continuat să se dezvolte prin cercetări colaterale scopului inițial, care era cel al traducerii automate. Progrese în mai multe arii au contribuit la relansarea traducerii automate la mijlocul deceniului 70 și realizarea sistemului Systran, folosit o perioadă îndelungată de Comisia Europeană. Ulterior, domeniul traducerii automate a evoluat atât în direcția procesării simbolice, atractivă pentru lingviști, care pot regăsi în aceste metode propriile intuiții lingvistice, cât și în cea a procesării statistice și, mai nou, al metodelor de învățare automată, cu precădere învățarea de profunzime (*deep learning*).

Eforturile de cercetare întreprinse inițial în acest domeniu au fost, mai întâi, de natură teoretică, potențate de necesitatea de a proba teoriile elaborate asupra limbajului natural prin metode cantitative. Odată cu răspândirea internetului au apărut însă imediat și interese comerciale, cum ar fi cel al creării și exploatării aplicațiilor care utilizează limbajul uman. Din ce în ce mai mult este dorit un computer capabil să „înțeleagă" conținutul documentelor, inclusiv a nuanțelor contextuale. Tehnologia poate apoi să extragă cu precizie informațiile conținute în documente, să clasifice și să organizeze documentele în sine. Aplicațiile informatice încep din ce în ce mai mult să se deplaseze din calculatorul aflat pe biroul clientului în telefonul lui portabil sau în dispozitivul mobil (tabletă sau laptop, conectate la rețelele de comunicații fără fir). În momentul de față clienții pot să facă uz de propoziții în locul cuvintelor izolate. Un exemplu de sector public deschis comunicațiilor în limbaj natural este cel al turismului, în care aplicațiile permit efectuarea de rezervări, dar intermediază și vânzarea pachetelor de vacanță printr-un dialog din care sistemul este capabil să se informeze asupra preferințelor solicitantului, dialogul desfășurându-se în limba maternă a clientului. Dar cele mai multe aplicații ale domeniului TLU încep să capete contur în exploatarea eficientă a oceanului informațional care este azi *web*-ul. Recent se observase că timpul de căutare a informațiilor era cel puțin la fel de mare ca și cel dedicat folosirii lor productive și aceasta în situația în care o bună parte din informația cunoscută azi se găsește într-o formă sau alta ca text pe *web*. Utilizatorii de azi ai *web*-ului sunt doritori să obțină informații, adresând întrebări în limba proprie, asupra oricărui domeniu ori sferă a cunoașterii. Mai mult chiar, opinia că sectorul care va rezista un timp indefinit de lung asaltului tehnologiilor informatice, el fiind apanajul exclusiv al oamenilor, anume zona de creativitate umană, a căpătat recent

[1] A se consulta, de exemplu: https://blog.pangeanic.com/alpac-report

o corecție majoră. În doar câteva luni, programe generative de inteligență artificială, gen ChatGPT[2], se dovedesc capabile să înțeleagă[3] și să genereze texte care nu numai că sunt aproape fără erori de gramatică, dar surprind prin coerență, bogăția de idei, diversitate. ChatGPT, de exemplu, permite utilizatorilor să solicite informații despre subiecte extrem de diverse, de la calculul cuantic și originea speciilor în teoria lui Darwin până la mecanica inimii umane sau principiile decorațiilor interioare. Această aplicație, precum și altele din aceeași categorie, cum ar fi, de exemplu, Microsoft Copilot[4], sunt capabile să genereze cod program plecând de la o cerere exprimată în limbaj natural, în felul acesta crescând de multe ori productivitatea firmelor de soft. ChatGPT poate, de asemenea, să compună poezii și muzică, să scrie scenarii de film ori povestiri pe subiecte pe care i le cere utilizatorul. Poate adăuga umor textelor generate sau poate adopta un ton formal. Asta nu înseamnă însă că limbajul natural a fost cucerit în totalitate de mașină. Programele generative de inteligență artificială consolidează în rețelele lor neuronale enorm de multă informație culeasă din oceanul mondial al internetului, însă, atunci când cererea depășește limitele lor de cunoaștere programele halucinează. Limbajul generat în aceste cazuri este cursiv și are aparența coerenței, dar este ilogic sau include generații false.

Foarte recent se poate identifica și un interes legat de dimensiunea păstrării identității limbilor naționale în societatea modernă. Contrar tendinței firești de uniformizare ce se presupune a fi caracteristica unei societăți globalizate politic, economic și informațional, limbile naționalităților mici și mijlocii nu sunt și nu trebuie să fie în pericol de diminuare ori dispariție prin asaltul limbilor mari, care sunt actualmente mai viguroase în comunicațiile planetare. Dimensiunea prezenței electronice în *web* a unei limbi a devenit o măsură a utilizării ei: o limbă este cu atât mai importantă cu cât este mai răspândită în mediile electronice și de aceea preocupările de a crea resurse lingvistice electronice în cantități cât mai mari și cele de dezvoltare a tehnologiei de prelucrare a limbii proprii sunt extrem de active azi în cadrul fiecărei comunități lingvistice.

Resursele lingvistice. Mesajele vorbite, înregistrate, sau cele tipărite reprezintă *surse lingvistice*, atunci când sunt utilizate pentru studiul limbii.

---

2 Creație a companiei OpenAI, accesibil la https://chat.openai.com/.

3 Cu sensul de a răspunde adecvat la o solicitare.

4 A se vedea, spre exemplu, cursul *Introducing GitHub Copilot: Your Brilliant AI Pair Programmer!* ținut de Florin Olariu în 13 septembrie 2023, în cadrul Școlii EUROLAN-2023, Bălți: https://eurolan.info.uaic.ro/2023/wp-content/uploads/2023/09/Florin-Olariu.zip

Ori de câte ori acestea sunt reprezentate într-un format electronic, ele sunt însă numite *resurse lingvistice*.

În principiu, orice informație de natură simbolică, statistică sau neuronală ce poate fi depozitată în calculator și care definește specificul unei limbi intră în categoria resurselor lingvistice. Astfel, pot fi considerate resurse:

- 0pcorpusurile lingvistice primare sau adnotate;
- dicționarele în format electronic și tezaurele lexicale digitale;
- modelele de limbă, care pot fi: colecții de reguli gramaticale, câmpuri de probabilități, frecvențe de apariție a n-gramelor, vectori sau tensori de valori, rezultați din antrenarea pe texte a modelelor neuronale etc.;
- colecții de imagini de pagini de documente vechi adnotate de experți sau de tehnologii lingvistice și multe altele.

Apariția corpusurilor electronice, cu toată tehnologia aferentă de adnotare și interpretare a lor, a provocat o adevărată cotitură în tehnologia limbajului. Definite de către John Sinclair[5] în anii ’60 ca fiind referențiale pentru studiul limbajului (Sinclair, 1987), corpusurile au fost utilizate la început pentru evidențierea faptică a intuițiilor lingvistice ale cititorilor. Ulterior, urmare a dezvoltărilor din domeniul învățării automate, corpusurile au căpătat valențe noi, fiind folosite ca surse de cunoștințe pentru agenți informatici specializați pe prelucrarea limbajului. Actualmente, există și continuă să se dezvolte programe din ce în ce mai sofisticate, care, utilizând precepte ale inteligenței artificiale, sunt capabile să învețe din corpusuri adnotate adecvat la diferite niveluri: fonologic, morfologic, lexical, sintactic, semantic, de discurs etc.

Orice teorie lingvistică azi trebuie susținută de o evaluare care să fie ancorată într-un corpus semnificativ de exemple. Corpusurile și metodele de evaluare a tehnologiilor dedicate limbii au devenit atât de importante în lingvistica contemporană încât, de mai mult timp, se organizează importante conferințe internaționale orientate pe această problematică. Un exemplu îl constituie seria de conferințe bienale LREC (*Language Resources and Evaluation Conference*), lansată în 1998 de Antonio Zampolli[6].

---

[5] John Sinclair (1933–2007) a fost pentru mult timp profesor la Universitatea Birmingham și directorul seriei de dicționare Collins COBUILD. El este creatorul conceptului modern de investigație lingvistică bazată pe corpus, a construit un imens corpus al limbii engleze moderne, a propus și dezvoltat instrumente și metode de lucru pe corpusuri. În ultima perioadă a activității lui a creat și condus *Word Tuscan Center*, un centru de educație în lingvistica corpusului, la Montecatini, Italia.

[6] Antonio Zampolli (1937–2003) a fost o personalitate remarcabilă a domeniului Lingvisticii Computaționale, inițiatorul școlilor de vară de la Pisa, ce au deschis calea colaborărilor dintre lingviști și informaticieni; a fost timp de 35 de ani director al Institutului de Lingvistică din Pisa.

Intuiția pe care se bazează utilizarea corpusurilor în tehnologiile lingvistice este că un număr limitat de contexte de apariție a unui cuvânt este suficient de reprezentativ pentru a cuprinde în ele sămânța utilizării lui în multe alte situații, așadar, că acestea nu reprezintă doar matricele folosirii lor în exact situațiile exemplelor acumulate în corpus, ci se pot constitui în patternuri generalizabile. Să ne amintim că și marile dicționare, precum *Dicționarul tezaur al Limbii Române* (DLR), așază la baza deciziei de selecție a exemplelor care să însoțească descrierile sensurilor cuvintelor exact aceeași presupoziție.

Corpusurile adnotate sunt cu precădere importante în cercetarea lingvistică, pentru că ele permit evidențierea, în paralel cu textul originar, a fenomenelor lingvistice, explicitate de experți sau, atunci când tehnologia a atins o acuratețe comparabilă cu a expertului uman, chiar de către mașină. Tehnologiile actuale permit separarea, la orice moment, a adnotării de textul originar, care, din motive lesne de înțeles, trebuie să conserve forma inițială.

Există două motive principale pentru care lingviștii informaticieni agreează corpusurile adnotate. În primul rând, expertizele lingvistice codificate în adnotări pot fi transferate programelor, prin procese de învățare. În al doilea rând, pe corpusuri adnotate pot fi căutate exemple ori contra-exemple pentru validarea/invalidarea teoriilor, pentru verificarea ipotezelor lingvistice sau pentru determinarea performanțelor sistemelor de prelucrare automată.

Iată un posibil scenariu de cercetare în lingvistica teoretică: lingvistul, ca urmare a observării unui număr de exemple de limbă pe care le consideră semnificative, are o intuiție asupra unui fenomen lingvistic și elaborează o teorie. Acele exemple, culese ici și colo din literatură, sunt marcate manual pe text, într-o format care împiedică confundarea textului originar cu adnotările adăugate. Rezultă astfel un prim corpus, în care manifestările teorii sunt puse explicit în evidență, urmând ca din el un program să poată extrage, prin învățare și generalizare, trăsăturile relevante ale teoriei, adică elaborând un așa-zis *model lingvistic*. Acesta va putea apoi fi utilizat pentru marcarea automată a fenomenului lingvistic pe exemple noi. Pe noile exemple puse în evidență de program, lingvistul își va putea apoi valida ori invalida teoria. Procesul poate să aibă un caracter iterativ, prin corecții și îmbunătățiri succesive, până la finisarea teoriei și statornicirea unui model care să satisfacă.

Astfel, un program de etichetare la parte de vorbire, de exemplu, va utiliza un corpus în care fiecare cuvânt este adnotat manual la parte de vorbire. Corpusul este, de regulă, împărțit în două secțiuni, una din care programul învață să producă aceeași gen de etichetare, și a doua, de dimensiune mai mică, pe care se verifică performanța programului. Pentru a fi eficiente, corpusurile trebuie să fie de mari dimensiuni. O adnotare de calitate este însă o operație

costisitoare, pentru că presupune elaborarea unui manual de adnotare adresat adnotatorilor umani, necesită resurse umane calificate și consumă timp. În plus, pentru a obține o bună acuratețe a modelului inferat de program, de obicei doi sau chiar trei adnotatori trebuie să lucreze separat asupra aceluiași text, rezultatele fiind apoi comparate și cazurile în care apar diferențe de opinie negociate. De aceea corpusurile adnotate sunt și dorite și urâte de cercetători.

O caracteristică a sistemelor moderne de prelucrări textuale o constituie separarea completă a codului program de detaliile fenomenelor lingvistice pe care le tratează. În general, programul dă viață unui algoritm general, care trebuie să fie ghidat în toate acțiunile lui de resurse lingvistice, externe lui, care descriu anumite fenomene lingvistice și sunt specifice unei anumite limbi. Doar alimentarea lui cu o resursă specifică îl face capabil să lucreze pe o limbă sau alta. Astfel, algoritmul de etichetare la parte de vorbire lucrează la fel, indiferent de limba pe care o procesează, ceea ce îl face aplicabil limbii engleze, cehe sau române fiind doar modelul corespunzător de limbă pe care îl accesează, învățat din corpusuri specifice.

Un exemplu de tezaur lexical extrem de larg utilizat în prezent în cercetare ca și în aplicații este WordNet[7], dezvoltat în Laboratorul de Științe Cognitive al Universității Princeton de către un colectiv condus de regretatul profesor George Miller. Inspirat de teoriile psiholingvistice moderne asupra memoriei umane, WordNet organizează substantivele, verbele, adjectivele și adverbele limbii engleze în serii sinonimice, numite **synset**-uri, fiecare reprezentând modalități de realizare lexicală în limba engleză a conceptului ce abstractizează înțelesul comun al seriei (Fellbaum, 1998). Anumite synset-uri pot fi legate prin relații de natură semantică, dintre care cele mai importante, aplicabile doar substantivelor și verbelor, sunt cele de hipernimie și hiponimie[8].

Ulterior dezvoltării Princeton WordNet au apărut din ce în ce mai multe *wordnet*-uri și pentru alte limbi. Astfel, în cadrul proiectului European EuroWordNet[9] (Vossen, 1998) au fost create wordnet-uri[10] pentru încă 10 limbi europene și, pentru prima dată, aceste rețele lexicale au fost aliniate între ele. Soluția tehnică pentru corelarea multilinguală a rețelelor semantice monolingve a fost definirea unui index lexical interlingual (ILI) ca subset al mulțimii de synset-uri a limbii engleze, care, considerat a fi independent de limbă, conține reprezentări conceptuale ale înțelesurilor lexicalizabile în

---

7 https://wordnet.princeton.edu/

8 Un concept lexical *C1* este hipernim al unui concept *C2* dacă *C1* este mai general decât *C2*. În acest caz, spunem că *C2* este un hiponim al lui *C1*.

9 http://www.illc.uva.nl/EuroWordNet/

10 Utilizat acum în toată lumea, numele acestui proiect este actualmente folosit ca nume comun.

limbile proiectului. Fiecare înțeles din oricare din limbile reprezentate în rețeaua semantică multilingvă este pus în corespondență, în general, cu un unic concept al indexului interlingual. Synset-urile din două sau mai multe limbi cărora le corespund același concept din ILI sunt considerate echivalenți de traducere. O altă inovație a proiectului EuroWordNet a fost adoptarea unei mulțimi de primitive semantice, independente de limbă, în termenii cărora unei subclase a conceptelor din ILI, considerată a conține **concepte de bază**, i s-a asociat descrieri ontologice. Prin importul acestor descrieri la nivelul lexicalizărilor (și, prin moștenire, la hiponimii acestora) în fiecare dintre wordnet-urile monolingve, EuroWordNet a devenit practic o ontologie lexicală multilingvă.

Proiectul Balkanet (Tufiș et al., 2004) a creat o nouă rețea de wordnet-uri pentru cinci limbi balcanice (turcă, greacă, bulgară, română și sârbă) toate aliniate cu Princeton WordNet 2.0. El aduce o seamă de noutăți, ca, de exemplu, definirea principiului păstrării ierarhiei (Tufiș și Cristea, 2002), care postulează că lanțurile de relații ierarhice se păstrează între limbi, chiar dacă ele nu prezintă peste tot aceeași densitate de concepte.

Tehnologiile limbajului natural scris:

a) Prelucrări sub-propoziționale. Prelucrările sub-propoziționale se realizează la nivelul cuvântului sau al grupurilor de cuvinte mai scurte de o propoziție. Analiza morfologică are ca obiectiv identificarea trăsăturilor morfologice ale cuvintelor flexionate în contextul apariției lor în propoziție. Fiecărui cuvânt flexionat din șirul de intrare i se asociază forma de bază (lema) și un set de trăsături ce sunt specifice părții de vorbire.

Un cuvânt flexionat, când e scos din context, în general este ambiguu din punct de vedere morfologic. O analiză morfologică ruptă de context e capabilă să pună în evidență toate posibilitățile de interpretare. Analize alternative sunt posibile datorită omonimiei categoriale, a alternanțelor morfologice finalizate identic, a funcțiilor multiple ale afixelor ori a granițelor incerte dintre morfeme în cuvintele compuse. Adesea, doar plasarea cuvântului în contextul restului cuvintelor propoziției poate realiza dezambiguizarea totală a nivelului morfologic. Această operație se numește cel mai adesea etichetare la parte de vorbire (*POS tagging*), deși rezultatul ei este mai complex decât determinarea neambiguă a părții de vorbire.

Tot în categoria prelucrărilor sub-propoziționale se plasează și parsarea[11] de suprafață (*shallow parsing*). Ieșirea unui astfel de *parser* evidențiază granițele anumitor grupuri, cum ar fi cele nominale nerecursive (grupuri ce nu conțin alte grupuri nominale) sau grupuri verbale, ca subconstituenți, fără a indica

[11] Un parser este un program capabil să extragă din text o structură pe un anumit nivel.

constituția lor structurală sau rolul lor în propoziție. Analiza de suprafață este dominată în prezent de metode computațional rapide, ca de exemplu analiza prin expresii regulate. Punerea în evidență a grupurilor sintactice constituie, de obicei, o primă etapă într-un lanț de prelucrare mai elaborat, cum ar fi, de exemplu, o analiză sintactică completă, determinarea rolurilor sintactice ale grupurilor nominale sau prepoziționale în jurul verbelor, rezoluția anaforelor etc.

b) Prelucrări sintactice. Prelucrările sintactice au ca scop determinarea structurii de constituenți sintactici ai propoziției/frazei. Raporturile existente între constituenții propoziționali sunt descrise ca seturi de constrângeri sintactico-semantice. O constrângere pozițională cu implicații asupra sensurilor cuvintelor este, de exemplu, cea dintre un adjectiv și substantivul pe care îl determină, în limba română. Plasarea adjectivului în fața sau în spatele substantivului poate, pentru anumite clase de adjective, determina sensul acestuia (comparați *o singură femeie* cu *o femeie singură*) [Cornilescu, 2004].

La baza prelucrărilor sintactice stau gramaticile. Elaborate inițial ca seturi neformale de reguli de constituire a compușilor, treptat, gramaticile capătă descrieri din ce în ce mai formale ce adaugă capacității lor de a descrie fenomenele sintactice complexe ce există în limbă și menirea de a servi drept suport parserelor sintactice. Din acest punct de vedere, pentru că notează cunoașterea gramaticală independent de un algoritm anumit de procesare, gramaticile, ca seturi de reguli dublate de constrângeri, se constituie în resurse lingvistice, la fel ca și corpusurile. În anumite cazuri, regulile de analiză gramaticală și constrângerile pot fi inferate din corpusuri adnotate la structura sintactică. Codificarea fenomenelor gramaticale poate fi făcute sub formă simbolică sau numerică. Diferența dintre cele două tipuri de notații este că, în primul caz, ea evidențiază cu claritate o semantică asociată regulilor, pe când în cel de al doilea, utilizat acum aproape exclusiv, noțiunea de regulă își pierde semnificația de notație simbolică explicită, ea fiind dizolvată într-un sistem de ponderi care influențează deciziile luate de o rețea neuronală. Indiferent de maniera de exprimare a cunoașterii gramaticale, o analiză sintactică trebuie să rezulte în construirea unei structuri care să exprime agregarea recursivă a cuvintelor în compuși și a acestora în propoziții și fraze. În sistemele robuste de analiză sintactică, anumite constrângeri pot fi încălcate fără a invalida în totalitate o analiză. Astfel, deși nerecomandate, greșeli gramaticale, precum încălcarea acordului subiect-predicat, anacolutul etc. pot fi permise, sistemul de analiză putând fi antrenat să semnaleze greșelile dar să accepte totuși exprimarea.

Astfel de sisteme de parsare își pot forma, din exemple adnotate, constrângeri similare regulilor sintactice sau semantice. De exemplu, unui verb tranzitiv i se pot asocia constrângeri de tipuri semantice referitoare la obiectele directe pe care le poate accepta.

c) Prelucrări lexical-semantice. La nivelul lexical-semantic cele două probleme fundamentale sunt descoperirea sensurilor cuvintelor în context și compunerea înțelesului exprimărilor din înțelesul cuvintelor componente.

Prima problemă, cunoscută sub numele de *dezambiguizarea sensurilor cuvintelor* (*word sense disambiguation – WSD*), presupune cunoscut un inventar al sensurilor cuvintelor, corespunzător unui dicționar sau unui tezaur lexical, și își propune să determine sensurile cuvintelor, în context, în conformitate cu acest inventar.

Compunerea (sau calculul) înțelesului exprimărilor este o problemă care atinge aspecte extrem de subtile și de greu de formulat riguros, pentru că limbajul natural abundă în interpretări metaforice, metonimice, ironice, în implicații conversaționale specifice cauzate de contexte pragmatice diferite etc. Dacă ne limităm la aflarea înțelesului intrinsec al exprimărilor, nealterat de fenomene ca cele menționate mai sus, cele mai multe abordări plasează acest efort în sfera logicii, în încercarea de a reduce calculul semantic la posibilitatea cunoașterii condițiilor în care o propoziție ar putea fi adevărată (ceea ce este altceva decât a ști dacă o propoziție este ori nu adevărată). Desigur odată pusă la punct o logică compozițională a condițiilor de adevăr aplicabilă aserțiunilor, ea trebuie să suporte generalizări pentru a fi aplicabilă întrebărilor și imperativelor, dar ideea este de a plasa deasupra ei un sistem inferențial bazat pe condiții de adevăr. Cu un astfel de sistem și o logică care calculează înțelesul compușilor sintactici plecând de la înțelesul constituenților, sensurile cuvintelor se pot combina în înțelesuri ale propozițiilor[12], iar din acestea se pot infera concluzii și supoziții. Primul model combinatoric aplicat limbajului natural care se îndepărtează de *logica predicatelor de ordinul întâi* (LPOI) a fost *logica intensională* a lui Montague (1973). Ulterior, modele mai sofisticate au încercat să ofere soluții computaționale și pentru exprimări în care apăreau concepte, cuantificatori și cuvinte imposibil de formalizat în LPOI, precum necesitatea, posibilitatea, verbe ca *a crede* sau *a ști*, expresii precum *cei mai mulți*, *mai mult de jumătate* etc.

În multe mesaje schimbate între oameni, conținutul strict informațional este codificat pe o purtătoare care transportă o încărcătură emoțională. Ca urmare, se pune problema dacă preocuparea de a descifra conținutul semantic al mesajului trebuie să fie completată cu străduința de a trezi în mașină o reacție la încărcătura lui emoțională. Poate această componentă rezona în vreun fel în mașină, sau e un balast inutil care va fi, inevitabil, ignorat de mașină? Întrebări de acest fel, cu toată aparenta lor apropiere de domeniul

[12] V. principiul compoziționalității, atribuit lui Frege (Frege, 1892).

science fiction, îi preocupă pe cercetători. Desigur, nu ne interesează să facem calculatorul să „suspine" la un text melancolic, dar implantarea abilității de a recunoaște latura emoțională într-un mesaj în limbaj natural poate ajuta mașina să sesizeze mai ușor intențiile unui interlocutor uman într-un dialog cu acesta și îi pot, de asemenea, atribui performanța de a produce texte care să producă emoții într-un receptor uman. Un comportament care simulează sensibilitate din partea mașinii poate face dialogul dintre om și mașină mult mai natural, partenerul uman putând chiar să ignore că vorbește cu un calculator, un obiect insensibil și apatic, ceea ce îl va face mult mai deschis la colaborare și îi va insufla încrederea că poate găsi în interlocutor, om ori mașină, soluțiile cele mai adecvate la cererile lui. Această latură a domeniului PLU se numește *analiza sentimentelor* și reunește eforturi intense, orientate spre scopuri pragmatice, cum ar fi recunoașterea în comentariilor consumatorilor a reacțiilor acestora asupra unor produse industriale, ceea ce ar putea influența designul, marketingul etc.

d) Structura discursului. Atunci când înțelesul unei propoziții nu poate fi recuperat integral din calculul atașat elementelor constitutive ale ei, deci când e nevoie să depășim granița de propoziție pentru a compune un înțeles, trecem în domeniul discursului. Privite în contextul mai larg al lingvisticii computaționale, teoriile discursului își încep investigația după ce morfologia computațională, sintaxa computațională ori semantica computațională și-au adus fiecare în parte aportul. Când structura compozițională a cuvântului și cea a frazei sunt descifrate, când sensurile cuvintelor din frază sunt știute și avem o reprezentare a modului în care aceste sensuri elementare se adună în cel global, al frazei, mai avem încă de elucidat, prin compunerea înțelesurilor frazelor constitutive, care este mesajul intenționat să-l transmită autorul textului și cum a reușit acesta să ne convingă de ceea ce a avut în intenție să ne transmită. Același lucru poate fi exprimat în multiple feluri. De ce alegem o formă și nu alta? Ce anume contribuie la coeziunea unui text, ce face un text să fie coerent? Putem utiliza un pronume (în general, o expresie referențială) oriunde? Care este legătura dintre structura de discurs și referențialitate? Acestea sunt întrebări la care încearcă să răspundă teoriile discursului.

Este important ca interpretarea discursului să fie realizată respectând secvența elementelor lui. Dacă structura sintactică poate fi elucidată, din punct de vedere computațional, ca un puzzle care poate fi început din orice capăt, discursul nu poate fi abordat decât în ordinea emiterii lui, pentru că elementele care-l compun își aduc contribuția în compunerea mesajului în exact secvența emiterii lor, considerarea unei alte ordini putând induce alte semnificații decât cele intenționate. Elementul constitutiv de bază al unui

discurs este *unitatea de discurs*. Cei mai mulți autori acceptă ca unitate de discurs propoziția (*clause*). În general, o propoziție comunică o situație, un eveniment, o stare. Întrucât aceste cărămizi constitutive ale discursului sunt uneori mai fine decât limita maximală a obiectului prelucrării sintactice, care este fraza, este firesc să ne gândim la o anumită redundanță între structurile generate de analiza sintactică și cele produse de interpretarea discursului. Această intersecție interpretativă trebuie să se manifeste ca o corespondență (*mapping*) între structura sintactică aflată între nivelul propozițional și frastic și substructura generată de analiza retorică a discursului la nivel frastic.

Comportamentul incremental al interpretării discursului (ce se impune atunci când la baza modelului interpretativ sunt plasate procesele cognitive ale creierului uman), poate fi obținut printr-o funcție de interpretare care să asambleze unitățile de discurs, într-o structură, în ordinea apariției lor, adică în ritmul lecturii sau al audierii mesajului. Amintim câteva fenomene lingvistico-cognitive care trebuie să-și găsească o modelare în această viziune:

– Care e structura de date care reprezintă cel mai bine discursul și cum lucrează un proces capabil să construiască această structură de reprezentare (un parser de discurs)? Majoritatea teoriilor care abordează problema reprezentării și procesării discursului acceptă ipoteza că structura discursul are o reprezentare arborescentă. În Teoria Structurilor Retorice (Mann și Thompson, 1988), de exemplu, nodurile interioare ale unui arbore de discurs reprezintă relații retorice între întinderi de text, pe când cele terminale reprezintă unitățile de discurs. Într-o modelare care descrie interpretarea discursului ca un proces incremental (Cristea și Webber, 1997), acest arbore este construit prin operații de atașare a unităților de discurs, în ritmul apariției acestora, în structura arborescentă aflate în dezvoltare.

– Cum se tratează expectativele (acele exprimări care obligă la continuări)? De exemplu, după clauza *Deși afară ploua*, plasată imediat după punct, trebuie să existe cu necesitate o urmare pentru ca discursul să nu fie defect. În (Cristea și Webber, 1997) se propune o generalizare a frontierei drepte a arborelui de discurs (Polanyi, 1985), care să permită rezolvarea expectativelor.

– Cum se rezolvă *referințele anaforice* (exprimări care nu pot fi interpretate prin ele însele, ci numai cu ajutorul unor elemente aflate în altă parte a textului)? De exemplu, în fraza „*Nici el nici ea nu agreează minciuna, însă, dintre cei doi părinți, Maria este cea care s-a ocupat de educația copiilor în spiritul adevărului.*” grupul nominal *cei doi părinți* este e expresie anaforică „ancorată” în setul entităților introduse de pronumele *el* și *ea*, în timp ce *ea* fiind o cataforă (referință pronominală care referă o entitate clarificată de textul care urmează), este considerat de unii lingviști a avea ca referent

pe Maria. În (Cristea și Dima, 2001), considerente de ordin psiho-lingvistic sunt argumente pentru un tratament unitar al anaforei și cataforei: indiferent de natura realizării unei entități de discurs (pronominală ori substantivală), prima menționare a ei duce la crearea unei reprezentări care poate apoi să sufere completări.

– Cum poate fi modelat procesul de rezumare a unui text? Nu tot ceea ce citim sau auzim ne rămâne în memorie, dar suntem capabili să producem un rezumat al discursului recepționat. Construcția acestui rezumat se realizează în ritmul însuși al lecturii și, din acest motiv, la orice moment al întreruperii unei lecturi, suntem capabili să rezumăm ceea ce am citit până în acel moment. Teoria nervurilor (Cristea et al, 1998), cu dezvoltările ulterioare Cristea et al, 2005), explică fenomene de coeziune și coerență a discursului prin evidențierea unei legături între structura de discurs și lanțurile referențiale.

Situarea limbii române. Cercetările românești în TLU sunt promițătoare și se aliniază curentelor aflate acum în colimatorul cercetărilor mondiale. Ca și domeniul de cercetare în sine, acestea pot fi grupate în două categorii mari: dezvoltarea de resurse lingvistice românești și dezvoltarea de instrumente de prelucrare a limbii, primele dedicate exclusiv limbii române, cele din urmă, în virtutea paradigmei moderne de separare a codului de date, putând fi generalizabile pentru a fi aplicabile cel puțin limbii române, după cum și invers, aplicațiile dedicate limbii române putând rezulta din particularizarea unor modele anterior antrenate pe alte limbi sau chiar în alte domenii.

Ambele preocupări, dezvoltarea de resurse lingvistice românești și de aplicații dedicate limbii române, sunt, în general, ancorate în proiecte finanțate din surse interne sau externe, uneori însă și prin străduința izolată a unor grupuri de cercetare sau în activități de laborator cu studenții. Mi-ar fi imposibil să fac aici o expunere exhaustivă a acestor preocupări și realizări, dar voi puncta doar câteva dintre ele, cu precădere dintre cele în care am fost implicat direct.

– Achiziționarea și punerea în circulație publică a unor corpusuri de limbă românească: proiectul Corpusului Limbii Române Contemporane COROLA[13] (Cristea et al., 2017; Kupietz et al., 2017), derulat între ani 2013–2017, un proiect al Academiei Române, s-a finalizat prin construirea unui corpus de aproximativ 1 mld. de cuvinte românești de limbă contemporană și 300 de ore de înregistrări vocale, adunate din 17 domenii și mai multe stiluri literare. Textele, oferite proiectului de deținătorii lor de drept și reprezentând romane, articole de presă, bloguri, scrieri științifice, piese de teatru etc., au fost curățate

[13] https://corola.racai.ro/

de elemente non-textuale, segmentate la fraze și cuvinte, acestea fiind adnotate automat cu informații morfosintactice. Interfața principală de acces online a fost oferită de Leibniz-Institut für Deutsche Sprache (IDS-Mannheim). Alte exemple de corpusuri sunt ParlaMintRo (Rebeja et al., 2022) – creat într-un context european (Erjavec et al., 2022) și MARCELL (Coman et al., 2019), care a construit o infrastructură durabilă pentru preluarea și prelucrarea semantică a documentelor legislative (legi, decrete, reglementări etc.) într-un context internațional, ce a inclus colective din Bulgaria, Polonia, România, Slovacia, Slovenia, Ungaria și Croația, scopul principal fiind ameliorarea sistemelor de traducere automată. Corpusul creat de echipa de la Institutul de Cercetări în Inteligență Artificială al Academiei Române din București (AR-ICIA) a prelucrat mai mult de 140.000 de documente, dens adnotate la parte de vorbire, lemă, analiza de dependențe, nume de entități, împreună cu metadatele corespunzătoare.

– Realizarea de tezaure lexicale, cu precădere RoWordNet, amintit deja, activitate derulată pe parcursul anilor 2001-2004, ca parte integrantă a proiectului FP5-IST Balkanet, în cadrul unui consorțiu în care au mai participat cercetători din Turcia, Bulgaria, Grecia, Serbia, Cehia și Franța. Echipele românești participante în proiect au fost cele de la AR-ICIA și Facultatea de Informatică a Universității „Al.I.Cuza" din Iași (UAIC-FII);

– Transpunerea în format electronic a marelui dicționar al Academiei Române a fost începută în cadrul proiectului eDTLR (Moruz și Cristea, 2016) și ea continuă ca temă internă, preluată de colectivele de la Institutul de Informatică Teoretică al filialei Iași a Academiei Române (ARFI-IIT) și de la Institutul de Filologie Română „Al. Philippide" al Filialei Iași a Academiei Române (ARFI-IFRI). Convingerea mea personală este că în afara tuturor beneficiilor evidente, publicarea marelui dicționar pe internet ar fi un efort răsplătit imediat prin revigorarea unui segment semnificativ de limbă uitată (un cetățean comun nu cunoaște mai mult de 30% din cuvintele limbii și putem spera doar ca aceste 30 de procente să nu fie aceleași pentru toată populația de limbă română…). Cu certitudine, segmente semnificative ale lexicului limbii române nu sunt cunoscute publicului larg. Acest inestimabil tezaur al limbii noastre, care concentrează activitatea de cercetare a celor mai reputate colective românești de lingviști care l-au creat în decursul a mai mult de un secol și care se continuă și astăzi pentru actualizarea lui în concordanță cu evoluția continuă a limbii, ar înceta să mai aibă aureola unui diamant închis în rafturile bibliotecilor academice și ar deveni un produs de larg consum, consultat nu numai în țară ci oriunde în lume, așadar deschis tuturor vorbitorilor și cercetătorilor de limbă românească de pe glob.

– Studii ancorate în ediția electronică a marelui dicționar vizează: recuperarea morfologiei vechi (Simionescu et al., 2012), recuperarea sensurilor cuvintelor în context (Nastase și Szpakowicz, 2001), interconectarea resurselor lingvistice de diferite tipuri (Cristea et al., 2017b; Scutelnicu, 2021).

– Realizarea de colecții de arbori sintactici (treebank-uri) și antrenarea de parsere de analiză sintactică (Perez, 2012; Mărănduc, 2022).

– Realizarea de colecții de arbori de discurs și elaborarea unei teorii a structurii discursului – Teoria nervurilor, amintită deja, care facilitează analiza automată a structurii discursului, din care pot fi generate rezumate focalizate pe entități de discurs (Postolache, 2001; Cristea et al., 2005). O zonă de cercetări care au captat atenția noastră mai mult timp a constituit-o rezoluția anaforelor, domeniu în care s-a dezvoltat un model cadru de realizare a modelelor de rezoluție (Cristea și Dima, 2001) și s-a realizat un motor general de rezolvare a referințelor anaforice, care a trecut testul de a fi aplicat la mai multe limbi (Cristea et al, 2002; Cristea și Postolache, 2005).

– În zona semanticii, realizarea de colecții de fraze adnotate la roluri semantice (Trandabăț, 2009), adnotarea la relații semantice între entități textuale a romanului Quo Vadis de Henryk Sienkiewicz (Bibiri et al., 2014; Cristea et al., 2015).

– Preocupări de analiză temporală a textelor beletristice, care au dus la elaborarea Teoriei traseelor temporale (*Time Yards Model*) (Cristea și Macovei, 2017; Cristea, 2017);

– Colecții de adnotări asupra documentelor românești vechi (Haja și Tamba, 2021; 2022) și tehnologia DeLORo de transcriere a documentelor conținând tipărituri și scriere uncială din scriptul chirilic în cel latin (Cristea et al., 2022; Cristea et al, 2023 în curs de apariție). La Institutul de Matematică și Informatică de la Chișinău (IMMI) au fost și sunt în continuare efectuate cercetări de digitizarea și transliterarea textelor românești tipărite cu caractere chirilice în sec. XVII-XX. Dificultățile în tratarea acestei teme sunt date de numărul mare de perioade din evoluția limbii, numărul relativ mic și foarte dispersat de resurse accesibile și mai ales marea diversitate a alfabetelor chirilice. Au fost elaborate mai multe modele de recunoaștere optică a caracterelor, orientate la diferite perioade istorice și diferite clase de fonturi, grație cărora s-a obținut o precizie foarte bună la recunoaștere. A fost, de asemenea, elaborat un algoritm de clasificare a fonturilor în baza unei rețele neurale multistrat, care funcționează cu o acuratețe de peste 96%. Pentru diferite perioade istorice au fost formulate seturi de reguli (libere și dependente de context) de transliterare din alfabetul chirilic românesc în cel modern obținându-se o acuratețe de 98%. Platforma de digitizare integrează toate etapele de prelucrare a unui

document, pornind de la fișierul scanat și finalizând cu cel puțin două fișiere editabile (în fonturi originale și transliterate). Recent, preocupările acestui colectiv s-au îndreptat și spre procesarea documentelor eterogene care, alături de textul propriu-zis, conțin și elemente de altă natură: formule, imagini, diagrame, partituri muzicale etc. Documentul este supus unei segmentări, cu identificarea tipului fiecărui segment, care apoi este recunoscut de motorul său specific (FineReader, Mathpix etc.). Rezultatul recunoașterii este constituit din text în format editabil și elementele eterogene prezentate în scriptul potrivit tipului segmentului: LaTeX pentru formule matematice și desene tehnice, MusicXML pentru partituri muzicale, CVS combinat pentru diagrame, MOL pentru structuri chimice etc.

O analiză detaliată a resurselor și tehnologiilor dedicate limbii române la nivelul anului 2012, poate fi găsită în (Trandabăț et al., 2012). Preocuparea de inventariere a resurselor lingvistice românești și a tehnologiilor capabile să proceseze limba română ar trebui să fie activă la orice moment, pentru a sprijini continuu cercetările românești, pentru a alimenta cu informații noile teze de doctorat ce vor apărea în acest domeniu, resursele trebuind să configureze baze de date și de calcul care să alimenteze noile proiecte, dar acest lucru nu se întâmplă, din păcate, în momentul de față. Comisia de Informatizare pentru Limba Română (CILR), creată în 2002 pe lângă Academia Română, își propunea, la înființare, să realizeze acest inventar și să îl țină la zi, dar după lucrarea din 2012 referită mai sus el nu a mai fost actualizat. Așteptăm o revigorare a activității acestei comisii la întâlnirea ei planificată a avea loc în cadrul celei de a 18-a ediții a conferinței *Linguistic Resources and Tools for Natural Language Processing* – ConsILR-2023, ce se va ține la Brașov și online între 11-13 decembrie. Invit cititorii să viziteze paginile acestui eveniment[14] și adresez o chemare către cercetătorii filologi basarabeni, cărora li se adresează primordial publicația ce include lucrarea de față, să participe, activ sau pasiv, în acest eveniment.

Portaluri ale resurselor și instrumentelor create în diverse centre de cercetare din România și din Republica Moldova există și aș menționa: pe cel de la ARFI-IIT[15], cel de la AR-ICIA[16], portalul SOLIROM de la ARFI-IFRI[17], precum și portalul IMMI-Chișinău[18] (funcțional cu valoare istorică).

În privința educației universitare și a evenimentelor care sprijină domeniul resurselor și tehnologiilor limbii române, pregătind specialiști lingviști-

[14] https://profs.info.uaic.ro/~consilr/2023/
[15] http://lsplr.iit.academiaromana-is.ro/
[16] https://www.racai.ro/
[17] https://solirom.ro/
[18] http://www.math.md/elrr/res_main.php

informaticieni pentru cercetare și industrie, lăsând la o parte multe altele cel puțin la fel de importante, menționez aici doar câteva: seria școlilor EUROLAN, derulată bienal fără întrerupere începând din 1993, care a avut în acest an cea de a 16-a ediție[19], masteratul de Lingvistică Computațională, înființat în 2001 la UAIC-FII[20], precum și cel în Procesarea Limbajului Natural de la Universitatea București[21], seriile de conferințe dedicate domeniului prelucrării vorbirii *Speech Technology and Human Computer Dialogue* – SpeD[22], ajunsă în acest an la cea de a 12-a ediție, organizată de Universitatea Națională de Știință și Tehnologie Politehnica București, și seria de conferințe ConsILR, menționată deja, organizată tradițional de UAIC-FII, ARFI-IIT și AR-ICIA.

Concluzii. Aș încheia această trecere grăbită în revistă a domeniului în care îmi desfășor activitatea cu o chemare însuflețită la unirea forțelor. Pe de o parte, a tagmei lingviștilor cu cea a informaticienilor și, pe de altă parte, a comunităților care activează în aceste ramuri ale științei pe cele două laturi ale Prutului. Suntem două națiuni unite prin aceeași limbă minunată. Ea însă trebuie susținută și tehnologic ca să reziste asaltului limbilor uriașe din internet sau a celor cu care este la concurență în societate. Acest asalt își face din ce în ce mai mult prezenta simțită astăzi când limba noastră este cotropită de influențe străine, cu precădere din partea englezei dar și a rusei. Nu e vorba numai de limbajul tehnic, unde e mai comod să copiezi decât să calchiezi, dar și de limbajul de zi cu zi. În același timp, segmente importante de limbă veche, de o frumusețe încântătoare, sunt uitate.

Există o mulțime de activități pe care putem să le facem împreună, care ar începe toate cu strângerea legăturilor, în conferințe, în școli de vară, în schimburi științifice. Exemplul l-am dat recent, organizând ediția a 17-a a ConsILR, în noiembrie 2022, la Chișinău[23] și ultima ediție a școlilor EUROLAN, în septembrie a acestui an, la Bălți. A fost pentru prima oară când am scos din România aceste două evenimente. Dar am putea face programe de cercetare care să aibă componente comune, am putea organiza activități de creare de resurse în comun[24], am putea face schimburi de studenți sau îndruma studenți în cotutelă. Și, evident, tot ceea ce rezultă din aceste activități comune ar fi date în consum cetățenilor ambelor noastre țări, iar limba noastră ar crește (sau, cel puțin, nu s-ar ofili…). Abordările comune pe care le am în vedere

[19] https://eurolan.info.uaic.ro/2023/
[20] https://www.info.uaic.ro/programs/lingvistica-computationala-en/
[21] https://nlp.unibuc.ro/master
[22] https://sped.pub.ro/
[23] https://profs.info.uaic.ro/~consilr/2022/
[24] Mă gândesc acum câtă nevoie am fi avut de un ajutor în adnotarea paginilor scanate de carte veche în proiectul DeLORo, sau în procurarea de resurse din Basarabia pentru proiectul COROLA!

nu ar forța limitele de comprehensibilitate ale informaticianului, nedeprins cu limbajul lingvistului, și nici pe ale acestuia care n-ar trebui să fie descurajat de limbajul informaticianului. Fiecare însă poate contribui cu metodele și cunoștințele sale la realizarea acestui deziderat comun.

Referințe bibliografice[25]:

BIBIRI, Anca-Diana, COLHON Mihaela, DIAC Paul, CRISTEA, Dan. Statistics Over A Corpus Of Semantic Links: „QuoVadis". În: Mihaela Colhon, Adrian Iftene, Verginica Barbu Mititelu, Dan Cristea, Dan Tufiș (eds.) *Proceedings of the 10th International Conference* „*Linguistic Resources And Tools For Processing The Romanian Language*, *Craiova*, *18-19 September 2014*". Iași: Editura Universității „Alexandru Ioan Cuza", 2014, p. 33-44.

COMAN, Andrei, MITROFAN, Maria, TUFIȘ, Dan. Automatic Identification and Classification of Legal Terms in Romanian Law Texts. În: Onofrei M., Bibiri A., Nicolae C., Tufiș D., Cristea D. *Proceedings of the 14th edition of the International Conference on Linguistic Resources and Tools for Natural Language Processing* – ConsILR-2019. Iași: Editura Universității „Alexandru Ioan Cuza", 2019.

CORNILESCU, Alexandra. The distribution of APs, The Adjectival Phrase inside the DP [online]. Disponibil: http://www.linguist.jussieu.fr/~mardale/ADJc.doc, 2004 [citat 03.10.2023].

CRISTEA, Dan. Resurse lingvistice și tehnologii ale limbajului natural. Cazul limbii române. În: *Prelegeri Academice*, Academia Romana – Filiala Iasi, vol. III, nr. 3, Iasi, 2005.

CRISTEA, Dan. The Time Yards Model: Rethinking the way to look at texts. În: *SYNASC-2016*. 2017, p. 15-22.

CRISTEA, Dan, CLEJU, Nicolae, REBEJA, Petru, HAJA,Gabriela, COMAN, Eduard, VASILESCU, Anca, MARINESCU, Claudiu, DASCĂLU, Andreea. Bringing the Old Writings Closer to Us: Deep Learning and Symbolic Methods in Deciphering Old Cyrillic Romanian Documents. In: *Memoirs of the Scientific Sections of the Romanian Academy*. 2023, in press.

CRISTEA, Dan, DIMA, Gabriela Eugenia. An integrating framework for anaphora resolution. În: *Information Science and Technology*. Bucharest: Romanian Academy Publishing House, vol. 4, no. 3-4, 2001, p. 273-291.

CRISTEA, Dan, GÎFU, Daniela, COLHON, Mihaela, DIAC, Paul, BIBIRI, Anca-Diana, MĂRĂNDUC, Cătălina, SCUTELNICU, Liviu-Andrei. QuoVadis: A Corpus of Entities and Relations. In: N.Gala, R.Rapp and G.B. Enguix (eds.): *Language Production*, *Cognition*, *and the Lexicon*. Springer International Publishing Switzerland, 2015.

CRISTEA, Dan, GÎFU, Daniela, MORUZ, Alex, ONOFREI, Mihaela, PISTOL, Laura, SCUTELNICU, Andrei, BOLEA, Cecilia. An Insight into the Corpus of Contemporary Romanian. In: *Memoirs of the Scientific Sections*/ *Memoriile Secțiilor Științifice*. Series IV, Tome XL. Bucharest: Publishing House of the Romanian Academy, 2017, p. 67-84.

[25] Toate referințele în care sunt autor sau coautor pot fi consultate în pagina https://profs.info.uaic.ro/~dcristea/publications.html.

CRISTEA, Dan, IDE, Nancy, ROMARY, Laurent. Veins Theory. An Approach to Global Cohesion and Coherence. In: *Proceedings of Coling/ACL '98*, Montreal, 1998.

CRISTEA, Dan, MACOVEI, Andreea. Why Time Resembles Rail Yards? A Way to Look at Time in Text Books. In: *Romanian Journal of Information Science and Technology*. Vol. 19, 1-2, 2016, pp. 31-43, 2017.

CRISTEA, Dan, POSTOLACHE, Oana Diana. How to deal with wicked anaphora. În: António Branco, Tony McEnery and Ruslan Mitkov (editori): *Anaphora Processing: Linguistic, Cognitive and Computational Modelling*. Benjamin Publishing Books, 2005.

CRISTEA, Dan, POSTOLACHE, Oana Diana, PISTOL, Ionuț. Summarisation through Discourse Structure, în: *Proceedings of CiCling 2005*. Springer LNSC, vol. 3406, 2005.

CRISTEA, Dan, REBEJA, Petru, PĂDURARIU, Cristian. Applying YOLOv5 Learning in Detecting Old Cyrillic-Romanian Characters. In: *Proceedings of the the 17th International Conference* "*Linguistic Resources and Tools for Processing The Romanian Language*" (Chișinău, 10-12 Nov. 2022). 2022, pp. 115-122.

CRISTEA, Dan, WANDL-VOGT, Eveline, ONOFREI, Mihaela, SCUTELNICU, Andrei. Towards a technology for dictionary intermediated dynamic alignment of multilingual corpora. A vision and some concrete steps. In: *COST ENeL Meeting*, 24-25 February 2017, Budapest, 2017.

CRISTEA, Dan, WEBBER, Bonnie Expectations in Incremental Discourse Processing. In: *Proceedings of ACL*. Madrid, 1997.

ERJAVEC, Tomas, OGRODNICZUK, Maciej, OSENOVA, Petya N. et al. The ParlaMint corpora of parliamentary proceedings. In: *Language Resources and Evaluation 57*, 2022, p. 415-448.

FELLBAUM, Christiane. *WordNet: An Electronic Lexical Database*. MIT Press, 21, 1998.

FREGE, Gottlob. Über sinn und bedeutung (translated as `on sense and reference'). În: Geach and Black (eds.), *Translations from the Philosophical Writings of Gottlob Frege*. Blackwell, Oxford, translation of the paper from 1960, 1892.

HAJA, Gabriela, TAMBA, Elena Isabelle. Modele de inteligență artificială (deep learning) aplicate în analiza sintactico-semantică a limbii române vechi (DeLORo). Perspectiva filologilor. În: *Caietele Sextil Pușcariu. Actele Conferinței Internaționale* „*Zilele Sextil Pușcariu*", vol. V. Ediția a V-a, Cluj-Napoca, 9–10 septembrie 2021. Cluj-Napoca: Editurile Scriptor, Argonaut, 2021, p. 103-117.

HAJA, Gabriela, TAMBA, Elena Isabelle. The Role of the Philologist in the Project Artificial Intelligence Models (Deep Learning) Applied in the Analysis of Old Romanian Language (DeLORo). În: *Proceedings of the 16th International Conference „Linguistic resources and tools for natural language processing"* (13-14 december 2021), eds. Petru Rebeja, Mihaela Onofrei, Dan Cristea, Dan Tufiș. Iași: Editura Universității „Alexandru Ioan Cuza", 2022, p. 49-58.

KUPIETZ, Marc, WITT, Andreas, BAŃSKI, Piotr, TUFIȘ, Dan, CRISTEA, Dan, VÁRADI, Tamás. EuReCo – Joining Forces for a European Reference Corpus as a sustainable base for cross-linguistic research. In: Bański, Piotr; Kupietz, Marc; Lüngen, Harald; Rayson, Paul; Biber, Hanno; Breiteneder, Evelyn; Clematide, Simon; Mariani, John; Stevenson, Mark; Sick, Theresa (eds.): *Proceedings of the Workshop on Challenges in the Management of Large Corpora and Big Data and Natural Language Processing*

(*CMLC-5+BigNLP*) *2017,* including the papers from the Web-as-Corpus (WAC-XI) guest section. Birmingham, 24 July 2017. Mannheim: Institut für Deutsche Sprache, 2017, p. 15-19.

LOCKE, W.N., BOOTH, D.A. (eds.) *Machine Translation of Languages*. Cambridge, Massachusetts: MIT Press. 1955, pp. 15–23.

MANN, W.C., THOMPSON, S.A… *Rhetorical structure theory: A theory of text organization. Text, 8:3*, 243-281, 1988.

MARCUS, Solomon. *Lingvistica matematică*. București: Editura didactică și pedagogică, 1966. 253 p.

MĂRĂNDUC, Cătălina. *Instruments for Processing the Old Romanian Language* (*Training Corpora*, *Models*, *Statistics*), teză de doctorat, UAIC-FII, iunie 2022.

MONTAGUE, Richard. The proper treatment of quantification in english. In: K.J.J. et al. Hintikka, editor, *Approaches to Natural Language*,. D. Reidel, Dordrecht, 1973. pages 221-242

MORUZ, Mihai Alex, CRISTEA, Dan. A Bootstrapping System for Dictionary Management and Parsing. În: M. Mitrofan, D. Gîfu, D. Tufiș, D. Cristea (eds.): *Proceedings of the 12th International Conference on Linguistic Resources And Tools For Processing The Romanian Language – ConsILR*, Mălini, 27-29 october 2016. Iași: Editura Universității "Alexandru Ioan Cuza", 2016, p. 153-162.

NASTASE, Vivi, SZPAKOWICZ, Stan. Word Sense Disambiguation in Roget's Thesaurus Using WordNet. In: *Proceedings of NAACL-2001*. Pittsburgh, PA, 2001.

PEREZ, Cenel-Augusto. *Resurse lingvistice pentru prelucrarea limbajului natural*, teză de doctorat, UAIC-FII, ianuarie, 2012.

POLANYI, Livia. A Theory of Discourse Structure and Discourse Coherence. În: Eilfort, W. H., Kroeber, p. D. and Peterson, K. L. (eds.) *Papers from the General Session at the Twentl-First Regional Meeting of the Chicago Linguistics Societ*y, Chicago, April 25-27, 1985.

POSTOLACHE, Oana Diana. *Rezumare automată*, lucrare de licență, Facultatea de Informatică a Universității „Al.I. Cuza" din Iași, 2001.

REBEJA, Petru, CHITEZ, Mihaela, ROGOBETE, Roxana, DINCĂ, Andreea, BERCUCI, Loredana. ParlaMint-RO: Chamber of the Eternal Future. In: *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference*. 2022, pp. 131-134.

SCUTELNICU, Andrei. *Representing and Interconnecting Lexical Resources*, teză de doctorat, UAIC-FII, august 2021.

SIMIONESCU, Radu, CRISTEA, Dan, HAJA, Gabriela. Inferring diachronic morphology using the Romanian Thesaurus Dictionary. In: A. Moruz, et al. (eds.).*Proceedings of the 8th International Conference* "*Linguistic Resources And Tools For Processing Of The Romanian Language*", *8-9 December 2011, 26-27 April 2012* (Muzeul Național al Literaturii Române, București). Iași: Editura Universității „Al.I. Cuza", 2012, p. 85-92 [online] . Disponibil: https://profs.info.uaic.ro/~dcristea/papers/Inferring Diachronic Morpho-ConsILR 2011.pdf. [citat 03.10.2023].

SINCLAIR, John. *Looking Up*. London & Glasgow: Collins ELT, 1987.

TRANDABĂȚ, Diana. *Natural Language Processing Using Semantic Frames*, teză de doctorat, UAIC-FII, 2009.

TRANDABĂȚ, Diana, IRIMIA, Elena, BARBU MITITELU, Verginica, CRISTEA, Dan, TUFIȘ, Dan. The Romanian Language in the Digital Age / Limba română în era

digitală. In: *White Paper Series*, eds. Georg Rehm and Hans Uszkoreit. Berlin: Springer, 2012, 87 p.

TUFIȘ, Dan, CRISTEA, Dan. Methodological issues in building the Romanian Wordnet and consistency checks in Balkanet. În: *Proceedings of the Workshop on Wordnet Structures and Standardization, and how these affect Wordnet Applications and Evaluation, workshop in conjunction with The Third International Conference on Language Resources and Evaluation, LREC-2002*. Las Palmas, Spain, 2002.

TUFIȘ, Dan, CRISTEA, Dan, STAMOU, Sofia. BalkaNet: Aims, Methods, Results and Perspectives. A General Overview. În: *Romanian Journal on Science and Technology of Information*, Romanian Academy, Bucharest, Romania, vol. 7, no. 1-2, special issue on the Balkanet project, July 2004.

VOSSEN, Piek (Ed*.) EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Dordrecht: Kluwer Academic publishers, 1998.

Note: Articolul a fost realizat în cadrul proiectului de cercetare *Modele, resurse și tehnici de prelucrare a limbajului natural aplicate limbii române în diacronie și sincronie*, etapa II pe 2023, din planul de cercetare al ARFI-IIT.

Legăturile web au fost toate testate ultima oară la data predării articolului, adică în 15 octombrie 2023.

Aduc mulțumiri doamnei Svetlana Cojocaru, m.c. AȘM, vicepreședinte al Academiei de Științe a Moldovei, președinte al Secției de științe exacte și inginerești, pentru transmiterea informațiilor asupra realizărilor IMMI-Chișinău.