Universitatea „Dunărea de Jos” din Galați
Școala doctorală de Științe Fundamentale și Inginerești

TEZĂ DE DOCTORAT
- REZUMAT -

Contribuții privind reprezentarea cunoștințelor contextuale în procesarea limbajului natural

Doctorand
Cristian Niculiță

Conducător științific	Prof. univ. dr. ing. Luminița DUMITRIU
Referenți științifici	Conf. univ. dr. Corina FORĂSCU CSII dr. Verginica BARBU MITITELU Conf. univ. dr. ing. Emilia PECHEANU

Seria I2: Calculatoare și tehnologia informației Nr 6
GALAȚI
2020

Seriile tezelor de doctorat susținute public în UDJG începând cu 1 octombrie 2013 sunt:

Domeniul fundamental ȘTIINȚE INGINEREȘTI

Seria I 1: **Biotehnologii**
Seria I 2: **Calculatoare și tehnologia informației**
Seria I 3: **Inginerie electrică**
Seria I 4: **Inginerie industrială**
Seria I 5: **Ingineria materialelor**
Seria I 6: **Inginerie mecanică**
Seria I 7: **Ingineria produselor alimentare**
Seria I 8: **Ingineria sistemelor**
Seria I 9: **Inginerie și management în agricultură și dezvoltare rurală**

Domeniul fundamental ȘTIINȚE SOCIALE

Seria E 1: **Economie**
Seria E 2: **Management**
Seria SSEF: **Știința sportului și educației fizice**

Domeniul fundamental ȘTIINȚE UMANISTE ȘI ARTE

Seria U 1: **Filologie- Engleză**
Seria U 2: **Filologie- Română**
Seria U 3: **Istorie**
Seria U 4: **Filologie - Franceză**

Domeniul fundamental MATEMATICĂ ȘI ȘTIINȚE ALE NATURII

Seria C: **Chimie**

Domeniul fundamental ȘTIINȚE BIOLOGICE ȘI BIOMEDICALE

Seria M: **Medicină**

Mulțumesc, profund recunoscător, doamnei Profesor dr. ing. Luminița Dumitriu, care mi-a canalizat ideile în direcția potrivită și a cărei experiență a fost esențială pentru finalizarea acestei lucrări de doctorat.

Le mulțumesc membrilor comisiei de evaluare și susținere a tezei, ale căror observații valoroase m-au ajutat să finisez părțile în care era nevoie de mai multă claritate.

De asemenea, mulțumesc tuturor colegilor care m-au susținut întotdeauna plini de înțelegere.

În final, mulțumesc, din tot sufletul, familiei mele pentru neîncetata încredere, răbdare și încurajare, pe care mi le-a acordat în tot acest timp.

Cuprins

Introducere

Studiul prezentat în această lucrare de doctorat izvorăște din fascinația pe care generații întregi de cercetători din domeniul științei calculatoarelor au manifestat-o asupra înțelegerii limbajului natural și de care nici noi nu am avut practic nici o șansă să scăpăm, lăsându-ne astfel luați și duși de valul entuziasmului de a studia acest vast domeniu.

Din perspectiva noastră, obiectivul principal al studierii limbajului natural este acumularea unei capacități de înțelegere suficiente a modului uman, intuitiv de analiză, astfel încât să facă posibilă transpunerea acestei cunoașteri într-un mod practic care să permită interacțiunea naturală cu dispozitivele create de om. Totuși, în ciuda succeselor mai mult sau mai puțin importante de a insufla viață mașinii, scopul final, obținerea unei inteligențe artificiale reale, care să poată rivaliza cu modul uman de analiză, rămâne încă inaccesibil chiar și în prezent.

Un mod metaforic de a privi toate aceste studii, aste acela că fiecare cercetător, analizând un anumit aspect punctual, a apucat și tras totodată de un fir virtual al unei țesături informaționale complexe și deocamdată impenetrabile, în încercarea de a o destrăma, astfel încât, în ultimă instanță, să obțină recompensa ascunsă în spatele ei, care este însăși esența modului de înțelegere a limbajului natural.

Pentru a ne aduce contribuția la efortul general în domeniul procesării limbajului natural, în această lucrare am ales să ”tragem” puțin de firul dedicat analizei definițiilor. Scopul principal a fost de a putea identifica definițiile în text pentru a extrage din ele informațiile esențiale legate de conceptele pe care le descriu. De asemenea, un aspect notabil referitor la frazele definiționale pe care l-am analizat în cadrul lucrării, este acela că de obicei sunt destul de importante pentru a oferi prin ele însele o imagine generală clară asupra contextului în care se găsesc și prin aceasta considerăm că au și o capacitate de sumarizare relativ ridicată.

Bineînțeles, detecția și extragerea de definiții a fost studiată anterior și numeroși cercetători au propus diferite metode practice. Dacă cele mai simple se plasează în zona identificării definițiilor cu ajutorul tiparelor lexico-sintactice definite manual, cele mai complexe folosesc tehnici avansate de machine learning ce implică utilizarea rețelelor neuronale și modalități sofisticate de codificare a structurii și caracteristicilor lexico-sintactice specifice definițiilor.

Din punctul de vedere al unui cercetător care urmărește să studieze detecția definițiilor în contextul limbii române, toate aceste metode prezintă însă un mare dezavantaj, și anume că au fost create pentru alte limbi, majoritatea adresându-se limbii engleze. Din câte cunoaștem, pentru limba română nu a fost creată și implementată o metodă specific dedicată acestui scop, motiv pentru care ne-am propus să dezvoltăm una care să țină cont de particularitățile limbajului.

Deși inițial am plecat de la premisa preluării unei metode pentru limba engleză în scopul adaptării ei ușoare la nevoile noastre, am constatat pe parcurs că diferențele mari de complexitate necesită dezvoltarea unor mecanisme specifice de prelucrare mult mai avansate, aspect care constituie contribuția cea mai importantă a acestei lucrări.

Privire de ansamblu asupra tezei

Capitolul 1 prezintă o imagine de ansamblu a stadiului actual în privința metodelor de detecție/extracție a definițiilor. În secțiunea 1.2 sunt discutate o serie de noțiuni generale care ajută la cristalizarea noțiunii de definiție.

Secțiunea 1.3 face o prezentare a metodelor de preprocesare și instrumentelor de preprocesare iar 1.4 discută modalitățile de abordare a analizei definițiilor atât din punct de vedere teoretic, dar în special practic.

În capitolul 2 este prezentată pe scurt metoda în limba engleză de la care s-a pornit, care este bazată pe tipare lexico-sintactice, precum și procesul de adaptare a corpusului de antrenare folosit de aceasta, pentru limba română.

Pentru aceasta s-au efectuat și o serie de modificări de sintaxă (secțiunea 2.3), care s-au dovedit necesare în contextul complexității crescute a limbii române. Secțiunea 2.4 introduce conceptul **cuvintelor de legătură** care este un aspect definitoriu al metodei în limba română, iar secțiunea 2.5 definește o serie de măsuri ce descriu particularitățile acestora.

Capitolul 3 prezintă un studiu făcut asupra instrumentelor de etichetare a părților de vorbire și de marcare a grupurilor substantivale (noun phrases) disponibile pentru limba română.

De asemenea, în secțiunea 3.2 sunt prezentate o serie de adaptări și îmbunătățiri aduse instrumentului de etichetare ales.

În capitolul 4 este introdus conceptul de **tipar de simplificare**, o componentă ce joacă un rol important în obținerea formei canonice a frazei analizate, pe baza căreia se creează un **tipar simplu** care va fi folosit în detecția definițiilor.

Pentru reprezentarea acestor tipare de simplificare a fost definit un limbaj special relativ simplu a cărui sintaxă este centrată pe componenta de tipar. Secțiunile 4.2 și 4.3 prezintă aspecte generale ale sintaxei fișierelor de configurare a tiparelor de simplificare, iar secțiunea 4.5 descrie în detaliu toate elementele sintactice asociate componentelor de tipar.

Capitolul 5 prezintă procesul de creare a tiparelor simple.

Secțiunea 5.2 descrie modul intern de reprezentare a tiparului frazei analizate în timpul procesului de simplificare.

În secțiunea 5.4 sunt prezentate o serie de elemente de preprocesare ale frazei inițiale, realizate în scopul unei simplificări temporare pentru a ajuta instrumentul de etichetare să asocieze anumitor termeni dificili părțile de vorbire corecte.

Începând cu secțiunea 5.5 sunt prezentate etapele de obținere a tiparului simplu.

Secțiunea 5.14 descrie la modul general etapa de simplificare finală, cea a grupurilor substantivale, care prezintă două faze: etapa de analiză a enumerațiilor și etapa de simplificare propriu-zisă.

În prima fază, de analiză (secțiunea 5.15) se urmărește detecția enumerațiilor care pot apare în cadrul grupurilor substantivale și care pot fi: enumerații de substantive, de adjective sau de adjective cu formă de participiu. Pentru detecția enumerațiilor este generată o întreagă gamă de posibile combinații relaționale între termeni care sunt evaluate pe baza unor *scoruri de fiabilitate*: scorul de **similitudine**, de **structură**, de **discrepanță** și de **distanță**.

În secțiunea 5.16 este prezentat pe larg procesul de simplificare folosind informațiile culese în faza de analiză a enumerațiilor.

În finalul capitolului, în secțiunea 5.17, este descrisă procedura de stocare a tiparelor simple obținute, realizată prin împărțirea în sub-tipare corespunzătoare unităților formale ale definiției.

În capitolul 6 se realizează validarea experimentală a clasificatorului și se studiază ipoteza potrivit căreia **conceptele definiționale** importante din cadrul documentului au de asemenea o frecvență relativă ridicată. Menționăm că aceste concepte, amintite mai sus, sunt cele care în definiție corespund *termenului definit* și *super-conceptului* său.

În secțiunea 6.2 sunt prezentate rezultatele clasificării calculându-se masurile obișnuite de evaluare a clasificatoarelor: precizie, rata de reamintire, acuratețe, scorul F, iar în secțiunea 6.3 se face o analiză a cauzelor erorilor de clasificare.

Secțiunea 6.4 discută despre corelația dintre frecvență (TF-IDF) și importanța conceptelor definiționale, descriind două variante algoritmice de a selecta aceste concepte, validarea rezultatelor făcându-se în urma unei analize umane.

Ultimul capitol prezintă concluziile finale, contribuțiile acestei teze, precum și direcțiile posibile de cercetare viitoare.

Figuri

Tabele

Listinguri

1 Stadiul actual în domeniul detecției și extragerii definițiilor

1.1 Introducere

Achiziția de cunoștințe a fost încă de la începuturi un obiectiv major al procesării limbajului natural. Dorința este de a ajuta computerele să poată citi textul și să exprime cunoștințele pe care acesta le conține printr-o reprezentare formală, potrivită pentru a răspunde la întrebări și a rezolva probleme. Cu toate acestea, progresul a fost dificil. Abordările cele mai vechi au fost manuale, dar efortul masiv pentru codificare cunoștințelor le-au făcut foarte costisitoare și limitate la domenii bine definite. Ulterior, includerea și dezvoltarea abordărilor bazate pe învățare automată a redus substanțial acest efort, și a făcut pași importanți pe calea găsirii unei soluții îndeajuns de automată care să ducă procesul de la un capăt la altul. Chiar și așa, este important să menționăm că învățarea supravegheată necesită date etichetate, care este, ea însăși, o activitate destul de costisitoare și se poate dovedi chiar imposibilă pentru achiziționarea de cunoștințe pe scară largă în domenii care nu sunt strict delimitate [1].

Natura textului analizat în scopul detecției construcțiilor definiționale are o influență majoră în succesul acestei acțiuni. În textele bine structurate, cum ar fi textele tehnice sau medicale, identificarea automată a definițiilor este posibilă chiar prin utilizarea structurii și, eventual, a cuvintelor cheie [2]. De exemplu, în majoritatea manualelor de matematică, definițiile sunt marcate în mod evident în text și, de obicei, au un format special. În schimb, în textele mai puțin structurate, identificarea construcțiilor definiționale este în general mult mai dificilă, deoarece ele sunt exprimate de obicei într-o formă mai liberă din punct de vedere lingvistic. În astfel de cazuri, chiar și o ființă umană întâmpină dificultăți în identificarea definițiilor, iar procesul este laborios, necesitând parcurgerea cu mare atenție a întregului text. Din păcate, metodele de detecție dezvoltate până în prezent nu reușesc să automatizeze decât într-un mod destul de ineficient acest proces pentru textele mai puțin structurate.

În mod formal, extracția definițiilor are ca scop detectarea unor perechi termen – definiție dintr-un text obișnuit. În funcție de nivelul de detaliu dorit putem avea ca scop doar detecția definițiilor, ceea ce presupune o clasificare generală a frazelor în definiții sau non-definiții sau se poate urmări identificarea precisă a părților componente ale unei definiții, caz în care este necesară o operație mai complexă de etichetare a acestora [3].

De asemenea, extragerea definițiilor este strâns legată de identificarea relațiilor concept - supra-concept. În multe cazuri, metodele de extragere a definițiilor sunt capabile totodată să identifice perechi concept – supra-concept (hiponim – hiperonim), deoarece se axează pe detectarea definițiilor care se conformează într-o mare măsură tiparului Aristotelian, cunoscut din literatura clasică. Fiind bine definită, structura acestui tipar facilitează identificarea definițiilor prin metode predominant lexico-sintactice.

După cum am amintit anterior, identificarea manuală a definițiilor într-un text mare, chiar și unul având o natură structurată, cere un efort îndelungat și, în mod ideal, ar trebui automatizată. Dat fiind faptul că nu există o metodă de detecție a definițiilor creată pentru limba română, în cadrul acestei lucrări vom urmări să dezvoltăm o asemenea metodă bazată pe tipare lexico-sintactice, care este o adaptare a celei descrise de Navigli și Velardi în [4]. Trebuie să precizăm că odată cu metoda am preluat și corpusul de antrenare folosit de aceasta care conține definiții în formatul clasic aristotelian ce va fi descris în secțiunea următoare.

1.2 Noțiuni generale legate de definiții și glosare

Scopul principal al unei definiții este acela de a stabili în mod explicit sensul în care un termen este folosit într-un document tehnic sau științific [5], oferind informații care ar putea fi utile în diverse situații.

Un prim pas pentru detectarea și extragerea definițiilor din corpusuri, trebuie să fie chiar definirea lor [6], pentru a ști clar a ce anume trebuie să căutăm, sub ce formă se găsesc în text informațiile care ne interesează. Tipologiile de definiții care au fost menționate în mod frecvent în literatura de specialitate [7], [8] pot fi grupate în trei categorii: definiția formală, semi-formală și cea non-formală.

Definiția formală corespunde tipului definițional aristotelian descris în [9] prin ecuația X = Y + C. În aceasta, X reprezintă *definiendum* (conceptul definit, hiponimul). Ansamblul Y + C este *definiens* și constituie explicația oferită pentru termenul definit, Y fiind clasa generică de care aparține X, *genus* (supra-conceptul, hiperonimul) iar C reprezentând caracteristicile specifice, *diferentia*, care descriu modul în care X este diferit de celelalte elemente aparținând lui Y. Semnul de egalitate stabilește relația de echivalență dintre aceste elemente, fiind expresia din text a verbului definițional numit *definitor*, care joacă un rol central în majoritatea metodelor de identificare/extragere a definițiilor. Deoarece vom folosi foarte frecvent termenul *hiperonimie*, vom oferi în continuare o definiție a sa:

> ***Hiperonimia*** este o relație semantică de supraordonare între înțelesuri ale cuvintelor.

În cazul nostru, vom folosi termenul de hiperonimie pentru a ne referi la relația dintre conceptul definit și supra-conceptul său.

O definiție semi-formală descrie *definiendum* numai prin caracteristicile specifice sau atributele sale [10]. Definițiile formale și semi-formale pot fi simple (exprimate printr-o singură propoziție), sau complexe (două sau mai multe propoziții).

O definiție non-formală încearcă să aibă o exprimare obișnuită, astfel încât cititorul să poată vedea elementul familiar din noul termen [7]. Aceasta poate fi asocierea cu un sinonim, o parafrazare sau o derivare gramaticală.

Punctul comun între toate aceste viziuni cu privire la aceeași structură lingvistică, referită prin apelativul comun de "definiție", este că toate au același scop didactic, și anume dezambiguizarea sensului unui element lexical. Aceste descrieri prezintă definiția ca asocierea dintre un termen și hiperonimul său (*genus*), sau între un anumit termen și caracteristicile sale.

Pe lângă acestea, în lucrările ce tratează tipologiile de definiții, sunt menționate și alte moduri de a exprima definițiile.

O tipologie unificată este propusă în [11], din care prezentăm, ca fiind de interes, următoarele trei categorii noi:

- definiții exprimate prin marcatori lingvistici de "nivel scăzut": includ semne de punctuație ca parantezele, ghilimelele, cratimă, două puncte;
- definiții exprimate prin marcatori lexicali: elemente lexicale lingvistice sau metalingvistice;
- definiții exprimate de marcatori lingvistici de "nivel înalt": modele sintactice, cum ar fi anafora sau apoziția.

Atunci când sunt complete, definițiile au, în general, o structură regulată, conținând o serie de tipare lexicale și metalingvistice ce pot fi recunoscute în mod automat [12]. Contextul unei definiții este un fragment specializat dintr-un text care este, în principiu, structurat pe termenul respectiv și explicația acestuia, iar ambele elemente sunt conectate prin tipare tipografice sau lexico-sintactice [13]. În principal, formele tipografice sunt date de semnele de punctuație

(virgule, paranteza), în timp ce formele lexico-sintactice includ verbe definiționale cum ar fi: "*a defini*" sau "*a semnifica*", precum și marcaje discursive, spre exemplu "*adică*", "*cu alte cuvinte*". În plus, contextele definiționale pot include și modele pragmatice, care menționează condițiile de utilizare ele termenului sau clarifică sensul său, ca "*în termeni generali*", sau "*în acest sens*".

În urma analizării formelor lingvistice pe care le pot avea construcțiile definiționale, se pot identifica aceste tipare specifice, care pot fi apoi puse în valoare în cadrul unor instrumente automate de căutare. Această abordare a fost pusă în practică cu diverse grade de succes. De remarcat faptul că rezultatele obținute pe textele tehnice sunt, în general, mai bune decât în cazul celor non-tehnice, unde nivelul este satisfăcător.

Doi factori limitează succesul acestor rezultate. Pe de o parte, importanța relativă a diferitelor forme lingvistice este dificil de evaluat, fiind astfel de obicei ignorată. Pe de altă parte, găsirea unor forme lingvistice eficiente care să fie în măsură să recunoască majoritatea definițiilor reale, dar care să nu accepte non-definiții, necesită timp și expertiză și s-a dovedit a fi extrem de dificilă [2].

Procesul de creare a glosarelor este în mod implicit dependent de cunoștințele de specialitate ale domeniului dat, de conceptele și de limbajul acestuia. În cazul domeniilor științifice care evoluează în mod constant, glosarele trebuie să fie în mod regulat actualizate și extinse. Crearea manuală a glosarelor de specialitate este, prin urmare, costisitoare. Ca alternativă, au fost propuse metode de creare automate și semi-automate.

În [14] sunt identificate două roluri ale glosarelor specializate. În primul rând, ele reprezintă resurse lingvistice care rezumă termenii de bază ai unui domeniu specializat și în al doilea rând, ele sunt resurse de cunoștințe, prin faptul că furnizează definiții ale conceptelor pe care le reprezintă termenii respectivi. Glosarele au o utilitate evidentă ca surse de referință. Un studiu privind utilizarea ajutoarelor lexicografice în traducerea specializată a arătat că glosarele se numără printre primele cinci resurse utilizate [15]. De asemenea, s-a arătat că glosarele facilitează înțelegerea textelor și dobândirea de cunoștințe în timpul studiului [16]. Gruparea definițiilor într-un glosar, permite găsirea rapidă a informațiilor despre cuvintele cheie, precum și contextul în care pot fi ele găsite, ușurând asimilarea cunoștințelor [2].

Din punct de vedere al procesării de către calculator, glosarele în format electronic s-au dovedit resurse valoroase în anumite activități de procesare a limbajului natural precum găsirea răspunsului la întrebări [17], dezambiguizarea sensului [18], învățarea de ontologii [19].

1.3 Categorii de texte și instrumente de preprocesare

Antrenarea și evaluarea metodelor de extragere a definițiilor și relațiilor de hiperonimie se realizează folosind diverse colecții de texte. Pentru obținerea rezultatelor optime, în faza inițială de creare a tiparelor sau de antrenare a algoritmilor de clasificare/clusterizare, sunt utilizate corpusuri al căror conținut este ales cu grijă și, în majoritatea cazurilor, adnotat manual. Datorită dificultății de obținere a lor, aceste corpusuri tind să aibă dimensiuni mici. În faza de testare sunt utilizate corpusuri de dimensiuni cât mai mari pentru a ameliora problema acoperirii reduse a datelor.

Structura textului induce două categorii: corpusuri *specializate/tehnice* care în general au un grad de structurare crescut [6], [13], [14] și corpusuri *eterogene* create din texte culese de pe net din diverse surse [20], [21], [22], [23], [24], [25]. O sursă de texte des utilizată care se situează undeva la mijloc este Wikipedia, în care informațiile au într-o anumită măsură un format previzibil.

La un capăt al spectrului se află metodele de extragere a definițiilor care în general au nevoie de colecții de text bine structurate pentru a obține rezultate bune [6], [13]. Succesul sistemului DEFINDER [26] se datorează într-o măsură considerabilă utilizării corpusului bine structurat

din domeniul medical, înregistrând valori de 0.87 pentru precizie și 0.75 pentru rata de reamintire. De partea cealaltă se află abordările de tip clusterizare pentru crearea de taxonomii care pot utiliza texte eterogene [20], [21].

Privite din punctul de vedere al **adaptabilității metodelor** în cadrul cărora sunt folosite, corpusurile pot fi *specifice* unei anumite limbi [6], [13], [2], *extensibile* pentru alte limbi [21], [22], [24], [27], sau sunt de la bun început *multilingvistice* [23]. De regulă metodele care se bazează în mare parte pe tiparele clasice sunt mai dificil de adaptat deoarece structurarea lexico-sintactică diferă de la o limbă la alta.

În general textele sunt analizate cel puțin cu ajutorul unui instrument de analiză a frazei care etichetează părțile de vorbire și eventual creează un arbore de dependențe. Aceste informații sunt utilizate atât pentru tiparele lexico-sintactice cât și pentru obținerea caracteristicilor în clusterizare sau clasificare. Câteva dintre instrumentele mai des utilizate sunt: Minipar [20], [21], Stanford POS Tagger [2], [25], Stanford Parser [24], Tree Tagger [4], TnT Tagger [14], etc.

1.4 Rezumat al modurilor de abordare a extragerii de definiții

Studiul extragerii automate a cunoștințelor de tip definiție a fost abordat atât din perspectiva teoretico-descriptivă cât și din cea practică.

Una dintre primele **lucrări teoretico-descriptive** este [12], în care este descris mediul contextual în care apar termenii. Aici se menționează că, atunci când autorii definesc un termen, folosesc pe de o parte modele tipografice pentru a evidenția vizual prezența termenilor și/sau definițiilor, precum și tipare lexicale și metalingvistice specifice, folosind structurile sintactice pentru conectarea elementelor contextului definițional. Această noțiune a fost întărită de [10], care susține că modelele definiționale pot oferi în plus, anumite chei ce permit identificarea tipului de definiție, un lucru foarte util în elaborarea de ontologii.

Lucrările practice pun în aplicare toate aceste elemente cu scopul de a dezvolta procedee de identificare și extragere automată a contextelor definiționale.

Din punct de vedere al modului de antrenare, putem grupa metodele în două categorii. Prima o constituie **abordările supravegheate**, în care putem plasa multe din sistemele bazate pe tipare și reguli create manual. Ca abordări mai speciale putem aminti metoda de reprezentare a tiparelor folosind un model de generalizare a laticilor de cuvinte prezentată în [4] unde se folosește un set de antrenare compus din definiții ale căror unități formale sunt adnotate manual. De asemenea, în [39] dependențele sintactice din frază sunt folosite pentru crearea unor caracteristici care să descrie adnotațiile semantice ale hiponimelor și hiperonimelor, în scopul antrenării unui clasificator SVM. Pentru a doua categorie, în care intră **abordările semi-supravegheate**, se pot menționa o serie de metode ce utilizează tehnica de bootstrapping precum [14], [23] sau [40] care, cu anumite mici particularități, aplică formula generală în care pornind de la un set inițial redus de termeni și tipare de definiții, colectează iterativ perechi de termeni/definiții, din care se extrag noi tipare pentru reluarea ciclului de căutare.

În funcție de scopul pentru care au fost elaborate, distingem de asemenea două categorii. Abordările de detecție a definițiilor dezvoltate în contextul mecanismelor de *tipul întrebare-răspuns* [41], [42], [43] sunt **centrate pe *definiendum***, deoarece caută definiții despre un anume termen. Al doilea tip de abordare este **centrat pe *definitor***, adică pe verbele care apar în mod obișnuit în definiții, scopul fiind de a găsi lista completă a tuturor definițiilor dintr-un corpus, indiferent de termenii definiți [5], [13], [44], [40], [3].

Marea majoritate a abordărilor legate de extragerea de definiții se bazează pe seturi de reguli sau tipare create manual pentru a identifica definițiile din text. Pe lângă acestea, un procent redus de autori au căutat să încorporeze și tehnici de machine learning, în efortul de a

îmbunătăți rezultatele obținute prin aplicarea anterioară a tiparelor/regulilor. Recent, chiar a apărut un curent de cercetare care renunță complet la utilizarea tiparelor, bazându-se în totalitate pe utilizarea rețelelor neuronale. In continuare vom face o trecere în revistă a metodelor mai importante, punctând pentru fiecare caracteristicile lor specifice.

Pentru categoria metodelor în care **tiparele** au fost **elaborate și ajustate manual** menționăm [26], în care este descris un sistem de căutare a definițiilor în domeniul medical, numit DEFINDER, alcătuit din două module. Primul modul de analiză lexicală aplica tehnici de potrivire a tiparelor utilizând expresii verbale tipice cum ar fi "*se numește*" și "*este definit ca*" precum și un set limitat de marcatori de text ”()” sau ”--”. Al doilea modul, de analiză gramaticală, poate analiza dependențele sintactice dintre cuvinte pentru a identifica definițiile exprimate prin construcții lingvistice mai complexe. În [6] scopul principal este identificarea construcțiilor definiționale exprimate cu ajutorul relațiilor de hiperonimie (”*un X este un Y care...*”) și sinonimie (”*echivalent cu*”), pentru care a fost creat un set de tipare lexico-sintactice folosind ca informații lema, partea de vorbire și funcția sintactică a cuvintelor. În [45] s-a dezvoltat un motor de căutare web care utilizează modele mai generale și mai complexe de tipul "*orașe, cum ar fi SubstantivPropriu (Început (ConstrucțieSubstantivală))*" care fac posibilă constrângerea rezultatelor prin proprietăți sintactice. În [5] se urmărește detecția și adnotarea definițiilor dintr-o colecție de texte tehnice în limba germană. Ei folosesc tipare centrate pe verbele definiționale (asociate frecvent cu definițiile). Tiparele specifică totodată și posibilitățile de poziționare a componentelor *definiendum* și *definiens*, astfel încât, după detecția construcțiilor definiționale acestea pot fi adnotate cu ușurință.

În general, o metodă bazată doar pe reguli/tipare are o precizie ridicată, dar o rată de reamintire scăzută, deoarece nu este capabilă să detecteze variabilitatea crescută a structurilor sintactice pe care le pot avea definițiile. Pentru a ameliora această problemă a reamintirii, [42] propune ideea încorporării în tiparele lexicale a unor termeni secundari care apar frecvent definițiile etalon, extrase din dicționare (WordNet, Enciclopaedia Britannica) precum și în alte resurse enciclopedice de pe internet. Este de remarcat faptul că această abordare este specifică pentru metodele centrate pe termeni și nu este scalabilă pentru terminologii și domenii arbitrare.

Pentru a evita sarcina de a crea manual setul de date antrenare, **tiparele** definiționale pot fi **extrase în mod automat** prin diverse metode.

În această direcție, un mecanism destul de des întâlnit pentru obținerea tiparelor este **tehnica de bootstrapping**. În [14] este prezentată o abordare ce implică estimarea automată (folosind metoda de bootstrapping) a tiparelor dintr-o colecție de texte, prin generalizarea structurii propozițiilor, folosind părțile de vorbire ale cuvintelor în combinație cu utilizarea unor măști de abstractizare a secvențelor neesențiale (”*.+*”, ”*.**”). La fiecare iterație tiparele noi sunt filtrate pe baza unui scor calculat în funcție de procentul termenilor de inițializare împreună cu care apar. La final, tiparele obținute sunt supuse unui proces de rafinare manuală. În [23] este elaborat un sistem de creare a glosarelor, numit Glossbot, ce utilizează web-ul ca sursă de informații, conceput astfel încât să nu depindă de limbă. Pentru a elimina necesitatea utilizării unui corpus de domeniu, se apelează la tehnica de bootstrapping, amorsând procesul de extracție al tiparelor cu doar câteva perechi sămânță termen-hiperonim. Achiziționarea iterativă a tiparelor de extracție a definițiilor din paginile web de tip glosar, prin exploatarea formatării HTML, poate acoperi variabilitatea mare a definițiilor textuale, care includ atât propoziții ce respectă tiparele lexico-sintactice obișnuite (ex. "*un corpus este o colecție de documente*"), precum și definițiile stil glosar (ex. "*corpus: o colecție de documente*"), toate acestea, indiferent de domeniul țintă. La fiecare sfârșit de iterație, definițiile noi sunt ordonate în funcție de relevanță prin intersecția mulțimii de cuvinte a definiției cu terminologia domeniului, fiind păstrate doar cele care se clasează pe primele poziții. De asemenea, tehnica de bootstrapping este folosită în [40] pentru a ușura la elaborarea unor tipare noi, însă într-un mod hibrid, fiind dublată la finalul fiecărei iterații de o verificare manuală a tiparelor în scopul rafinării lor și, la final, pentru eliminarea celor având o precizie redusă. Tiparele utilizate sunt de două tipuri: lexicale - care extrag din

text fraze candidat și de analiză sintactică - utilizate pentru a decide dacă într-adevăr fraza conține o definiție identificând totodată termenul definit, explicația sa, precum și informațiile complementare acestora.

O altă posibilitate de automatizare a achiziției tiparelor este folosirea unui **corpus adnotat** în prealabil într-un mod specific. Acest lucru este realizat în [4] care dezvoltă o metodă numită latici de clase de cuvinte ce permite atât extragerea definițiilor cât și a perechilor hiponim – hiperonim din cadrul acestora. Prin utilizarea unor așa numite tipare stea se ameliorează problema variabilității frazelor definiționale, oferind totodată o modalitate flexibilă de extragere automată a hiperonimelor din ele. Pentru a face acest lucru sunt creați o serie de clasificatori bazați pe latici, folosind un set de definiții textuale. Propozițiile de antrenare sunt grupate automat pe baza similarității și, pentru fiecare asemenea grup, se creează un clasificator sub formă de latice care modelează toate variantele pentru tiparul de bază al grupului. O latice este un graf aciclic orientat, o subclasă de automate cu stări finite non-deterministe. Scopul structurii laticii este de a codifica (într-o formă compactă) diferențele importante din cadrul secvențelor distincte. Studii ulterioare ale acelorași autori au demonstrat că această abordare dă dovadă de precizie ridicată în mai multe domenii [46].

După cum am menționat anterior, o serie de lucrări folosesc **metode de machine learning** pentru a îmbunătăți rezultatele obținute în faza de potrivire a tiparelor/regulilor. În cazul în care aceste rezultate înregistrează o precizie scăzută datorită permisivității prea mari a tiparelor utilizate, este posibilă utilizarea modulului de machine learning pentru filtrarea lor în scopul eliminării cazurilor fals pozitive. În [47] se folosește un clasificator de entropie maximă pe un corpus compus din pagini medicale extrase din Wikipedia în limba daneză, din care au extras propoziții pe baza unor caracteristici sintactice. Tiparele lexico-sintactice sunt combinate în [48] cu un clasificator Bayes naiv în scopul extragerii de glosare din textele tutorial în limba daneză.

În încercarea de a rezolva problema variabilității limbajului, [17] au propus o metodă bazată pe tipare lexico-sintactice probabilistice, folosind modele n-gram, care pot modela dependențele locale secvențiale între cuvinte. Comparativ cu tiparele obișnuite, ele sunt capabile de generalizare și permit potrivirea parțială, prin calcularea probabilității unui așa numit "grad de potrivire generativă" dintre o instanță de test și un set de instanțe de antrenare. În acest caz, tiparele nu sunt folosite efectiv în procesul de detecție al definițiilor ci doar pentru a ajuta modelul să recunoască secvențele de cuvinte specifice acestora.

O abordare foarte diferită și unică în felul ei, a fost propusă de [2] care utilizează un algoritm genetic pentru ponderarea unui set de caracteristici definiționale identificate de către experți umani, în scopul obținerii unui filtru eficient de selecție a definițiilor.

Rămânând în sfera tehnicilor de machine learning trebuie să menționăm un curent recent de cercetare care folosește doar rețelele neuronale pentru modelarea relațiilor dintre termen (*definiendum*) și explicație (*definiens*) folosind reprezentări ale cuvintelor de tip *word embedding*. Această abordare se dovedește capabilă de a îmbunătăți scorul de reamintire (engl. recall), care este destul de scăzută la metodele bazate pe tipare. Caracteristicile de nivel înalt din definițiile de antrenare sunt codificate prin intermediul filtrelor convoluționale (CNN) și sunt apoi furnizate ca date de intrare pentru rețeaua neuronală. În [49] este prezentată o metodă de procesare a informațiilor lexico-sintactice dintre termen și explicație prin intermediul rețelelor neuronale, pentru învățarea unor caracteristici de nivel înalt capabile să identifice structuri definiționale similare în text. [3] duce mai departe această idee, creând o metodă complexă de analiză a similarității atât la nivel local între termen și explicație cât și global între ansamblul termen-explicație și fraza din care acestea provin.

1.5 Motivația alegerii tehnologiei folosite

În scopul realizării clasificatorului pentru limba română pe care ni l-am propus, am ales ca

sursă de inspirație metoda cu latici de clase de cuvinte prezentată în [4], care se bazează pe utilizarea tiparelor. Am dorit o metodă bazată pe tipare pentru a putea avea acces la mecanismele interne ale clasificatorului. Tiparele, spre deosebire de un model bazat pe machine learning, pot fi înțelese dacă sunt păstrate într-un format corespunzător. Un alt avantaj pe care l-am considerat extrem de atractiv a fost posibilitatea modificării manuale a acestor tipare după faza de antrenare. Se pot aplica astfel în mod direct cunoștințele lingvistice umane pentru îmbunătățirea tiparelor obținute. De asemenea, faptul că metoda se bazează pe crearea automată a tiparelor pe baza unui corpus de definiții, ușurând astfel enorm procesul lor de achiziție, a fost un alt aspect pozitiv. Pentru a valorifica acest lucru am preluat și corpusul de definiții în engleză pe care l-am tradus în limba română făcând toate ajustările de adnotare necesare acestui scop.

1.6 Concluzii

Definiția constituie un mecanism de descriere prin care oamenii pot înțelege orice termen necunoscut. Aspectele esențiale ce descriu un anumit domeniu pot fi sintetizate într-un mod foarte succint printr-o serie de definiții ale termenilor principali ai domeniului, cel mai adesea grupate într-un glosar al domeniului.

Acest capitol face o trecere în revistă a modurilor în care a fost abordată extragerea definițiilor de-a lungul timpului, începând cu modele destul de simple create manual, bazate pe recunoașterea tiparelor lexico-sintactice, evoluând treptat spre modele care urmăresc să automatizeze procesul de creare a tiparelor în diferite moduri:

- crearea automată studiind structura frazelor dintr-un corpus de antrenare
- crearea automată folosind tehnica de bootstrapping prin identificarea în text a unor perechi concept – supra-concept deja cunoscute

Abordările bazate pe tipare, care până la această dată se dovedesc a avea precizia cea mai mare, suferă totuși de dezavantajul de a avea o capacitate de reamintire destul de scăzută.

Pentru a ameliora acest fapt, a început recent să se dezvolte un curent de cercetare care pune în valoare puterea de modelare a rețelelor neuronale, special concepute pentru a procesa reprezentările termenilor în formatul *word embeddings*. Acest format are capacitatea de a codifica într-o formă numerică vectorială similaritățile semantice dintre cuvinte.

Fiind o subproblemă a prelucrării limbajului natural, extragerea definițiilor este o sarcină cu grad de dificultate foarte ridicat, motiv pentru care anumite tipuri de definiții rămân până la această dată aproape imposibil de detectat. Printre acestea menționăm cazul definițiilor care se extind dincolo de limitele unei fraze, în care termenul definit și definiția sa se află în fraze diferite, fiind conectate semantic prin expresii referențiale. Totuși în ultimii ani, linia de cercetare bazată pe rețele neuronale pare că va avea mai mult succes în această direcție, corelată fiind cu elaborarea unui corpus definițional [38] care conține și adnotează într-un mod foarte precis aceste cazuri dificile.

Toate abordările prezentate în acest capitol sunt fie exclusiv axate pe o anumită limbă fie au un caracter general din acest punct de vedere. Dat fiind acest aspect, în această lucrare, ne propunem ca obiectiv principal dezvoltarea unui sistem de extragere a definițiilor, special optimizat pentru limba română. Acest sistem este bazat pe tipare lexico-sintactice create printr-un procedeu automatizat, pe baza unei colecții de definiții de tip aristotelian, manual adnotată în acest scop.

2 Contribuții privind preprocesarea textului

2.1 Introducere

În cadrul acestei lucrări se urmărește adaptarea metodei de detecție a definițiilor bazată pe latici de clase de cuvinte (word class lattices - WCL), prezentată în [4]. Aceasta va fi descrisă pe scurt în secțiunea următoare.

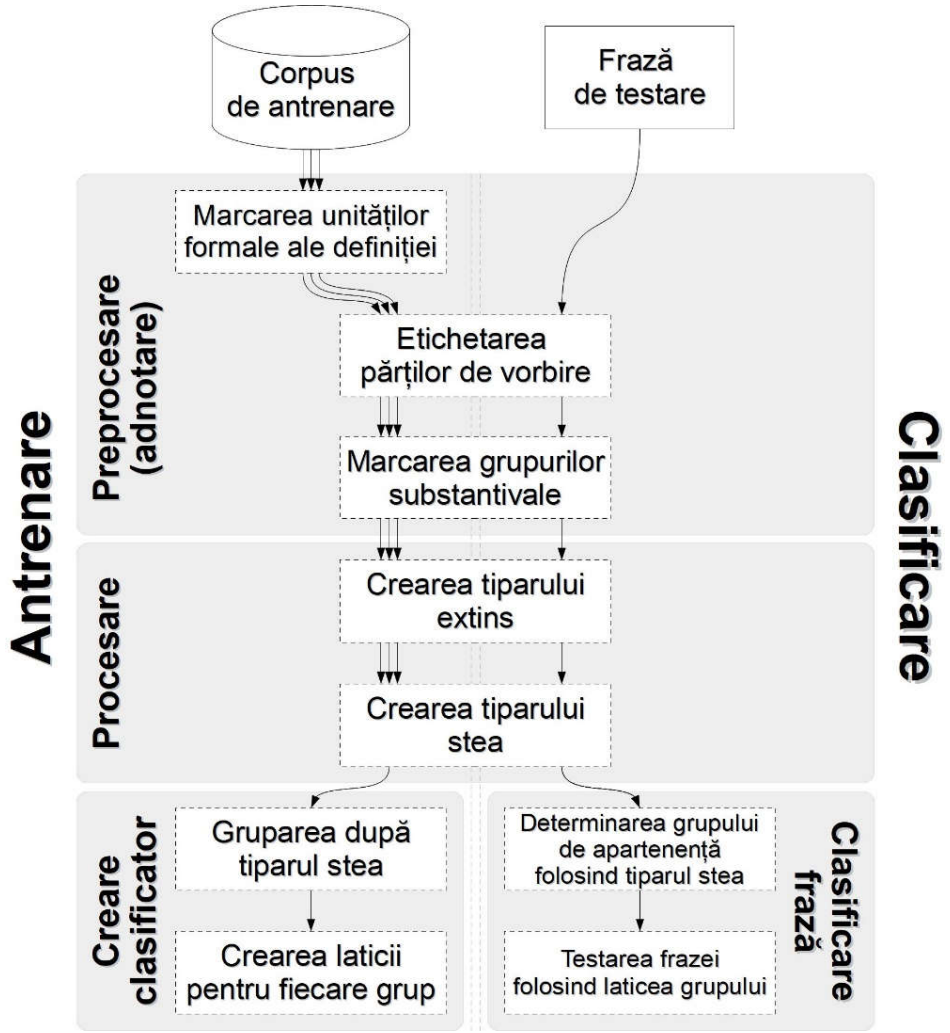


Figura 2-1 Diagrama clasificatorului WCL

În esență WCL (figura 2-1) este o metodă de detecție bazată pe tipare. Dificultatea principală în adaptarea acesteia la limba română constă în faptul că în limba engleză frecvența apariției prepozițiilor în cadrul grupurilor substantivale este mult redusă ceea ce permite crearea unor așa numite *tipare stea*. Modul de creare a tiparelor stea începe cu identificarea clasei fiecărui cuvânt din fraza analizată. O clasă a unui cuvânt este o generalizare a lui, care se face urmând două reguli. Dacă acel cuvânt este foarte frecvent în corpus (ex. ”pe”, ”la”, ”de”), el este considerat cuvânt cheie și rămâne nemodificat. În caz contrar cuvântul este înlocuit de partea sa de vorbire (ex. majoritatea substantivelor, adjectivele, etc.). Tiparul stea este obținut prin înlocuirea tuturor secvențelor de clase ale cuvintelor comune prin caracterul ”*” (figura 2-2).

Fraza:	In	geography	,	a	country	is a	political	division .
Tipar extins:	In	NN	,	a	NN	is a	JJ	NN .
Tipar stea:	In	NN	,	a	NN	is a	*	.

Figura 2-2 Tipar stea

2.2 Eliminarea tiparelor stea

Încercarea de a crea tipare stea pentru limba română are un rezultat cel mult mediocru, abilitatea de a grupa definițiile fiind drastic limitată de utilizarea frecventă a prepozițiilor. Eficiența tiparelor stea este dată de capacitatea lor de generalizare care pentru limba engleză este suficientă. Prin contrast, în limba română, nu este un fapt neobișnuit ca un grup substantival să conțină două, trei sau chiar mai multe prepoziții fapt care determină fragmentarea ei atunci când se urmărește crearea unui tipar stea.

2.3 Reformatarea setului de definiții pentru limba română

Pentru antrenarea clasificatorului în limba română s-a decis utilizarea setului de definiții de antrenare utilizat pentru clasificatorul WCL în engleză. Ideea inițială care a determinat luarea acestei decizii a fost dorința de a putea realiza o paralelă cât mai bună între cele două clasificatoare. O altă motivație a venit din faptul că având părțile de definiție deja etichetate, în urma traducerii puteau fi direct utilizate fără alte eforturi suplimentare de etichetare manuală.

Această operațiune a întâmpinat însă două probleme serioase:

- la traducerea grupurilor substantivale apăreau inversiuni de cuvinte, se intercalau prepoziții, rezultatul fiind că formatările pentru conceptul definit și supra-conceptul acestuia deveneau în multe cazuri nepotrivite necesitând ajustări chiar importante
- pentru adnotarea inițială nu se urmărise un set de reguli de adnotare bine stabilit deoarece existau situații în care două definiții similare ca structură erau adnotate diferit, fapt care ar fi dus la ambiguități în faza de antrenare

Având în vedere aceste aspecte am hotărât că este necesară o reetichetare a definițiilor care să țină cont de particularitățile limbii române, urmând un set de reguli de etichetare bine stabilit care să acopere majoritatea cazurilor dificile.

Definițiile din corpusul de antrenare au fost adnotate manual cu o serie de etichete care delimitează părțile definiției după modelul clasic propus de Aristotel.

Etichetele folosite în adnotarea unităților formale ale definițiilor sunt următoarele:

- **<RGET>** - conține cuvântul definit inclusiv anumite informații suplimentare
- **<TARGET>** - marchează exact termenul definit
- **<VERB>** - precizează construcția verbală centrală a definiției. Aceasta poate fi doar o formă simplă a verbului ”*a fi*”, sau poate avea o structură complexă care conține mai multe (de obicei două) verbe specifice definițiilor
- **<GENUS>** - cuprinde porțiunea de text care este explicația cuvântului definit
- **<HYPER>** - precizează cuvântul sau expresia care desemnează categoria din care face parte entitatea sau aspectul definit
- **<REST>** - desemnează o parte a textului definiției care conține informații adiționale care pot lipsi fără să aibă un impact asupra modului de înțelegere a explicației definiționale

Diferențe față de formatul folosit pentru limba engleză includ:

1. introducerea perechii <TARGET> ... </TARGET> – în engleză acest grup nu există, fiind folosit un sistem de înlocuire al cuvântului definit (țintă) prin cuvântul cheie TARGET. Modificarea a fost necesară pentru a suporta marcarea conceptelor definite formate din mai multe cuvinte.

2. modificarea etichetei <GENUS> prin adăugarea unei etichete terminale </GENUS> pentru a putea prelucra mai bine cazurile în care sunt furnizate mai multe explicații pentru cuvântul definit.

3. gruparea <HYPER> ... </ HYPER> nu mai apare de mai multe ori în cadrul aceleiași unități de tip <GENUS>. Enumerațiile simple sunt grupate într-o singură unitate <HYPER>, iar cazurile complexe (în care grupurile substantivale conțin prepoziții cu grad redus de coeziune) sunt tratate prin folosirea grupărilor <GENUS> multiple menționate anterior.

4. tiparele verbale au fost pe cât posibil simplificate prin introducerea cuvintelor speciale în gruparea <GENUS>.

2.4 Considerații privind adnotarea definițiilor în limba română

Ideea principală pe care se bazează re-adnotarea definițiilor este folosirea unor așa-zise cuvinte de legătură (prepoziții, anumite adverbe și pronume) care introduc grupuri substantivale (engl. noun phrases – NP).

În cadrul definiției, aceste grupuri substantivale marchează, prin substantivul principal, cuvântul definit și hiperonimul corespunzător. În plus, pentru delimitarea unităților formale ale definiției, se pot folosi ca elemente ajutătoare aceste cuvintele de legătură, ce au în cele mai multe cazuri, un tipar de utilizare destul de bine definit.

De obicei în text avem de-a face cu fraze definiționale care conțin o definiție precum și anumite explicații suplimentare, exemple etc. de care ne putem lipsi. Scopul pe care ni-l propunem este găsirea numărului minim de cuvinte care formează partea principală a unei definiții, prin utilizarea cuvintelor de legătură.

Pentru a explica procedura de delimitare a unităților formale ale definiției (RGET, TARGET, VERB, GENUS, HYPER, REST), vom considera cuvintele de legătură ca și elemente de decizie care pot hotărî rămânerea într-o anumită unitate formală sau tranziția la următoarea unitate validă, dictată de aplicarea regulilor specifice ale categoriei din care face parte cuvântul de legătură (ex. TARGET -> RGET, VERB -> GENUS, HYPER -> GENUS, etc.)

Caracteristicile cuvintelor de legătură pot fi analizate ținând cont de două aspecte:

- gradul de coeziune
- completitudinea definiției

Relația asociată unui cuvânt de legătură are un *grad de coeziune* dictat de coerența semantică a grupului substantival respectiv și se poate clasa în două categorii:

a. relații indivizibile – grupul nu poate fi împărțit în subunități fără a-și pierde înțelesul original (ex. apă ***de*** mare)
b. relații divizibile – pot fi împărțite (ex. formațiuni de rocă ***în*** pământ)

Din punct de vedere al *completitudinii definiției*, scopul este de a include volumul minim de informații care face ca definiția să fie suficient de informativă și inteligibilă. În acest sens se disting trei tipuri de informații:

a. obligatorii
b. informative

c. dispensabile

Considerând cele două aspecte descrise mai sus, (1) **gradul de coeziune** și (2) **importanța informației introduse**, putem grupa cuvintele de legătură în patru categorii generale.

Obs. Trebuie precizat faptul că, în acest context, importanța informației este într-o anumită măsură evaluată în mod subiectiv pe baza contribuției aduse la înțelesul global a definiției, fiind reflectată în adnotarea manuală a setului de definiții.

Cele patru categorii însoțite de caracteristicile (regulile) lor implicite sunt prezentate în cele ce urmează:

STRONG – Informația introdusă de cuvântul de legătură (ex. prepoziția ”de”) este obligatorie și nu poate fi separată de substantivele de care aparține, formând relații indivizibile. Din acest motiv, deși pot fi găsite oriunde în definiție, aceste cuvinte de legătură nu determină, în mod normal, trecerea la o altă unitate formală.

MEDIUM – Și în acest caz informația are o contribuție importantă la gradul de înțelegere al definiției, fapt ce o face să aparțină de grupa obligatorie. Cuvintele de legătură din acest grup (ex. pronumele genitival ”al”) sunt asociate cu relațiile divizibile. Regula implicită în cazul lor este că determină tranziția de la TARGET la RGET și de la HYPER la GENUS, dar niciodată de la GENUS la REST. În mod normal ele nu ar trebui să se găsească în TARGET sau în HYPER.

WEAK – Informația introdusă se clasează în categoria informativă, deoarece, în pofida faptului că ajută la nuanțarea caracteristicilor conceptului definit, poate fi omisă fără a avea un impact major asupra înțelegerii de ansamblu a definiției. De asemenea această categorie de cuvinte de legătură (ex. prepozițiile ”din”, ”pentru”) corespunde relațiilor divizibile. La fel ca și categoria MEDIU, ele determină trecerea de la TARGET la RGET și de la HYPER la GENUS. Diferența este că, dacă ne aflăm în GENUS, orice al doilea cuvânt de legătură întâlnit va face tranziția în REST. În urma aplicării regulilor implicite rezultă că aceste cuvinte de legătură nu pot apare decât în RGET, GENUS și REST.

STOP – Informația introdusă nu are un impact notabil asupra înțelegerii definiției, făcând parte din categoria dispensabilă. Cuvintele de legătură (ex. adverbul ”cum”, conjuncția ”iar”) corespund relațiilor divizibile și întâlnirea lor determină trecerea imediată la unitatea formală REST, indiferent dacă ne aflăm în HYPER sau în GENUS.

Cuvintele de legătură se conformează mai mult sau mai puțin regulilor de tranziție implicite menționate pentru fiecare categorie generală. Pentru tratarea excepțiilor, destul de numeroase, au fost elaborate reguli specifice fiecărui cuvânt de legătură.

2.5 Contribuții privind descrierea tiparelor de utilizare a cuvintelor de legătură

Folosind următoarele trei statistici ale cuvintelor de legătură: numărul de apariții (*Nr_total*), numărul de tranziții (*Nr_tranz*) și numărul de excepții (*Nr*_exc) au fost definite trei măsuri care descriu tiparele de utilizare:

1. ***Fitness*** - gradul de potrivire al cuvintelor de legătură la categoria lor

Gradul de potrivire măsoară gradul în care cuvântul de legătură se conformează regulilor stabilite pentru categoria de care aparține. El se măsoară în procente, formula de calcul fiind următoarea:

$$Fitness = \left(1 - \frac{Nr\_exc}{Nr\_total}\right) \cdot 100$$

Cu ajutorul acestei măsuri se poate determina cât de potrivită este încadrarea unui cuvânt de legătură într-o anumită categorie și eventual se poate decide asupra reîncadrării sale într-o categorie mai potrivită. Dintr-o perspectivă probabilistică această măsură poate fi interpretată ca fiind puterea de predicție a respectivului cuvânt de legătură referitor la regulile implicite de adnotare ce ar trebui aplicate.

2. ***Mobility*** - mobilitatea cuvintelor de legătură

Mobilitatea este o măsură a predispoziției unui cuvânt de legătură de a determina trecerea de la o unitate formală la alta în cadrul definiției. Ea se calculează făcând raportul dintre numărul de tranziții și numărul total de apariții după cum urmează:

$$Mobility = \frac{Nr\_trans}{Nr\_total}$$

Din formulă se poate observa că valorile sale variază în intervalul [0, 1] în care 0 indică imobilitatea, iar 1 mobilitatea totală.

3. ***PredError*** - eroarea de predicție a tranziției pentru cuvintele de legătură

Eroarea de predicție este proporția în care cuvintele de legătură apar în situații care încalcă regulile implicite ale categoriei din care fac parte, fapt care determină definițiile respective să dobândească un caracter mai specific sau mai general (în funcție de tipul excepției în cauză).

Regulile implicite, care sunt destul de generale și au rolul de directive globale, nu pot să acopere toate particularitățile fiecărui cuvânt de legătură. Din acest motiv aplicarea exclusivă a regulilor implicite, care sunt în mod intenționat simple, tinde să aibă ca rezultat extragerea unor definiții cu caracter vag. Pe de altă parte, adnotațiile făcute în setul de antrenare iau în considerare individualitățile cuvintelor de legătură, creând astfel diferențele măsurate de eroarea de predicție, scoțând în evidență cuvintele de legătură care prezintă caracteristici speciale.

Din punct de vedere informațional, deviațiile cuvintelor de legătură pot avea două tendințe legate de includerea informațiilor (descrise de regulile implicite):

- mai multe în comparație cu acțiunea implicită - mărește astfel specificitatea definiției
- mai puține - face ca definiția să aibă un caracter mai general

În practică sunt calculate două valori pentru eroarea de predicție, care corespund celor două tendințe descrise mai sus:

$$PredError_{+} = \frac{Nr\_exc_{+}}{Nr\_total}$$

$$PredError_{-} = \frac{Nr\_exc_{-}}{Nr\_total}$$

Valorile pentru Nr_exc_{+} și Nr_exc_{-} sunt calculate folosind următoarele două formule:

$$Nr\_exc_{+} = Nr\_exc_{TARGET\_RGET} + Nr\_exc_{HYPER\_GENUS} + Nr\_exc_{GENUS\_REST}$$

$$Nr\_exc_{-} = Nr\_exc_{RGET\_TARGET} + Nr\_exc_{GENUS\_HYPER} + Nr\_exc_{REST\_GENUS}$$

Calculând erorile de predicție pentru o serie de cuvinte de legătură frecvente se observă tendința aproape exclusivă către includerea de informații care este perfect naturală deoarece o definiție trebuie să fie cât mai precisă (specifică), pentru a furniza în mod eficient explicația intenționată.

Deoarece valorile $PredError_+$ și $PredError_-$ sunt normalizate, putem defini gradul de încredere (TrustDegree) ca fiind măsura inversă erorii totale de predicție:

$$TrustDegree = 1 - (PredError_+ + PredError_-)$$

2.6 Concluzii

Lista de contribuții din capitolul 2:

- adaptarea modului de adnotare a unităților formale ale definițiilor pentru limba română
- identificarea cuvintelor de legătură și gruparea lor în patru categorii: STRONG, MEDIUM, WEAK, STOP
- elaborarea unor măsuri de descriere a tiparelor de utilizare a cuvintelor de legătură
- elaborarea unui set cuprinzător de reguli de adnotare a unităților formale ale definițiilor ținând cont de caracteristicile particulare importante ale cuvintelor de legătură

În acest capitol sunt descrise adaptările făcute metodei de extragere a definițiilor, prezentată de Navigli și Velardi în [4] și care se bazează pe utilizarea tiparelor grupate folosind latici de clase de cuvinte. Ei numesc tiparele extrase drept *tipare stea* dat fiind faptul că înlocuiesc anumite secvențe de părți de vorbire (pentru creșterea gradului de generalitate) prin intermediul caracterului ”*”.

Pentru limba română utilizarea metodei bazate pe tiparele stea este ineficientă datorită frecvenței crescute a prepozițiilor. Pentru obținerea unui tipar alternativ eficient, este introdus conceptul de *cuvânt de legătură*. Sunt identificate patru categorii de cuvinte de legătură în funcție de *gradul coeziunii* asociate (relații indivizibile și, respectiv, divizibile) și în funcție de *importanța informațiilor* introduse de acestea (obligatorii, informative, dispensabile). Aceste categorii au fost numite STRONG, MEDIUM, WEAK și STOP, fiecare având asociată o combinație specifică de valori pentru cele două trăsături menționate mai sus.

Adaptarea metodei prezentate de Navigli și Velardi presupune câteva modificări ce țin în esență de modul de marcare al conceptelor definite și supra-conceptelor.

Aceste diferențe, precum și studierea atentă a corpusului în limba engleză (care a fost tradus pentru a fi utilizat pentru metoda adaptată la limba română) au născut necesitatea elaborării unui set de reguli de adnotare a unităților formale ale definiției, care să asigure extragerea unor tipare având un format consecvent.

De asemenea, pentru a caracteriza cât mai detaliat cuvintele de legătură au fost elaborate o serie de măsuri și anume: gradul de potrivire (*fitness*), mobilitatea (*mobility*), eroarea de predicție (*predError*) calculate pe baza unor date numerice statistice extrase din corpusul de antrenare, cum ar fi: numărul de apariții, numărul de tranziții, numărul de excepții. Toate aceste valori sunt calculate pentru fiecare cuvânt de legătură prezent în corpus.

3 Contribuții privind instrumentele de etichetare morfo-sintactică în limba română

Una dintre primele etape ale prelucrării limbajului natural este etichetarea părților de vorbire (PoS tagging). Majoritatea abordărilor sunt în general axate pe prelucrarea secvențială cuvintelor. Această acțiune presupune asocierea unei părți de vorbire (clasa cuvântului, categoria lexicală) fiecărui cuvânt. Acest proces poate fi privit ca o formă simplificată de analiză morfologică.

Din punct de vedere lingvistic, specialiștii sunt în mare parte de acord că există trei părți de vorbire principale: substantiv, verb și adjectiv [52]. La crearea unui set de etichete, ideea de bază este de a eticheta cu părți de vorbire toate clasele de cuvinte care au comportament gramatical diferit.

Aproape toate sistemele de etichetare pentru părți de vorbire funcționează cu un set de etichete bine determinat. Dându-se o propoziție, sistemul trebuie să atribuie câte o etichetă pentru fiecare cuvânt al ei. Această acțiune întâmpină două mari dificultăți:

- cuvinte ambigui - cuvinte cărora li se pot atribui mai multe etichete
- cuvinte necunoscute - cuvinte care nu apar în corpusul de antrenare

O altă problemă în etichetarea părților de vorbire, este concordanța setului de etichete. Folosirea unui set mare permite codificarea mai multor aspecte legate de structura morfo-sintactică a cuvintelor, dar în același timp face dificilă distincția între etichetele similare. Această distincție poate fi uneori atât de subtilă încât chiar oamenii pot să nu cadă de acord în această privință după cum arată un experiment făcut în [53] pe Penn Treebank în care experții care au adnotat au avut păreri diferite în aproximativ 7.2% din cazuri.

Acuratețea maximă obținută în etichetare (numărul de cuvinte etichetate corect raportat la numărul total de cuvinte) este în prezent de aproximativ 96%-97% pentru majoritatea limbilor indo-europene. Este demn de notat faptul că este posibilă obținerea unei acurateți mari (în jur de 90%) folosind doar metode extrem de rudimentare, cum ar fi doar alegerea celor mai probabile etichete pentru cuvinte [54]. Metodele sofisticate urmăresc acoperirea ultimelor zece procente.

Acuratețea de 96%-97% poate fi totuși privită ca fiind foarte ridicată și orice încercare de îmbunătățire a performanței peste aceste valori se dovedește extrem de dificilă.

În [55] se realizează o clasificare a tipurilor de erori care împiedică instrumentele de etichetare să atingă acuratețea maximă (tabelul 3-1). Acestea includ și erorile umane prezente în seturile de antrenare.

Tabelul 3-1 Tipuri de erori la etichetarea părților de vorbire

Nr. crt.	Clasă	Frecvență
1	Formă inexistentă în lexicon	4.5%
2	Cuvânt necunoscut	4.5%
3	Posibil să fie etichetat bine	16%
4	Context lingvistic dificil	19.5%
5	Imprecizie/neclaritate	12%
6	Inconsecventă/lipsă standard	28%
7	Standard etalon greșit	15.5%

3.1 Analiza instrumentelor existente

Toate sistemele de procesare analizate se bazează într-o formă sau alta, pe setul de etichete de tip descriere morfo-sintactică (morpho-syntactic description - MSD) [56] dezvoltat în cadrul proiectului Multex-East (www.nl.ijs.si/ME).

Pentru operația de etichetare a părților de vorbire și partiționare în grupuri sintactice am evaluat trei instrumente de etichetare:

1. **UAIC POS tagger** [57] este un instrument de etichetare hibrid ce combină un model statistic cu un sistem bazat pe reguli care este elementul principal al acestei abordări, fiind utilizat pentru a reduce ambiguitatea cuvintelor înainte de atribuirea etichetelor. Aceasta facilitează foarte mult acțiunea de corectare a anumitor tipuri de erori, deoarece permite includerea directă a cunoștințelor lingvistice în sistem. Pe lângă instrumentul de etichetare a fost dezvoltat și un instrument de identificare a grupurilor substantivale [58].

2. **MLPLA - Modular Language Processing framework for Lightweight Applications** [59] este un instrument de procesare modular. El conține un modul de intrare, o serie de module de procesare (etichetator de părți de vorbire, lematizor, procesor de unități sintactice, etc.) a căror funcționare este independentă între ele, oferind posibilitatea creării unor linii de procesare flexibile în funcție de necesități. De asemenea, conține un modul de ieșire care formatează informațiile. MLPLA folosește un set redus de etichete derivate din MSD, numit CTAG [60] în care anumite informații cum ar fi genul sunt excluse din etichete deoarece pot fi inferate la nevoie pe baza formei cuvântului. Numărul de etichete este astfel redus la 78, dar acest fapt determină creșterea gradului de ambiguitate la etichetare.

Noile instrumente dezvoltate în cadrul RACAI, MLPLA v2 și RELATE, sunt antrenate folosind metodologia și adnotările specificate în cadrul proiectului dependențelor universale (Universal Dependencies – universal.dependencies.org). Din acest motiv multe dintre etichetele folosite de aceste două instrumente corespund cu cele din limba engleză. În momentul scrierii acestei lucrări ele nu sunt accesibile decât sub formă de serviciu web, ceea ce pentru necesitățile de procesare ale metodei dezvoltate în această lucrare constituie un impediment major.

3. **TreeTagger** [30] a fost conceput ca răspuns la problema ce punea dificultăți aproape tuturor instrumentelor de etichetare anterioare și anume, estimarea cu acuratețe a probabilităților cu valori mici, în condițiile în care setul de antrenare era de dimensiuni reduse. Tehnica propusă pentru evitarea problemei rarității datelor se bazează pe un arbore decizional pentru estimarea, cu grad de încredere ridicat, a probabilităților de tranziție.

Deși nu a fost dezvoltat special pentru limba română el poate fi antrenat ca orice alt instrument de acest tip pentru limba de interes și poate obține rezultate bune în cazul în care corpusul de antrenare este corespunzător ca dimensiune, relativ la setul de etichete țintă.

TreeTagger nu oferă facilități pentru detecția grupurilor sintactice.

4. **UDPipe** [60] este un instrument antrenabil pentru tokenizare, etichetare, lematizare și analiza dependenței textelor în format CoNLL-U. În crearea lui s-a urmărit în primul rând simplitatea de utilizare, scop pentru care s-au avut în vedere următoarele aspecte:

- furnizarea celor mai bune instrumente de tokenizare, analiză morfologică, etichetare a părților de vorbire și analiză a dependențelor
- crearea unui instrument unic care să folosească un singur model (pentru fiecare limbă)
- crearea pe cât posibil de modele pentru cât mai multe limbi
- evitarea utilizării de informații specifice unei anume limbi

Precizia UDPipe pentru limba română (versiunea 2.0) este de 81% pentru etichetarea părților de vorbire, 75% pentru lematizare.

Pentru verificarea instrumentelor de etichetare a textului au fost alese o serie de definiții din corpusul de antrenare WCL în limba engleză. În scopul exemplificării tipului de informații furnizate la ieșire de fiecare instrument și a modului de prezentarea a acestora, am ales următoarea definiție:

In Greek mythology, Callisto was a nymph of Artemis.

Definiția a fost mai întâi tradusă în limba română folosind Google Translator și apoi prelucrată cu fiecare dintre instrumentele de etichetare menționate anterior în scopul comparării rezultatelor și determinării fiabilității fiecăruia:

UAIC POS tagger

Instrumentul de etichetare hibrid dezvoltat la UAIC furnizează un set cuprinzător de informații, iar ca și capacitate de etichetare se descurcă foarte bine, el neînregistrând, de asemenea, nici o eroare pentru definiția prezentată. Pentru marcarea grupurilor substantivale se poate folosi direct rezultatul obținut la faza de etichetare a părților de vorbire.

MLPLA și MLPLA v2 / RELATE

Instrumentul se descurcă surprinzător de bine, ținând cont și de faptul că în elaborarea lui, principalul parametru restrictiv a fost micșorarea cât mai mult posibil a dimensiunii modelului și a costurilor computaționale. În cazul definiției analizate acest instrument înregistrează o eroare de etichetare.

Noua versiune a acestui instrument, MLPLA v2, precum și RELATE, antrenate folosind Universal Dependencies, au la etichetare rezultate comparabile cu cele obținute de UAIC POS tagger.

De asemenea, faptul că folosesc un set de etichete asemănătoare cu standardul folosit pentru limba engleză îi conferă un avantaj, dat fiind faptul că în cadrul prezentei abordări în care se dorește păstrarea unui grad cât mai ridicat de paralelism între metoda de clasificare originală și adaptarea ei pentru limba română. În plus el oferă o gamă mai largă de informații care este extrem de utilă în procesarea frazelor definiționale, inclusiv o analiză a dependențelor sintactice.

TreeTagger

Informațiile furnizate de TreeTagger sunt minime, ele incluzând doar eticheta MSD și lema cuvântului. De asemenea sunt înregistrate două erori de etichetare.

Un alt neajuns este faptul că fiind o aplicație de sine stătătoare, necesită lansarea unui proces extern pentru a funcționa, transferul de informații intrare/ieșire necesitând o interacțiune indirectă și greoaie prin intermediul unor fișiere pe disc.

UDPipe

UDPipe oferă, de asemenea, un set generos de informații morfo-sintactice. Ca și MLPLA v2 / RELATE a fost de asemenea antrenat folosind Universal Dependencies, furnizând pe lângă etichetele MSD și pe cele de tip UD. Ca și acestea el poate face și analiza dependențelor sintactice între termenii din frază.

În ordinea importanței vom enumera prioritățile metodei pe care dorim să o dezvoltăm:

1. posibilitatea integrării instrumentului de etichetare în cadrul unei aplicații de sine

stătătoare
2. performanța în etichetarea părților de vorbire

Referitor la primul punct, precizăm că MLPLA v2 / RELATE sunt disponibile doar ca servicii web ceea ce nu permite integrarea lor în cadrul unei aplicații. TreeTagger, fiind el însuși o aplicație de sine stătătoare, necesită lansarea unui proces extern pentru a funcționa, transferul de informații intrare/ieșire necesitând o interacțiune indirectă și greoaie prin intermediul unor fișiere pe disc. UDPipe deși scris în C++ pune la dispoziție un sistem de referințe pentru a permite accesarea din Java. În schimb, MLPLA și UAIC POS tagger vin sub formă de librării Java care pot fi integrate ușor în orice proiect, în special ultimul care poate fi utilizat fără nici o ajustare prealabilă.

Din punctul de vedere al performanțelor de etichetare, MLPLA v2, RELATE și UDPipe pot fi considerate pe același nivel, urmate îndeaproape de UAIC POS tagger. Ținând cont de faptul că MLPLA v2 și RELATE au fost eliminate din cursă, comparația finală se face între UDPipe și UAIC POS tagger.

În mod normal, câștigător ar fi fost UDPipe datorită performanțelor bune și a posibilității de integrare relativ simple. Motivul pentru care în final nu a fost utilizat în cadrul tezei constă din faptul că acest instrument a fost descoperit ulterior, moment la care se depusese deja un volum de muncă important pentru integrarea instrumentului ales anterior, **UAIC POS tagger**. În plus, deși la testare se descurcase bine, procentele de succes în cazul etichetării părților de vorbire și lematizării, așa cum erau prezentate pe site, apăreau ca fiind inferioare celor corespondente lui UAIC POS tagger. Din aceste motive s-a considerat ca timpul care ar fi trebuit alocat reintegrării unui nou instrument de etichetare să fie utilizat pentru activități mai apropiate de scopul demersului de cercetare.

3.2 Îmbunătățirea preciziei operației de etichetare pentru limba română

Ușurința de integrare a instrumentului de etichetare ales, UAIC POS tagger, în cadrul unei aplicații Java nu constituie nici pe departe cea mai importantă facilitate a sa. După cum a fost precizat anterior, procesarea informațiilor se realizează în două etape:

- o etapă statistică – realizată cu ajutorul unui clasificator antrenat prin metoda entropiei maxime
- o etapă bazată pe reguli care procesează rezultatele anterioare

Această a doua etapă este un element extrem de important, deoarece regulile pot fi create și ajustate manual pentru a obține creșterea preciziei de etichetare.

Regulile sunt de fapt o serie de tipare complexe, reprezentate sub formă de grafuri direcționate, fiecare componentă de tipar fiind poziționată într-un nod al grafului. Anumite noduri pot decide alegerea unor variante de părți de vorbire (KEEP) sau eliminarea altora (REMOVE) în funcție de necesități. Parcurgerea grafului se face de la stânga la dreapta, iar acolo unde există mai multe variante de continuare, acestea sunt încercate pe baza unui număr de ordine asociat conexiunii. Căutarea eșuează în cazul în care nu se găsește nici o cale care să parcurgă graful de la început și până la sfârșit.

Deși regulile implicite cu care vine instrumentul de etichetare erau în general adecvate, am observat că în analiza frazelor definiționale, apăreau totuși anumite erori frecvente atât de identificare a părților de vorbire cât și de marcare a grupurilor substantivale, fapt ce a determinat demararea unui proces de adăugare și modificare de reguli pentru ameliorarea acestor neajunsuri.

Obs. În cele de urmează referirea la formele *directe* semnifică cazurile de nominativ/acuzativ, iar cele *oblice*, cazurile de genitiv/dativ.

Reguli de corectare a etichetelor părților de vorbire

Toate ajustările aduse regulilor de identificare a părților de vorbire determină modificarea rezultatului analizei unui număr de 1084 de definiții din totalul de 1773.

Nu putem spune dacă într-adevăr toate acestea constituie modificări pozitive, deoarece, oricât de atent este analizată noua regulă, pot apare efecte colaterale negative. În marea majoritate a cazurilor analizate, rezultatele obținute prin implementarea noilor reguli sunt cu certitudine pozitive.

Vom enumera în continuare câteva modificări care au avut un impact demn de menționat:

- regulă pentru corectarea erorii prin care conjuncția ”*sau*” era confundată aproape întotdeauna cu forma fără diacritice a determinantului ”*său*”, lucru extrem de grav în condițiile în care conjuncția *sau* joacă un rol fundamental în identificarea enumerațiilor – 501 erori
- regulă pentru identificarea corectă a formelor de infinitiv a verbelor cărora li se atribuia etichetă de verbe la indicativ – 172 erori
- regulă pentru tratarea erorii prin care articolul genitival ”*al*” era etichetat ca și articol demonstrativ ”*cel*”, fiind confundat cu forma fără diacritice a regionalismului ”*ăl*”, aspect foarte grav deoarece articolul genitival este considerat cuvânt de legătură de tip MEDIUM – 115 erori
- regulă pentru ameliorarea erorii prin care cuvintele aflate pe prima poziție în frază, care nu există în dicționarul tagger-ului, sunt etichetate drept adverbe, în condițiile în care aceste cuvinte sunt în majoritatea cazurilor substantive proprii sau comune – ~90 erori
- regulă pentru corectarea încadrării cuvântului ”*multe*” în categoria incorectă oblică, fapt ce determina marcarea eronată a grupurilor substantivale – 33 erori
- regulă pentru corectarea erorii prin care adjectivul ”*mari*” era identificat greșit ca fiind forma fără diacritice a substantivului ”*mări*” – 24 erori
- regulă pentru corectarea încadrării substantivelor cu formă directă precedate de determinant oblic în categoria celor oblice – 11 erori

Prelucrări de corectare a marcajelor grupurilor substantivale

Modificările din această categorie au fost canalizate pe îmbunătățirea regulilor deja existente în scopul acceptării unor noi posibile combinații de termeni în cadrul grupurilor substantivale. În final toate aceste ajustări au avut ca rezultat corectarea marcajelor pentru aproximativ 560 de grupuri substantivale.

Deoarece nu putem vorbi de reguli distincte vom descrie etapele succesive care au dus la rezultatul final:

- includerea adverbelor aflate după adjective și posibilitatea existenței mai multor determinanți înaintea unui substantiv – 48 de corecturi
- includerea unor situații suplimentare legate de gradele de comparație aflate înaintea substantivelor și a adjectivelor cu formă directă asociate substantivelor cu formă oblică – 222 de corecturi
- includerea articolelor nehotărâte despărțite de substantiv prin adjective sau determinanți, a numeralelor ordinale aflate înaintea substantivelor – 164 de corecturi
- includerea determinanților aflați după substantive oblice, îmbunătățirea integrării substantivelor introduse de articolul genitival ”*al*” – 86 de corecturi
- includerea substantivelor care fac parte dintr-o enumerație dar le lipsește articolul genitival – 35 de corecturi
- îmbunătățirea integrării dintre substantivele comune și cele proprii – 40 de corecturi

3.3 Concluzii

Lista de contribuții din capitolul 3:

- îmbunătățirea regulilor de corectare a etichetelor corespunzătoare părților de vorbire
- îmbunătățirea marcării grupurilor substantivale (noun phrases) în text

Acest capitol discută despre caracteristicile instrumentelor de etichetare precum și de cauzele care determină apariția erorilor în procesul de etichetarea a părților de vorbire.

Având acest lucru în vedere, se realizează o analiză a caracteristicilor și performanțelor a cinci instrumente de etichetare: UAIC POS Tagger, MLPLA – Modular Language Processing framework for Lightweight Applications), MLPLA v2 / RELATE, UDPipe și TreeTagger, pentru a determina care dintre ele este cel mai potrivit, îndeplinind cel mai bine necesitățile aplicației de detecție a definițiilor pentru limba română, ce este dezvoltată în această lucrare.

În urma realizării testelor este ales UAIC POS Tagger datorită flexibilității sale și a ușurinței de a putea fi integrat într-o aplicație independentă. Acest instrument de etichetare are particularitatea de a avea o natură hibridă, funcționarea sa realizându-se în două etape: o etapă statistică care are un motor de regresie liniară antrenat pe principiul maximei entropii și o etapă bazată pe reguli care permite ajustarea rezultatelor anterioare. În felul acesta utilizatorul poate influența într-o mare măsură rezultatele obținute în prima etapă prin ajustarea acestor reguli.

Ca și contribuții în această direcție, au fost create o serie de reguli noi care determină atât îmbunătățirea preciziei în operația de atribuire a părților de vorbire, cât și creșterea eficienței operației de marcare a grupurilor substantivale.

4 Contribuții privind limbajul de reprezentare a tiparelor de simplificare

4.1 Introducere

Metoda prezentată în această lucrare are la bază un sistem de testare cu ajutorul unor *tipare definiționale*. Acestea sunt create folosind o suită de *tipare de simplificare* pe baza informațiilor extrase dintr-o colecție de texte formată în exclusivitate din definiții. Procesul de simplificare lexicală a frazelor este realizat în scopul aducerii lor la o *formă canonică*, astfel încât tiparele definiționale rezultate să fie cât mai generale și mai simple, păstrând însă caracteristicile esențiale ale unei definiții.

În paragraful anterior am menționat două tipuri de tipare:

1. **tipare de simplificare** – folosite la antrenare, pentru obținerea formei canonice a frazelor
2. **tipare definiționale simple (tipare simple)** – create la antrenare pe baza formei canonice a definițiilor și care vor fi folosite în faza de clasificare pentru detecția/extragerea definițiilor din text

Aceste tipare de simplificare pot fi privite ca niște reguli de simplificare ce vor fi aplicate secvențial pentru obținerea formei canonice a frazelor. Ele sunt specificate în fișiere de configurare pentru a facilita editarea lor. Pentru definirea regulilor a fost conceput în mod special un limbaj de reprezentare relativ simplu, având o sintaxă compactă, centrată pe componenta de tipar. În consecință, putem spune că cea mai importantă caracteristică a limbajului este *flexibilitatea*, extrem de utilă în faza de elaborare (experimentare) a tiparelor de simplificare.

Pentru a controla procesul de simplificare este prevăzut și un mecanism de adnotare al cuvintelor prin care se pot atașa într-un mod codificat informații pentru regulile ce urmează să fie testate în continuare. Spre deosebire de regulile de simplificare ce sunt asociate cu întreg tiparul, regulile de adnotare sunt asociate doar cu o anumită componentă a tiparului.

Clasificatorul descris în această lucrare utilizează trei astfel de fișiere de configurare a tiparelor de simplificare:

- pentru reprezentarea tiparelor asociate cuvintelor de legătură (LINKWORDS)
- pentru reprezentarea tiparelor pentru regulile de simplificare și adnotare (MIXED)
- pentru definirea componentelor de tipar complexe (a seturilor) (SETS)

Dat fiind faptul că îndeplinesc funcții specifice, fiecare dintre ele are anumite particularități de sintaxă lucru foarte evident în cazul fișierului SETS.

Fiecare fișier este alcătuit din două secțiuni principale:

1. prima dintre ele definește directive de preprocesarea a regulilor.
2. a doua conține tiparele propriu-zise (LINKWORDS, MIXED) sau definițiile de seturi (SETS).

În cadrul acestui capitol vom discuta pe larg despre tiparele de simplificare și vom face numeroase referiri la componentele acestora pentru a explica tipurile și modurile lor de acțiune.

Din acest motiv, pentru a asigura o cât mai mare claritate a textului vom face următoarele notații simbolice pentru tiparul de simplificare și componentele sale:

- ***tiparS*** – tipar de simplificare
- ***componentăTs*** – componentă de tipar de simplificare

4.2 Aspecte generale legate de sintaxa limbajului de reprezentare a tiparelor

În fișierele de configurare, secțiunile sunt specificate prin intermediul unor cuvinte cheie precedate de caracterul ”@”. Ele sunt structurate, după cum am amintit anterior, în două secțiuni principale, care la rândul lor conțin o serie de subsecțiuni.

Celor două secțiuni principale le corespund următoarele directivele:

1. @preprocess – pentru preprocesarea textului
2. @structure – pentru partea de reguli

Sintaxa completă a fiecăreia conține și numele nodului rădăcină (ex. @structure:MIXED)

Ordinea de apariție în fișier a celor două secțiuni este, în mod obligatoriu, cea menționată mai sus.

Sfârșitul oricărei secțiuni poate fi marcat prin introducerea cuvântului cheie *@end*, dar acest lucru nu este obligatoriu. Utilizarea sa este necesară doar atunci când sunt definite o serie de secțiuni imbricate.

Oriunde în cadrul fișierului pot apare comentarii:

- pe o singură linie, fiind precedate de secvența “//”
- pe mai multe linii, începutul fiind marcat de “/**”, iar finalul prin “**/”

4.3 Contribuții privind reprezentarea informației la preprocesarea textului

Regulile de preprocesare sunt grupate într-o secțiune subordonată, în partea de preprocesare, numită *@adjustments*. În versiunea curentă, a fost adăugată doar o singură regulă care se ocupă cu tratarea cuvintelor ce conțin diacritice.

De asemenea, ca metodă de preprocesare, există posibilitatea de a înlocui anumite secvențe de text cu altele. Aceste reguli sunt grupate in secțiunea *@replacement*, fiecare fiind scrisă pe câte o linie care conține textul ce va fi înlocuit și textul înlocuitor.

Ele sunt folosite pentru a genera în fișier secvențe de text complicate care se repetă în mai multe locuri. Astfel ele contribuie la îmbunătățirea lizibilității textului.

Aceste reguli de preprocesare lucrează la nivel de șiruri de caractere.

1. **LINKWORDS** - Reprezentarea tiparelor asociate cuvintelor de legătură

Cuvintele de legătură, așa cum am prezentat în capitolele anterioare, constituie principalele elemente de legătură în frază. Dată fiind importanța lor, tiparele asociate lor au fost definite într-un fișier separat pentru a putea satisface mai bine necesitățile lor de preprocesare.

Deși nu există nici o restricție în utilizarea integrală a facilităților de reprezentare a limbajului, tiparele sunt simple, deoarece cuvintele de legătură sunt formate din unul, rareori mai multe cuvinte. La acest aspect contribuie într-o bună măsură faptul că instrumentul de etichetare a părților de vorbire grupează prepozițiile complexe cu ajutorul caracterului tilda ”~” creând astfel un singur cuvânt compus (ex. ”*astfel~încât*”).

Tiparele cuvintelor de legătură sunt specificate în secțiunea *@structure:LINKWORDS* și sunt grupate după cele patru categorii principale descrise anterior: STRONG, MEDIUM, WEAK,

STOP folosind cuvântul cheie *@type*. De exemplu pentru categoria STRONG avem următoarea secvență:

```
@type:STRONG
  ...
@end
```

Limbajul permite definirea discontinua a unei categorii dacă acest lucru se dorește.

Definirea tiparelor și încadrarea lor în categorii se bazează pe observațiile obținute din analiza definițiilor din corpusul de antrenare. Majoritatea cuvintelor de legătură sunt prepoziții și cele mai utilizate au fost culese din corpusul de antrenare. Pentru obținerea unei liste complete s-a recurs la extragerea restului de prepoziții din dicționarul folosit de tagger (care există sub forma unui fișier text), folosind interogări de tip REGEX pentru izolarea cuvintelor de interes.

2. **MIXED** – Reprezentarea regulilor de simplificare și adnotare

Similar cazului anterior, fișierul de definire a regulilor de simplificare prezintă în secțiunea principală de structură, *@structure*, o serie de subsecțiuni în care regulile sunt grupate în funcție de scopul utilizării sau de momentul utilizării lor în cadrul procesului de creare a formei canonice pentru frazele prelucrate.

Sintaxa este aceeași ca în cazul definirii categoriilor de cuvinte de legătură. O subsecțiune de acest fel este marcată prin același cuvânt cheie *@type* urmat de denumirea ei (ex. @type:REMOVABLES) care identifică rolul jucat în cadrul procesului de transformare a frazei.

Pe lângă regulile de simplificare există și reguli ajutătoare, numite reguli de adnotare. Ele nu simplifică în mod direct fraza, ci doar etichetează anumite cuvinte sau construcții de cuvinte astfel încât să poată fi identificate ulterior și prelucrate în mod specific (ex. 4.1). În general, un tipar de simplificare va conține pe lângă regulile specifice de simplificare și unele de adnotare, cele două tipuri fiind de cele mai multe ori inseparabile.

Ex. 4.1 l:egal->OVRD ?= cu

Modalitatea de utilizare a tiparelor definite ține cont de ordinea în care acestea apar în fișierul de configurare. Acțiunea de obținere a formei canonice este un proces secvențial și în general aplicarea cu succes a unor reguli depinde de anumite etape anterioare.

3. **SETS** – Definirea componentelor de tipar complexe

Pe lângă fișierele principale de configurare a tiparelor, există un al treilea fișier secundar care permite definirea unor componente de tipar complexe (seturi de componente de tipar simple).

Trebuie precizat faptul că limbajul de definire a tiparelor furnizează și o modalitate localizată de a crea asemenea seturi, însă acestea sunt anonime, având valabilitate doar în cadrul tiparului din care fac parte.

Pe de altă parte, modalitatea de definire a seturilor în fișierul SETS deși este mai complexă prezintă următoarele avantaje notabile:

- asocierea unui nume pentru fiecare set astfel încât să poată fi folosit de oricâte ori este necesar, eliminând astfel restricția utilizării locale
- pe lângă elementele obișnuite care fac parte din seturile anonime și anume: cuvinte, leme și părți de vorbire, un set poate include alte seturi definite în fișierul de seturi.

Ca dezavantaj relativ putem considera faptul că definirea într-un fișier separat îngreunează înțelegerea imediată a tiparului în care este folosit un astfel de set.

Din punct de vedere al sintaxei, fișierul de definire a seturilor prezintă multe asemănări cu cele două fișiere de tipare, având:

- o parte de preprocesare care este identică cu cea a fișierelor de tipare
- o parte de definire a seturilor care este similară doar din punct de vedere al structurii

4.4 Tipare de simplificare/adnotare

Tiparele sunt alcătuite dintr-o serie de componente de tipar despărțite prin spații.

Există două categorii de componente:

1. **de control** - introduc diverse informații care influențează modul de interpretare a tiparelor
2. **de structură** - constituie elementele care vor fi efectiv folosite în cadrul procesului de testare a secvenței de cuvinte analizată

Din punct de vedere structural tiparele sunt formate din trei părți:

- precondiție
- tipar propriu-zis
- postcondiție

Dintre acestea, doar tiparul propriu-zis este obligatoriu să fie prezent, precondiția și post-condiția fiind opționale. Deși aceste trei părți sunt identice din punct de vedere al sintaxei, ele diferă prin modul de interpretare, ce ține de interacțiunea cu cuvintele din fraza analizată, în cadrul procesului de aliniere.

De obicei tiparele sunt suficient de scurte pentru a încăpea pe o singură linie, însă limbajul oferă facilitatea de a scrie tiparul pe mai multe linii folosind un operator special.

4.5 Sintaxa de reprezentare a tiparelor

ComponenteleTs de control

Acestea controlează modul în care tiparul funcționează ca ansamblu, motiv pentru care le-am putea privi ca fiind meta-elemente.

În tabelul 4-1 sunt prezentate pe scurt aceste componente de control:

Tabelul 4-1 ComponenteTs de control

Componentă de tipar	Descriere
?=	precede precondiția sau post-condiția
=	precede partea principală a tiparului
#<nume>	asociază un nume tiparului.
@<directiva>	asociază o anumită directivă

ComponentaTs ”=” este necesar să apară doar în cazurile în care se face o tranziție între două părți ale tiparului (ex. tiparul este format din precondiție și tipar propriu-zis), în caz contrar ea poate lipsi. Pe de altă parte, în implementarea curentă, ”*?*=” este necesar să fie prezentă întotdeauna pentru a marca începutul precondiției și al post-condiției.

Dacă dorim să utilizăm un același tipar în mai multe locuri în cadrul altor tipare, fără a fi nevoiți să-l scriem de fiecare dată în întregime, este necesar să i se atribuie un nume prin intermediul

componenteiTs de control ” #<*nume*>”. Componenta de nume este necesar să se găsească pe prima poziție în cadrul tiparului.

ComponentaTs de control precedată de ”@” apare de asemenea pe prima poziție. Aceasta are un statut mai special deoarece prezintă un caracter polimorfic prin faptul că poate introduce o serie de directive care îndeplinesc diverse funcții extrem de diferite:

- @debug<n> – unde <n> este un număr natural care marchează tiparul respectiv pentru a putea fi identificat în momentul activării sale
- @repetable – modifică modul de aplicare a tiparului încercându-se reaplicarea sa de câte ori este posibil înainte de a trece la următorul tipar

ComponenteTs structurale

ComponenteleTs de structură sunt de trei feluri:

1. de tip sub-tipar - constă doar dintr-o referință de nume către un tipar definit anterior și marcat cu nume pentru a fi folosit ulterior mai ușor
2. de tip set - grupuri de *componenteTs* simple
3. *componenteTs* simple - componentele de bază ale *tiparelorS*

O *componentăTs* simplă conține nucleul, o serie de modificatori și adnotații.

Nucleul este partea principală a unei *componenteTs* simple. La rândul său, acesta conține următoarele elemente:

- un prefix care specifică tipul *componenteiTs* structurale – poate fi ”*w*”, ”*l*” sau ”*p*”
- cuvântul, lema, partea de vorbire în cauză
- o etichetă MSD prin care se poate stabili precizia de comparare

Pentru a conecta și totodată a face distincție dintre prefix și cuvânt, lema sau partea de vorbire se folosește caracterul ”*:*” (ex. p:ADVER)

Utilizarea prefixului ”*w*” pentru cuvânt este opțională deoarece în cazul în care nu există prefix se consideră implicit că respectiva *componentăTs* este de tip cuvânt.

Specificarea părților de vorbire se face utilizând doar primele 5 caractere din denumirea lor. Dacă denumirea e mai scurtă, ea rămâne netrunchiată.

Ca sintaxă de integrare a etichetei MSD în cadrul componentei se folosește caracterul ”/” care încadrează eticheta fără spații înainte și după. În cadrul operației de testare, potrivirea etichetelor MSD este realizată prin comparații de tip REGEX.

Pentru cazurile când nu ne interesează tipul nucleului, se poate folosi tipul necunoscut. În implementarea curentă este permisă utilizarea acestui tip doar în combinație cu un test de adnotare care să verifice existența unei anumite etichete.

Modificatorii sunt operatori care controlează modul de interpretare a *componenteiTs*.

Există două categorii de modificatori care pot fi identificați în funcție de plasamentul lor relativ la nucleul componentei:

- de control – localizați înainte de partea principală
- operativi – plasați după partea principală

În tabelul 4-2 sunt enumerați modificatorii disponibili.

Tabelul 4-2 Modificatori ai componentelorTs

Categorie	**Funcție**	**Simbol**
de control	*componentăTs* țintă	>>
	componentăTs exclusă	!
operativi	*componentăTs* recurentă	+
	componentăTs opțională	?

Modificatorul *componentăTs* **țintă** marchează componenta principală a tiparului respectiv. El poate fi folosit doar în partea principală a tiparului., având două utilități:

- în cazul tiparelor care simplifică fraza, această componentă marchează cuvântul principal folosit ca element de substituție
- în cazul anumitor tipare de adnotare cu etichete implicite, el marchează cuvântul care va fi adnotat

Obs.
1. Într-un tipar nu poate exista decât o singură componentă țintă
2. Este perfect posibil ca într-un tipar să se aplice cuvântului țintă atât operația de simplificare, cât și cea de adnotare.

Modificatorul *componentăTs* **exclusă** marchează componentele care vor fi ignorate la operația de simplificare descrisă în cadrul unui tipar.

Modificatorul *componentăTs* **recurentă** este un element care permite potrivirea cu oricâte cuvinte de un același tip.

Modificatorul *componentăTs* **opțională** stabilește faptul că este permis ca respectiva componentă să fie ignorată la testare dacă este necesar.

Adnotațiile asigură o modalitate de păstrare a unor informații rezultate în urma aplicării unui tipar.

Utilizarea adnotațiilor presupune două tipuri de operații:

- definirea – o anumită etichetă este atribuită unui cuvânt
- testarea – se verifică dacă un anumit cuvânt are o etichetă anume

Tabelul 4-3 prezintă operatorii implicați în procesul de adnotare.

Tabelul 4-3 Operatori de adnotare

Tip operație	Funcție specifică	Simbol
definire	aditiv	->
	de suprascriere	->*
	definire element de tipar	->:
testare	pozitivă	#
	negativă	#!

Operatorii **de definire** sunt cei care pot asocia anumite informații simbolice prin adnotare, cuvintelor din fraza de simplificat. Un cuvânt poate avea un număr nelimitat de adnotații. Modalitatea de adăugare aditivă este principala operațiune care facilitează acest lucru.

Prin contrast, operatorul de suprascriere elimină toate adnotațiile anterioare, înlocuindu-le cu cea nouă furnizată ca parametru.

Pe lângă operația de adnotare obișnuită, cu ajutorul operatorului de definire se poate forța definirea șirului de caractere ce va constitui eticheta de tipar simplu, anulând astfel aplicarea procedurii obișnuite de obținere a acesteia.

Operatorii **de testare** verifică prezența (în cazul operatorului de testare pozitivă) sau absența (pentru cel de testare negativă) unei adnotații.

Obs. Sintaxa permite definirea sau testarea simultană a mai multor adnotații, folosind operatorul pipe ”|”. În cazul testării, fiecare etichetă testată poate fi negativă sau pozitivă după necesități. În acest sens, operatorul ”!” va fi asociat fiecărei adnotații pentru care se cere un test negativ:

Ex. 4.2 p:NOUN#ART_U|!OBLQ

4.6 Concluzii

Lista de contribuții din capitolul 4:

- crearea unui limbaj de reprezentare a tiparelor de simplificare a frazelor

În acest capitol este prezentat un limbaj special creat pentru reprezentarea unor tipare care permit simplificarea frazelor astfel încât să fie aduse la o formă canonică ce va permite crearea unor tipare cu grad de complexitate mai redus.

Acest limbaj definește o serie de elemente sintactice centrate pe termen care specifică modul de interacțiune și potrivire dintre elementele tiparului și termenii din fraza analizată.

Regulile de simplificare a frazei se află în trei fișiere de configurație, fiecare cu specificul propriu, denumite LINKWORDS, MIXED și SETS.

Fișierul LINKWORDS conține tiparele asociate cuvintelor de legătură. Ele sunt grupate pe secțiuni în funcție de tipul cuvântului de legătură.

Fișierul MIXED conține tiparele pentru regulile de simplificare și adnotare. Acestea sunt reguli de complexitate ridicată și sunt grupate în funcție de etapa pentru care au fost concepute. În general între aceste etape există o ordine bine stabilită deoarece rezultatele anumitor procesări pot și, în general, vor fi utilizate în etapele ulterioare. Spre deosebire de ordinea etapelor care este dictată programatic în cadrul aplicației, ordinea regulilor în cadrul unei etape este stabilită pe baza succesiunii în care apar în fișierul de configurare.

În fișierul SETS pot fi definite componente de tipar complexe ce pot fi folosite ulterior în cadrul tiparelor. Acestea sunt colecții de date complexe care pot conține cuvinte, părți de vorbire, precum și alte seturi.

Sintaxa tiparelor definește două tipuri de componente de tipar: *de control* - care controlează modul de funcționare al tiparului și *de structură* - care sunt elementele propriu-zise de comparație.

Componentele de structură au un nucleu care determină tipul de comparație, precum și o serie de modificatori care stabilesc modul în care se va realiza această procedură. Fiecare componentă structurală poate efectua operații de adnotare asupra termenilor din frază asociați ei sau poate verifica dacă termenul din frază are o anumită adnotație.

5 Contribuții privind procesul de obținere a tiparelor simple

5.1 Introducere

Etapele principale ale metodei de detecție a definițiilor pentru limba română sunt în mare măsură aceleași cu cele ale metodei WCL a cărei diagramă a fost prezentată în capitolul 2 (figura 2-1). Diferențe demne de menționat se înregistrează în doar două dintre aceste etape.

Prima diferență și de departe cea mai importantă, este, după cum am mai precizat, faptul că nu se mai creează tiparul stea, ci un tipar echivalent numit *tipar simplu*. Prezentăm în figura 5-1 sub-etapele procesului de creare a tiparului simplu, ce constituie o contribuție importantă a acestei lucrări.

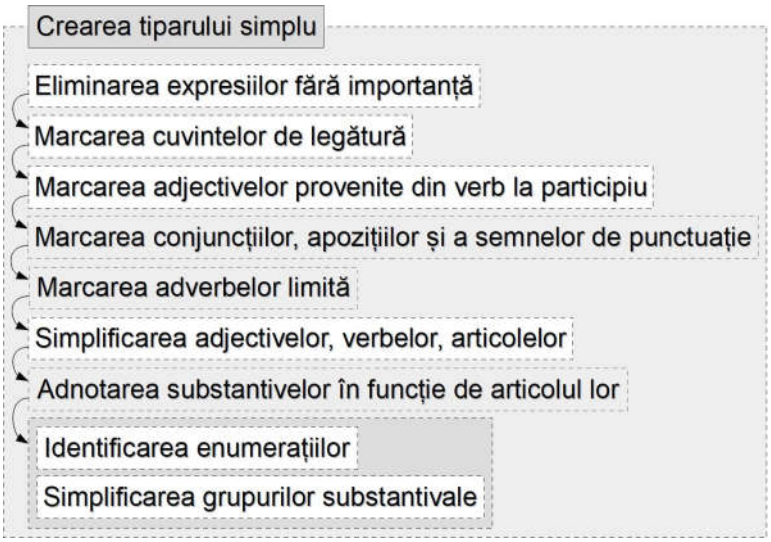


Figura 5-1 Etapele de creare a tiparului simplu

A doua diferență importantă se află în cadrul preprocesării, în etapa de etichetare a părților de vorbire. Înainte de a trimite fraza către instrumentul de etichetare spre a fi analizată se realizează după caz operații de ajustare specifice asupra anumitor cuvinte, secvențe de cuvinte sau semne de punctuație care au rolul de a ajuta la identificarea corectă părților de vorbire corespondente (figura 5-2).

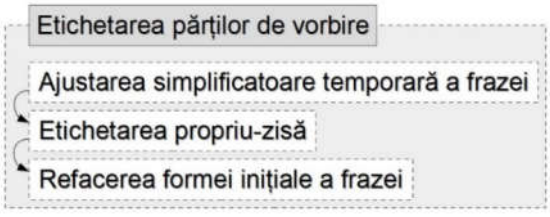


Figura 5-2 Operații speciale de preprocesare

În acest capitol vom vorbi foarte mult despre două tipuri de tipare. Primul este *tiparul de simplificare* despre care am discutat și în capitolul 4, iar cel de-al doilea este *tiparul în*

prelucrare care se referă la iterația curentă a tiparului ce corespunde frazei analizate, pornind de la tiparul inițial, numit *tipar extins* și până la forma finală de *tipar simplu*.

Similar cu notațiile pe care le-am efectuat în capitolul 4 vom utiliza forme prescurtate, simbolice pentru a face referire la tipar și la componentele sale:

- ***tiparP*** – tipar în prelucrare
- ***componentăTp*** – componentă de tipar în prelucrare

De asemenea pentru a face referire la o componentă de tipar generică vom folosi notația ***componentăT***.

5.2 Reprezentarea interna a *tiparuluiP* - structură, operații

Pentru crearea tiparului unei fraze este necesară o structură care să permită aplicarea operațiilor de simplificare necesare, fiind capabilă să ”memoreze” toate etapele succesive de transformare, începând cu tiparul extins până la forma de tipar simplu. Această structură a fost special concepută în acest scop, fiind implementată prin intermediul unei latici de *componenteT*, care inițial corespund termenilor din fraza analizată (figura 5-3). Această latice este un graf direcționat aciclic care are un singur punct de început. Nodurile sale sunt de două tipuri:

- noduri de date – conțin *componenteleT*
- noduri de legătură

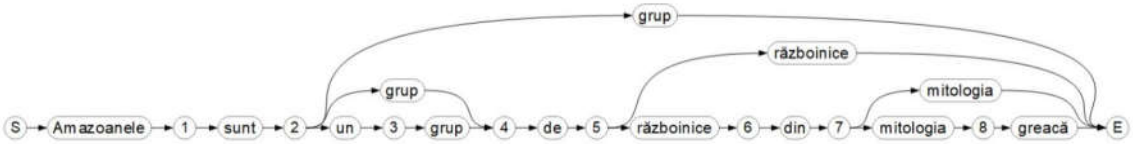


Figura 5-3 Latice de componenteT

Fiecare nod de legătură are cel puțin o conexiune de intrare și una de ieșire. În momentul când se realizează o operație de simplificare, se creează un nou nod care moștenește datele *componenteiTp* țintă, fiind adăugat în graf între cele două noduri de legătură care se află la extremitățile secvenței ce a fost simplificată.

Forma curentă a tiparului pe care am denumit-o *tiparP* este formată întotdeauna din componentele aflate pe nivelurile cele mai de sus.

Reprezentarea internă suportă traversarea structurii atât către dreapta cât și către stânga, fapt ce permite realizarea operațiilor care altfel ar fi greu, dacă nu chiar imposibil de realizat.

Pentru realizarea operațiilor necesare procesului de obținere a tiparului formei canonice au fost definite o serie de operații simple de transformare a structurii laticii. Ele acționează în principal asupra nodurilor de date, iar nodurile de legătură, care din punct de vedere al informației utile sunt transparente, suferă modificări colaterale conexe.

Având în vedere faptul că aceste operații acționează și modifică structura de *componenteTp* există restricția care împiedică orice fel de operație asupra unor componente care nu se află pe ultimul nivel.

Operațiile atomice definite sunt următoarele:

1. Simplificarea unei secvențe de noduri
2. Eliminarea de noduri
3. Adăugarea de noduri
4. Interschimbarea de noduri

5. Repoziționarea unui nod înainte/după un nod țintă

5.3 Activarea regulilor de simplificare a tiparului

Testarea tiparelor este operația prin care tiparele definite în fișierele de configurare a regulilor (în primul rând MIXED folosind seturile definite în SETS, dar și LINKWORDS) sunt activate în mod succesiv și testate pentru a vedea dacă se potrivesc cu o anumită secvență de cuvinte din frază.

În procesul de comparare a *tiparuluiS* cu *tiparulP*, operația de testare este finalizată cu succes în momentul în care toate testele parțiale specificate de elementele componente ale *tiparuluiS* sunt îndeplinite. Deoarece există elemente de tipar opționale și/sau repetitive nu se poate determina apriori numărul de *componenteTp* care vor fi afectate în final.

Oricare ar fi structura unui *tiparS*, există patru moduri de a iniția testarea. Acestea sunt determinate de poziția punctului de ancorare a căutării, care determină *componentaTs* de unde începe testarea, element care va corespunde *componenteiTp* de lucru curente.

Cele trei părți structurale (precondiție, tipar propriu-zis, post-condiție) sunt testate separat, modul în care acest lucru se realizează depinzând de tipul de testare ales.

În funcție de tipul de testare, mai precis de modul de ancorare, *componenteleTp* selectate de *tiparulS* pot fi grupate în două categorii:

- active – acele *componenteTp* care sunt considerate prelucrate (ex. cele care suferă o operație de simplificare). Ele nu vor mai fi accesibile comparațiilor ulterioare.
- martor – sunt *componenteTp* care au rolul de context pentru prima categorie. Absența lor invalidează aplicarea tiparului pe secvența țintă. De asemenea ele rămân disponibile pentru aplicarea altor *tipareS*.

O *componentăTs* poate specifica unul dintre următoarele trei tipuri de comparație, enumerate în ordinea specificității: nivel cuvânt, nivel lemă, nivel parte de vorbire.

Sintaxa curentă a limbajului de descriere a regulilor nu permite asocierea mai multor tipuri de comparație aceluiași element de tipar. Pentru a nuanța totuși această restricție se poate folosi un test asupra etichetei compacte MSD care include toate caracteristicile cuvântului (*componentăTp*) respectiv.

În practică, ordinea de testare a este: lemă, cuvânt, parte de vorbire. Această ordine aparent neoptimizată se explică prin faptul că un număr mult mai mare de *componenteTs* folosesc comparații la nivel de lemă.

Pe lângă testele de structură se aplică și testele logice care implică adnotațiile. Într-o *componentăTs* se poate cere ca în vederea potrivirii, *componentaTp* trebuie să dețină anumite adnotații.

Testarea potrivirii cu un set de componente este o operație secvențială în care se testează pe rând, după procedura specificată mai sus, fiecare componentă simplă a setului până la găsirea unei potriviri sau până la epuizarea tuturor opțiunilor.

În final, testarea unui tipar inclus într-o *componentăTs* presupune o apelare recursivă a metodei inițiale de testare a tiparelor.

5.4 Preprocesare

Această operație de preprocesare este realizată în scopul îmbunătățirii operației de etichetare a părților de vorbire. S-a observat că rezultatul acestei operații este afectat negativ într-un mod direct proporțional cu complexitatea frazei analizate. Acest aspect poate fi îmbunătățit prin

efectuarea anumitor modificări, care să afecteze cât mai puțin posibil structura frazei ce urmează să fie analizată.

Modificările menționate mai sus pot fi grupate în următoarele trei categorii:

1. eliminarea unor construcții
2. ajustarea anumitor expresii sau cuvinte
3. reformatarea semnelor de punctuație

Construcțiile care sunt menționate la punctul (1) pot fi incluse în categoria informațiilor suplimentare introduse prin intermediul parantezelor și intervin într-un mod negativ în procesul de analiză al părților de vorbire.

Expresiile și cuvintele de la punctul (2) sunt modificate astfel încât identificarea părților lor de vorbire să poată fi realizată cu succes. Ele nu vor fi eliminate din frază, asupra lor se face doar o operație de transformare simplificatoare. Această formă simplă este furnizată apoi instrumentului de etichetare care va putea astfel să identifice corect partea de vorbire. La finalul procesului se restaurează forma lor complexă inițială.

Punctul (3) se referă la anumite semne de punctuație care pot crea confuzii. Ca exemple putem menționa caracterele apostrof folosit în cadrul numelor proprii (ex. D'Artagnan), sau punctul (".") și virgula (","), folosite în cadrul numeralelor scrise sub formă numerică (ex. 23.564).

În cadrul etapei de etichetare tagger-ul furnizează o serie de informații din care sunt extrase doar următoarele:

- cuvântul
- lema
- partea de vorbire
- eticheta specială MSD
- marcajele grupurilor substantivale

În final, după cum am menționat mai sus este refăcut textul original.

5.5 Etapele procesului de simplificare a frazei

Operația de obținere a tiparului simplu, care corespunde formei canonice a frazei, presupune aplicarea succesivă a etapelor specificate în secțiunilor din fișierul de reguli MIXED, însoțită de tiparele ce permit identificarea cuvintelor de legătură aflate în LINKWORDS.

Etapele care contribuie la obținerea tiparului simplu corespund uneia dintre categoriile de reguli definite în fișierele de configurare a regulilor. Pentru a fi ușor accesibile aceste categorii sunt reprezentate sub forma unor structuri de tip arbore, având două feluri de noduri:

- noduri de control – specifică informațiile din subordinea lor
- noduri de date – conțin câte un *tiparS* de lucru

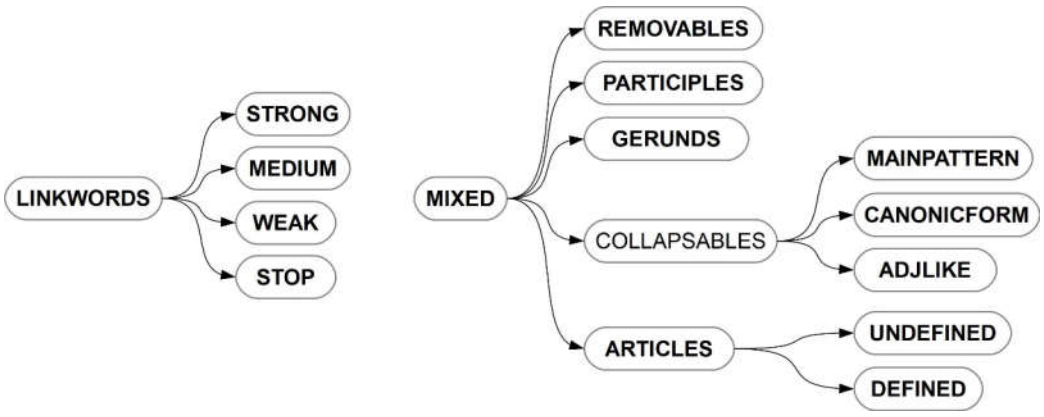
În implementarea curentă sunt create două asemenea structuri de tip arbore numite *linkwords* și *mixed* după numele fișierelor de reguli de unde sunt preluate datele (figura 5-4).

În figura 5-4 sunt prezentate categoriile principale folosite în acest moment în cadrul metodei. Fiecărui nod i se asociază la creare calea în arbore sub formă de șir de caractere așa cum se vede în exemplul Ex. 5.1. Aceasta este stocată într-o listă ce poate fi ușor interogată.

Ex. 5.1 MIXED/COLLAPSABLES/MAINPATTERN

Ambele tipuri de noduri din arbore sunt în esență identice având aceleași date componente, doar că sunt dezvoltate în mod specific rolului lor:

- nodurile de control primesc un nume și vor avea o listă de noduri de control subordonate și una de noduri de date.
- nodurile de date nu au nume și nici noduri subordonate, în schimb fiecare conține un *tiparS*.



(a) (b)

Figura 5-4 Structura fișierelor de configurare a regulilor

5.6 Eliminarea expresiilor fără aport informațional

Etapa utilizează tiparele din categoria MIXED/REMOVABLES.

Această primă etapă presupune eliminarea anumitor cuvinte și expresii care nu au nici o contribuție, din punct de vedere informațional, în cadrul unei fraze definiționale. Acestea sunt în mare parte anumite adverbe sau locuțiuni adverbiale folosite pentru a sublinia sau puncta anumite idei (ex. de~asemenea, în~special, de regulă, etc.). Deoarece nu fac efectiv parte din text, ele întrerup succesiunea naturală a cuvintelor, având un impact negativ asupra analizei frazei.

5.7 Marcarea cuvintelor de legătură

Această etapă utilizează toate tiparele din categoria LINKWORDS.

În fișierul de configurare nu există tipare definite direct în această categorie generală. Toate tiparele rezidă în categoriile STRONG, MEDIUM, WEAK și STOP care sunt incluse în ea. Există însă un mecanism prin care se pot accesa toate tiparele incluse într-o categorie, inclusiv cele definite în categoriile subordonate, fapt utilizat în această etapă, deoarece se dorește marcarea tuturor cuvintelor de legătură, indiferent de categoria din care fac parte.

5.8 Marcarea adjectivelor provenite din verb la modul participiu

Categoria corespondentă acestei etape este MIXED/PARTICIPLE.

Acest tip de adnotație nu se referă neapărat la partea de vorbire în special, ci mai curând marchează rolul pe care-l joacă în frază acest tip de adjective, din punct de vedere al înțelesului. Din acest motiv au fost create reguli de recunoaștere pentru anumite cuvinte cu un rol similar (ex. *capabil* de, *egal* cu).

Pe lângă verificarea tiparului specific, este necesară verificarea suplimentară pentru a se determina dacă adjectivul provine din verb la participiu, motiv pentru care a fost creată o listă de aproximativ 8000 de cuvinte extrasă din dicționarul instrumentului de etichetare.

5.9 Marcarea conjuncțiilor și a semnelor de punctuație

Acestea sunt de fapt două etape distincte dar, dat fiind faptul că sunt foarte simple și în esență identice, vor fi prezentate împreună. Fiind atât de simple, nici una dintre ele nu are asociată o categorie de tipare, ci în schimb marcarea cuvintelor țintă se face doar pe baza verificării părții de vorbire. Ca semne de punctuație ne interesează doar virgula.

5.10 Marcarea adverbelor limită

În cadrul acestei etape sunt adnotate anumite adverbe care fac parte din grupurile substantivale, pentru a ajuta la identificarea anumitor puncte care pot constitui limita dintre substantivul care este supra-conceptul definiției și atributele sale care nu sunt considerate destul de importante pentru a face parte din supra-concept.

5.11 Operații de simplificare pregătitoare

Acestei etape îi corespunde categoria MIXED/COLLAPSABLES/MAINPATTERN.

Tiparele definite aici sunt responsabile cu simplificarea intermediară a atributelor care stau pe lângă substantive. Ele sunt grupate în funcție de părțile de vorbire pe care le afectează în principal.

Scopul urmărit în aplicarea tiparelor unei părți de vorbire este obținerea unei *componenteTp* simplificatoare care să cuprindă toate cuvintele direct conectate semantic de respectiva parte de vorbire.

Obs. Deși am vorbit de conexiune semantică informațiile folosite pentru determinarea acestui aspect sunt totuși de natură sintactică, motiv pentru care deciziile de simplificare, deși optimizate pe cât posibil la maxim, pot și vor da în anumite cazuri limită rezultate incorecte.

Adjective

Prima categorie de părți de vorbire afectată sunt adjectivele deoarece sunt principala formă de exprimare a atributelor atât pentru substantive cât și pentru verbele la diateza pasivă.

În urma fiecărei operații de simplificare se obține o *componentăTp* simplificatoare de tip adjectiv. Sunt tratate gradele de comparație, simplificarea adverbelor ce țin de adjective.

Verbe

A doua grupă abordată în cadrul acestei etape sunt verbele. În afara verbelor specifice definițiilor care sunt foarte precis consemnate, restul nu au o importanță prea mare, dar necesită totuși un minim de prelucrare. În general, chiar dacă apar într-o frază care conține o definiție, ele nu ajung să facă parte din tiparul extras.

Simplificările care se aplică verbelor sunt consolidarea formelor verbale care au mai multe cuvinte: timpul viitor, trecut, etc.

Articole

O altă grupă importantă, cea a articolelor, procesează articolele hotărâte și nehotărâte.

Spre deosebire grupele anterioare, deși atenția este axată asupra acestei categorii de părți de vorbire, *tipareleS* vor face simplificarea prin intermediul *componentelorTp* de tip substantiv.

Datorită faptului că articolele nehotărâte dispar, este necesară adnotarea *componenteiTp* simplificatoare rezultată într-un mod specific, pentru a nu pierde această informație. De

asemenea, pentru a avea o uniformitate și ușurință în prelucrare, sunt adnotate similar substantivele articulate cu articol hotărât chiar dacă această formă nu se pierde, fiind inclusă în structura lor.

Conjuncții

Deoarece prelucrarea conjuncțiilor este limitată, transformările descrise nu au un impact notabil ca cele de până acum. Secvențele vizate sunt virgula urmată de conjuncțiile ”*dar*” și ”*însă*”, precum și cazurile în care aceleași conjuncții preced conjuncția ”*și*”.

5.12 Marcarea unor limite în grupurile substantivale

Categoria de tipare asociată acestei etape este MIXED/NPLIMIT.

Această etapă de adnotare este necesară pentru a marca o limită de natură semantică în cadrul grupurilor substantivale, ce delimitează o porțiune de interes crescut de restul expresiei care este considerată mai puțin importantă.

5.13 Adnotarea substantivelor în funcție de articolul asociat

Această etapă este alcătuită din două sub-etape:

1. substantivelor articulate cu articol nehotărât – MIXED/ARTICLE/UNDEFINED
2. substantivelor însoțite de articol hotărât – MIXED/ARTICLE/DEFINED

Dat fiind faptul că această etapă are un scop bine determinat și totodată simplu deoarece *componenteleTp* implicate au suferit deja o serie de alte transformări pregătitoare, tiparele din această categorie sunt foarte simple. Adnotațiile aplicate substantivelor țintă (formă articulată/nearticulată) sunt implicite datorită împărțirii în cele două sub-etape.

5.14 Generalități legate operația finală de simplificare a grupurilor substantivale

Acestei etape nu îi corespunde o anumită categorie în fișierul de configurare a tiparelor. Fiind o operație foarte complexă, ea necesită mai multe iterații, precum și păstrarea unor informații de stare de la o sub-etapă la alta. Implementarea finală a acestei operații de simplificare necesită metode de recursivitate dublu înlănțuite. Evident această necesară complexitate a făcut imposibilă implementarea operației folosind doar mecanismele simple oferite de limbajul de reprezentare a tiparelor. După cum am arătat anterior acesta este centrat pe componenta de tipar și deși s-ar putea face anumite de artificii la nivel de tipar, el nu are facilități de a transmite o stare între două operații de potrivire a tiparelor, iar despre recursivitate nici nu poate fi vorba.

Etapa de simplificare a grupurilor substantivale constă din două sub-etape:

- etapă de detecție a enumerațiilor în text
- etapa de simplificare propriu-zisă

În etapa de detecție a enumerațiilor este analizată fraza pentru a găsi grupuri de substantive și adjective asemănătoare din punct de vedere al rolului pe care-l îndeplinesc, Cu alte cuvinte se dorește o grupare pe considerente semantice a acestora. Gruparea lor permite tratarea acestor cuvinte ca singură unitate, fapt ce simplifică analiza ulterioară a frazei.

Faza de simplificare propriu-zisă utilizează informațiile dobândite la detecția enumerațiilor pentru a simplifica progresiv mai întâi grupurile substantivale, iar apoi fraza în sine, ceea ce duce la atingerea scopului final, obținerea formei sale canonice și a tiparului său simplu.

5.15 Etapa de identificare a enumerațiilor

În această etapă se urmărește, așa cum s-a menționat anterior gruparea cuvintelor cu același rol. Deoarece nu avem efectiv acces la informații de natură semantică, se face apel la orice fel de informații de natură sintactică sau lexicală, care sunt disponibile, pentru a deduce pe cât posibil rolul cuvintelor și legăturile dintre ele, prin intermediul descoperirii unor tipare comune de utilizare.

Părțile de vorbire care prezintă interes în etapa de detecție a enumerațiilor sunt:

- substantivele – elemente principale în jurul cărora se grupează restul cuvintelor
- adjectivele – aceasta este etapa în care adjectivele vor fi în final simplificate prin înglobarea în substantivul care-l deservesc
- virgulele – cu important rol în controlul procesului de creare a enumerațiilor
- conjuncțiile specifice *și*/*sau* – elemente de control terminale în enumerații

Metoda descrisă aici face distincție între trei tipuri de enumerații:

1. enumerații de grupuri substantivale
2. enumerații de adjective
3. enumerații de adjective cu formă de verb la participiu

Distincția care se face între adjectivele obișnuite și cele provenite din participiu ține de faptul că ultimele au un rol specific în procesul de creare a tiparului simplu.

Pe lângă identificarea enumerațiilor, această metodă urmărește reevaluarea dependențelor între termenii de tip substantiv din cadrul unui grup substantival complex. Analiza tuturor grupurilor substantivale din frază constituie astfel o primă fază.

După aceasta este necesară analiza globală a frazei pentru a se determina dacă nu cumva există enumerații între grupurile substantivale principale. Această a doua fază se realizează tot cu ajutorul metodei de identificare a enumerațiilor, folosind un set modificat de constrângeri.

Dată fiind secvența de termeni care trebuie analizată, metoda de detecție a enumerațiilor urmărește să găsească potrivirile optime între acești termeni pentru a determina dacă este necesară gruparea lor și în caz afirmativ care este cea mai bună posibilitate de a-i grupa. Pentru aceasta vor fi generate mai multe *structuri candidat* care reflectă aceste variante de grupare și care vor fi evaluate pe baza unor scoruri de fiabilitate.

În linii foarte generale metoda de detecție a enumerațiilor are următoarelor etape:

1. se parcurg termenii de analizat până la întâlnirea unui substantiv sau adjectiv, construind astfel o sub-secvență de lucru
2. se lansează procedura de adăugare a sub-secvenței respective – care se consideră că adaugă un nou termen principal
3. se evaluează toate variantele nou obținute, pentru actualizarea scorurilor parțiale
4. se reia de la etapa 1 până la epuizarea tuturor termenilor din secvență
5. se ordonează variantele finale pentru aflarea celei optime

Scoruri de fiabilitate

Metoda de detecție a enumerațiilor generează în mod recursiv toate structurile candidat posibile rezultate în urma diverselor moduri de asociere a termenilor din secvența de intrare, pe baza unui set de constrângeri. Fiecare structură candidat primește scoruri de fiabilitate pentru o serie de cinci categorii:

1. scor de similitudine – măsoară gradul de asemănare dintre două substantive

2. scor de structură – evaluează părțile componente ale enumerației și modul în care sunt ele aranjate
3. scor de discrepanță – este un handicap cumulativ acordat enumerației pentru cazurile în care se fac asocieri suboptimale între substantive
4. scor de distanță – se calculează în funcție de distanța în frază dintre noul substantiv și cel cu care este asociat
5. scorul general – este calculat însumând scorurile de similitudine, structură și discrepanță

Calculul acestor scoruri este controlat prin intermediul unor constante numerice și comutatoare booleene elaborate empiric. În implementarea curentă există două asemenea seturi de constante specifice celor două faze distincte ale operației finale de simplificare:

- faza de simplificare a conținutului grupurilor substantivale – constantele și comutatoarele sunt mai permisive oferind o mai mare libertate de asociere a termenilor
- faza de simplificare a frazei – anumite restricții sunt impuse pentru a obține enumerații de substantive cu grad de similitudine maxim

În anexa II este dat listingul celor două grupe de constante numerice.

1. **Scorul de similitudine**

Scorul de similitudine se calculează între două substantive, luând în calcul o serie de caracteristici care sunt prezentate în tabelul 5-1 și care, după cum se poate observa, sunt de natură lexico-sintactică.

Tabelul 5-1 Caracteristici de comparare pentru scorul de similitudine

Caracteristică	**Valori**
tip substantiv	comun, propriu, necunoscut
tip cuvânt de legătură	tare, mediu, slab, stop, deloc, necunoscut
tip de articol	definit, nedefinit, inexistent, necunoscut
formă (caz) substantiv	direct (nominativ-acuzativ), oblic (dativ-genitiv)

Scorul final obținut are valori în intervalul 0 (nepotrivire totală) și 100 (potrivire totală).

Dacă cele două substantive sunt însoțite chiar de același cuvânt de legătură, scorul de similitudine primește un bonus, având în vedere faptul de a nu se depăși limita maximă admisă.

Este, de asemenea, important să precizăm că această operație de calcul a similarității nu este comutativă.

$$SIMILARITY(subst1, subst2) \neq SIMILARITY(subst2, subst1)$$

2. **Scorul de structură**

Acest scor este calculat în mod iterativ pe măsură ce elementele sunt adăugate la enumerație. În acest scop au fost definite o serie de reguli care contribuie prin scoruri parțiale pozitive sau negative care se însumează pentru obținerea valorii finale.

Tabelul 5-2 prezintă setul de reguli precum și contribuția lor valorică.

Tabelul 5-2 Reguli pentru scorul de structură

Nr. Crt.	**Regulă**	**Valoare**
1.	Dacă există două sau mai multe conjuncții *și*/*sau* de același fel	-100

2.	Dacă există atât conjuncția *și* cât și conjuncția *sau*	-2000
3.	Dacă există virgulă fără a avea una dintre conjuncțiile *și*/*sau*	-50
4.	Dacă există o virgulă urmată de conjuncție se adaugă un bonus pentru potriviri cu substantive mai îndepărtate	25
5.	Dacă enumerația conține un singur substantiv despărțit prin virgulă de o altă enumerație al cărei tip compus este foarte similar cu acest substantiv	-500

3. **Scorul de discrepanță**

În faza de detecție a grupurilor substantivale, instrumentul de etichetare analizează cuvintele de la stânga la dreapta. Asocierile dintre cuvinte se fac doar pe baza genului, numărului și a cazului. Un termen nou este întotdeauna asociat cu ultimul termen cu care se potrivește, motiv pentru care rezultă o structură în care un termen nou este aproape întotdeauna inclus în cel de dinaintea lui, ceea ce poate duce la erori de asociere.

Metoda se bazează pe faptul că, în general, oamenii tind să folosească într-o aceeași enumerație termeni care seamănă foarte mult între ei, această asemănare fiind descrisă de scorul de similaritate.

Scorul de discrepanță poate avea valoarea zero în cazul unei potriviri care este considerată optimă sau valori negative, acestea crescând pe măsură ce nepotrivirile se acumulează.

4. **Scorul de distanță**

Corelat cu scorul de discrepanță se calculează și un scor de distanță care măsoară depărtarea dintre indexul de termen al elementului care se adaugă și indexul celui cu care este asociat. Acest scor a fost introdus în ideea că, dacă avem aceeași similaritate față de două elemente, elementul cel mai potrivit este cel mai apropiat.

Procedura de adăugare a unui nou termen la structurile candidat

Termenii care determină lansarea procesului de adăugare sunt substantivele sau adjectivele. Orice altceva este tratat ca element de control (virgule, conjuncții ”*și/sau*”) sau nu manifestă nici un interes.

Elementele cheie complexe în această structură în care termenii se adaugă secvențial sunt de tip *enumerație* și *grup substantival*. Nici un element simplu nu poate fi adăugat altfel decât ca făcând parte dintr-unul dintre acestea două.

Substantivele marchează întotdeauna începutul unui nou grup substantival deci ele pot fi adăugate imediat într-o enumerație sau ca element subordonat al unui alt grup substantival. Pe de altă parte adjectivele trebuie să fie adăugate unor grupuri substantivale deja existente.

Procesul de adăugare se realizează sub forma unei căutări recursive în structura de enumerații și grupuri substantivale creată până în acel moment, deoarece fiecare nouă structură este inclusă în una mai veche și în general este necesar să se ajungă la ”nivelul” cel mai profund.

Un anumit termen poate fi atașat în mai multe feluri în cadrul structurii, determinând la fiecare pas creșterea numărului de variante care vor trebui explorate la pașii următori.

Adăugarea se desfășoară diferit în funcție de anumiți factori de control, precum și de tipul termenului de adăugat:

1. adăugarea adjectivelor – după cum am spus ele nu pot fi adăugate decât pe lângă substantivul de care aparțin și nu se reevaluează dependența față de acesta. Structura este parcursă recursiv până la găsirea grupului substantival părinte, in care adjectivul este adăugat la sfârșit.

2. adăugarea substantivelor – se disting trei cazuri în funcție de elementele de control de tip virgulă și conjuncție ”*și/sau*” care îl însoțesc.
   a. dacă nu există nici virgulă, nici conjuncție – aceasta înseamnă că respectivul substantiv nu poate fi decât un atribut al unui alt grup substantival, iar adăugarea se face în grupul substantival părinte (determinat de instrumentul de etichetare).
   b. dacă există și virgulă și conjuncție – substantivul va fi adăugat unei enumerații și se activează un indicator care precizează faptul că asocierile cu grupuri substantivale ”îndepărtate” trebuie să primească un bonus.
   c. dacă există doar virgulă sau doar conjuncție – substantivul este adăugat unei enumerații, fără condiții sau opțiuni suplimentare.

Selecția structurii candidat optime

În finalul procesului de adăugare, după ce toți termenii au fost analizați, se obține o listă de structuri candidat. Acestea corespund variantelor posibile de aranjare a termenilor din secvența inițială, ținând cont de constrângerile prezente.

Pentru determinarea variantei optime se face o sortare în funcție de scorurile prezentate anterior. În ordinea importanței acestea sunt:

- scor de structură – acest scor trebuie să fie pe prima poziție dacă ne dorim obținerea unei enumerații bine formate
- scor de discrepanță – din testele efectuate, s-a observat faptul că asocierea incorectă a termenilor este o cauză foarte importantă pentru obținerea unor structuri suboptimale
- scor de distanță – ajută suplimentar la ameliorarea neajunsului de asociere incorectă
- scor de similaritate – ajută într-o anumită măsură la diferențierea între variante.

Calcularea scorului de similaritate pentru enumerații

Scorul de similaritate al unei enumerații cuantifică potrivirea între grupurile substantivale din care este alcătuită. Pentru a obține această valoare, sunt calculate mai întâi scorurile de similaritate dintre oricare două grupuri substantivale, după care se face media acestor valori pentru a obține similaritatea enumerației.

Pentru a calcula practic similaritatea enumerației este necesar să avem o listă cu toate grupurile substantivale în ordinea în care apar ele în enumerație. Dacă enumerația conține o enumerație inclusă se va apela o funcție specială care creează un substantiv de sumarizare a acelei enumerații.

În acest context, similaritatea dintre două grupuri substantivale NP1 și NP2 este definită ca fiind media aritmetică a similarităților dintre oricare două grupuri substantivale consecutive dintre NP1 și NP2.

Exemplul Ex. 5.2 prezintă două enumerații. În cea de la punctul (a) observăm că NP2 face parte discordantă ceea ce se reflectă destul de amplu în scorurile de adiacență și în scorul final. Pentru a avea un etalon este furnizată enumerația de la (b) în care grupurile substantivale sunt identice din punctul de vedere al caracteristicilor analizate.

Vom calcula doar similaritatea pentru punctul (a):

$$Similarity_{NP1-NP3} = \frac{76.67 + 80}{2} = 78.34$$

Ex. 5.2

a. cartea, creion și caietele

NP1	NP2	NP3
76.67	80	

b. cartea, creionul și caietele
NP1 NP2 NP3
100 100

Crearea substantivelor de sumarizare

Un substantiv de sumarizare exprimă într-un mod concis caracteristicile unei enumerații. El dobândește caracteristicile dominante ale grupurilor substantivale componente.

Determinarea fiecărei caracteristici se face prin analizarea tuturor caracteristicilor de acel tip din fiecare grup substantival, alegerea finală fiind valoarea cu număr de apariții majoritar. În cazul în care nu se obține această majoritate, se va alege valoarea corespunzătoare primului element din enumerație. Acesta este considerat cel mai important element deoarece imprimă enumerației o anumită amprentă distinctă.

5.16 Etapa de simplificare propriu-zisă a grupurilor substantivale

În cadrul acestei etape este pusă în practică analiza făcută în faza de detecție a enumerațiilor, prin care s-a obținut varianta optimă de asociere a termenilor din secvență. Ea descrie modalitatea de simplificare finală a grupurilor substantivale și a frazei. Și această etapă este bazată pe un proces recursiv care parcurge structura în profunzime până ajunge la elementele de la bază, pe care începe să le simplifice progresiv de jos în sus.

Tratarea elementelor de tip enumerație de substantive

Fiecărei enumerații simplificate trebuie să i se atribuie un cuvânt de legătură simbolic pe baza cuvintelor de legătură ce însoțesc grupurile substantivale aflate în componența sa. Pentru aceasta se utilizează un mecanism prin care se determină tipul de cuvânt de legătură cu influență maximă.

Tabelul 5-3 prezintă valorile coeficienților asociați fiecărui tip de cuvânt de legătură.

Tabelul 5-3 Coeficienții tipurilor de cuvinte de legătură

Cuvânt de legătură	Coeficient
STRONG	1
MEDIUM	-0.6
WEAK	-0.4
STOP	-1000

Deoarece corespund cazurilor oblice de genitiv/dativ, cuvintele de legătură de tip MEDIUM nu pot apare împreună cu alte tipuri cărora le corespund cazurile directe, nominativ/acuzativ. În practică, singurele tipuri de cuvinte de legătură care pot fi combinate într-o enumerație sunt STRONG și WEAK. Datorită tiparelor sale de utilizare, tipul STOP nu poate apare într-o enumerație alături de alte tipuri.

Valorile din tabel au fost alese după următoarele principii:

$$|1S| > |2W|$$

$$|1S| > |1M|$$

Obs.
1. În relațiile de mai sus S, M, W sunt inițialele tipurilor STRONG, MEDIUM și WEAK.
2. Tipul STOP care trebuie să determine tranziția imediată nu intră în calcul având după cum se vede în tabel o valoare considerabilă pentru a forța acest lucru.

Dându-se o serie de coeficienți c_i cu $i = \overline{1,n}$ formula de calcul a scorului corespunzător acestei serii este:

$$Score = \sum_{i=1,n} c_i \left(1 - \frac{i-1}{n}\right)$$

Prin această formulă se urmărește acordarea unei importanțe crescute cuvintelor de legătură de la începutul seriei, primul având chiar coeficientul nealterat, importanța scăzând progresiv pe măsură ce ne apropiem de final.

Tratarea elementelor de tip grup substantival

Prelucrarea grupurilor substantivale are un grad crescut de complexitate deoarece este constituită din trei faze:

1. simplificarea adjectivelor provenite din participiu
2. simplificarea enumerațiilor de adjective din participiu
3. simplificarea grupului substantival

Ordinea acestor operații nu poate fi alterată deoarece rezultatul obținut într-o anumită fază necesită utilizarea în fazele ulterioare.

1. Pentru obținerea *componentelorTp* simplificatoare ale adjectivelor din participiu sunt căutate în vectorul de elemente secvențe de cuvinte având următoarele părți de vorbire:

- ADJECTIV(PARTICIPIU) SUBSTANTIV
- ADJECTIV(PARTICIPIU) ADPOZIȚIE SUBSTANTIV

2. Găsirea enumerațiilor de adjective din participiu presupune parcurgerea vectorul de elemente în căutarea unui prim asemenea adjectiv. La găsirea sa, se lansează o altă căutare care urmărește să extragă o eventuală listă de asemenea adjective ce începe din acea poziție. Căutarea se va realiza atât timp cât sunt întâlnite adjective la participiu, virgule sau conjuncții ”*și/sau*”. De remarcat că în această fază a execuției putem întâlni doar adjective din participiu sau adjectivele simplificatoare ale secvențelor prelucrate în faza anterioară.

3. În final este posibilă realizarea ultimei faze care presupune crearea *componenteiTp* simplificatoare pentru grupul substantival. Tot acum este creată eticheta componentei de tipar simplu, un șir de caractere care să descrie succint caracteristicile principale:

- forma articolului – este considerat tipul elementului de tipar
  - NPDEF (articol hotărât)
  - NPUNDEF (articol nehotărât)
  - NP (fără articol)
- cazul substantivului – se adaugă la sfârșit sufixul ”OBL” (ex. NPUNDEF_OBL) pentru cazurile oblice de genitiv/dativ. Formele directe sunt implicite, deci componenta de tipar nu suferă nici o modificare.

5.17 Stocarea tiparelor simple

Acest proces implică crearea unor liste de tipare grupate pe mai multe categorii care sunt corespunzătoarea unităților formale ale definițiilor: RGET, VERB, GENUS, REST.

Pentru unitățile TARGET și HYPER nu se aplică acest procedeu din două motive:

- sunt întotdeauna alcătuite dintr-un substantiv simplu sau un grup substantival, fapt pentru care analiza lor nu ar aduce absolut nici un plus de rezoluție, deoarece în final am obține doar câte un singur tipar pentru fiecare unitate

- ar împărți unitățile RGET și GENUS în câte două părți ceea ce ar determina creșterea complexității de prelucrare și analiză

În schimb, pozițiile lor în cadrul unităților formale părinte sunt extrem de importante și sunt marcate, astfel încât să poată fi utilizate la momentul potrivit.

În cadrul fiecărei categorii de unități formale tiparele sunt grupate în mod adițional după forma tiparului simplu pentru a reduce numărul de verificări necesare la testare.

Listingul 5-1 prezintă un exemplu de tipar simplu pentru unitatea formală GENUS și câteva din secvențele de definiții luate din corpusul de antrenare care îi corespund.

NPUNDEF_hyp	care_gen	VERB_gen	ADPOZ_gen	NPDEF_gen
o componentă software	care	rulează	în	contextul alt unui program
un program de calculator	care	ajută	la	pregătirea unui microprogram
o specie de broască de copac	ce	trăiește	în	estul Australiei
un tip de speleothem	care	atârnă	de~pe	tavanul sau zidul peșterilor de calcar

Listingul 5-1 Tipar simplu

Pe prima linie sunt componentele tiparului simplu la care s-a adăugat, pentru a facilita înțelegerea lor, un postfix ce conține prescurtarea unității formale a definiției din care fac parte.

Pentru simplitate, în cele ce urmează, ne vom referi la aceste componente de tipar simplu prin notația generică *componentăT*.

Împărțirea în subtipare se face grupând secvențele de *componenteT* în funcție de unitățile din care fac parte. În cazul în care anumite *componenteT* se extind în două unități consecutive, ele vor fi introduse în prima unitate.

Obs. Apartenența la două unități formale survine datorită faptului că spre deosebire de procesul de simplificare a frazei care urmează o serie de reguli stricte, acțiunea de adnotare a unităților formale din corpus are o natură mult mai organică ce ține cont și de sensul cuvintelor.

În acest context, două tipare simple sunt considerate identice dacă au același număr de componente și toate *componenteleT* corespondente se potrivesc. De remarcat faptul că potrivirea între componente nu implică faptul ca acestea să fie perfect identice.

În momentul în care tiparul simplu există deja în baza de tipare se încearcă doar adăugarea noului text la cele existente. Pentru aceasta se face o altă verificare extrem de strictă, de această dată la nivel de termen, pentru a preîntâmpinarea adăugarea unui text identic cu unul deja existent.

Componenta de tipar simplu

Această componentă determină, în funcție de tipul ei, modul în care este privit termenul corespondent din fraza analizată. Vom prezenta în continuare aceste tipuri însoțite de o serie de comentarii explicative:

- clasa substantivelor directe (caz nominativ/acuzativ) (NOUN, NP, NPDEF, NPUNDEF, NP_ENUM, NPDEF_ENUM, NPUNDEF_ENUM)
- clasa substantivelor oblice (caz genitiv/dativ) – la tipurile ce fac parte din clasa substantivelor directe se adaugă postfixul ”_OBL” (ex. NPDEF_OBL)
- clasa adjectivelor din participiu (PARTICIPLE, PERTICIPLE_ENUM)
- clasa informațiilor morfologice (WORD, LEMMA, POS_TYPE)
- LINKWORD_TYPE – tip special pentru marcarea cuvintelor de legătură. În plus, față de elementele simple, conține tipul cuvântului de legătură.

- CUSTOM – tip special pentru tratarea cazurilor în care elementul de tipar este impus prin mecanismul de adnotații

Obs.
1. NOUN și NOUN_OBL sunt elemente de tipar simplu intermediare. Ele nu apar în forma finală a tiparului.
2. În cazul claselor de substantive forma pentru enumerație este considerată identică cu forma simplă a aceluiași tip de articol (ex. NPDEF == NPDEF_ENUM). Vom considera aceste tipuri înrudite ca făcând parte din același tip extins. Acest aspect este valabil și în cazul adjectivelor provenite din participiu.

5.18 Concluzii

Lista de contribuții din capitolul 5:

- crearea unui mod de reprezentare internă a tiparelor corespondente frazelor analizate și implementarea operațiilor necesare de funcționare
- crearea unui motor de interpretare și comparare între tiparele de simplificare și tiparele frazelor
- implementarea algoritmilor care valorifică regulile de simplificare definite în fișierele de configurare specifice
- dezvoltarea unei modalități de identificare în text și de simplificare a enumerațiilor

Acest capitol descrie algoritmul prin care sunt analizate frazele în vederea detecției definițiilor. Rezultatul analizei frazelor este obținerea unui tipar cât mai simplificat, însă suficient de detaliat pentru a putea face distincția dintre frazele obișnuite și definiții.

Procesul de analiză a unei fraze trece prin mai multe etape, și anume: preprocesare, etichetare, simplificare și în final crearea tiparului simplu.

Prima dintre etapele de mai sus, presupune preprocesarea textului atât la nivel morfologic cât și sintactic pentru a ajuta instrumentul de etichetare să ia deciziile cele mai corecte în anumite cazuri dificile. Pentru aceasta sunt operate anumite simplificări temporare înainte de etichetare.

Etapa de simplificare prezintă o complexitate foarte ridicată și a presupus elaborarea, pornind de la zero, a algoritmilor și rutinelor necesare procesării frazei.

În timpul analizei, fraza este stocată într-o latice de componente de tipar, care este capabilă să rețină toate fazele succesive de simplificare. Ea permite aplicarea unor operații de adăugare, eliminare, repoziționare, simplificare a nodurilor.

Pentru testarea cu ajutorul tiparelor de simplificare, definite în cele trei fișiere de configurație, a fost creat un motor specific de interpretare și comparare care verifică succesiv posibilitatea aplicării fiecărui tipar. În cazul unei operații de potrivire încheiate cu succes, asupra termenilor corespondenți din frază sunt aplicate toate operațiile de adnotare și simplificare stabilite în cadrul tiparului.

Simplificarea frazei conține mai multe sub-etape cum ar fi: eliminarea anumitor expresii neimportante în contextul analizei definițiilor, marcarea cuvintelor de legătură, marcarea anumitor adjective care au formă de verb la participiu, operații de simplificare intermediara, etc.

Cea mai importantă sub-etapă este simplificarea finală a grupurilor substantivale care poate reduce structuri complexe de termeni la un singur substantiv. În paralel cu această operațiune se realizează identificarea enumerațiilor din cadrul grupului substantival. În acest scop a fost elaborat un set de reguli bazate pe constante numerice și comutatoare booleene prin care sunt controlate mecanismele de asociere a termenilor. Alegerea variantei optime de asociere se

realizează print utilizarea unor scoruri de fiabilitate care analizează (a) similitudinea dintre substantive, (b) elementele de structură ale enumerațiilor, (c) gradul de discrepanță și (d) distanța dintre două substantive asociate.

Ultima etapă, de creare a tiparelor, presupune împărțirea tiparului simplu obținut după simplificare în funcție de unitățile formale ale definiției, creându-se câte un sub-tipar pentru RGET, VERB, GENUS, REST. Aceste sub-tipare vor fi adăugate fiecare la câte o listă specifică.

6 Detecția definițiilor – Validare experimentală

Procesul de detecție a definițiilor constă din următoarele etape:

1. procesarea frazei care urmează să fie testată pentru a fi creat tiparul simplu asociat
2. testarea cu sub-tiparele create pentru unitățile formale ale definiției

În etapa întâi, procedura de creare a tiparelor simple este comună atât fazei de antrenare cât și celei de testare și are loc aceeași operație de procesare care a fost aplicată și definițiilor din corpusul de antrenare.

În etapa a doua se urmărește găsirea unui set de sub-tipare care să acopere cât mai extins textul frazei. Vorbim de set de sub-tipare deoarece acestea nu este obligatoriu să provină de la o aceeași definiție din corpusul de antrenare.

Prin utilizarea unor sub-tipare pentru fiecare dintre unitățile formale ale definiției sunt atinse două deziderate:

- se îmbunătățește capacitatea de detecție a definițiilor prin mărirea gradului de acoperire a configurațiilor de termeni ce pot apare într-o definiție
- devine posibilă identificarea unităților formale ale definiției, fapt ce facilitează marcarea în cadrul definiției a conceptului și a supra-conceptului

6.1 Studiu de caz

Pentru testarea clasificatorului a fost utilizată o colecție de texte extrase din Enciclopedia Științei [62], carte care conține informații din diverse domenii ale lumii științifice. Această colecție de texte este compusă din 1128 de fraze (din care 112 sunt definiții), informații ce acoperă un număr de 17 subiecte, fiecare grupând în jur de 50-60 de fraze.

Pentru prelucrarea textelor pe mai multe nivele fiecare dintre cele 17 subiecte a fost considerat un document astfel încât să fie posibilă atât o analiză separată a fiecăruia, cât și analiza generală a tuturor.

Acest studiu a fost realizat având în vedere două obiective:

1. în primul rând testarea performanței clasificatorului
2. utilizarea rezultatelor obținute la testare pentru studia o posibilă corelație între termenii ce constituie conceptele și supra-conceptele din definiții, pe de o parte, și frecvența lor relativă (sau TF-IDF).

Mai precis s-a urmărit confirmarea ipotezei că dacă se iau toate conceptele dintr-un document și se ordonează în funcție de valorile frecvenței (sau după TF-IDF), cele mai importante concepte și supra-concepte din definiții se vor găsi pe primele poziții.

Pentru concizie, în continuare, vom utiliza sintagma ***concept definițional*** pentru a face referire în mod colectiv la termenii care în definiție constituie conceptul definit și supra-conceptul acestuia.

Deoarece conceptele definiționale pot fi obținute cu ușurință folosind marcajele furnizate de clasificator, vom prezenta modul de identificare a restului de concepte din document.

Acestea sunt grupe de maxim trei cuvinte extrase din frază, folosind un anumit set de reguli care urmărește să obțină doar conceptele valide, spre deosebire de extragerea unor grupe oarecare de cuvinte consecutive.

În acest context, prin concepte valide ne referim la entități exprimate prin intermediul substantivelor care pot fi însoțite de eventuale atribute.

Regulile care guvernează crearea acestor grupuri de cuvinte sunt următoarele:

- este obligatoriu să existe cel puțin un substantiv în secvență
- orice alt cuvânt prezent în secvență altul decât substantivul principal trebuie să aparțină logic de acesta sau să conecteze un cuvânt care aparține logic de acesta

6.2 Rezultatele clasificării

În urma analizării frazelor din corpusul de test sunt obținute rezultatele prezentate în matricea de confuzie din tabelul 6-1.

Tabelul 6-1 Matricea de confuzie

		Real	
		Definiție	Non-definiție
Clasificare	Definiție	72	40/16
	Non-definiție	27	987/1010

Din tabel se observă că doar 72 din totalul de 112 definiții au fost corect identificate de către clasificator. Pe lângă acestea sunt clasificate drept definiții un număr de 27 de fraze obișnuite.

Testul 1	Testul 2
TP (true positives) = 72	
FN (false negatives) = 40	**FN** (false negatives) = 16
FP (false positives) = 27	
TN (true negatives) = 987	**TN** (true negatives) = 1010

Măsurile care descriu performanța clasificatorului sunt prezentate în cele ce urmează:

a. **Precizia** (rata cazurilor adevărat pozitive) – măsoară proporția de definiții corect clasificate din totalul frazelor care au fost clasificate drept definiții

$$Precision = \frac{TP}{TP + FP}$$

b. **Reamintirea** (rata cazurilor adevărat negative) – măsoară proporția de definiții corect clasificate din totalul definițiilor

$$Recall = \frac{TP}{TP + FN}$$

c. **Acuratețea** – măsoară proporția predicțiilor corecte ale clasificatorului

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

d. **Măsura-F** – o altă măsură a acurateței de testare care furnizează un singur scor pe baza valorilor obținute pentru precizie și reamintire, reprezentând media armonică a celor două

$$F1\text{-}score = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2TP}{2TP + FP + FN}$$

Utilizând formulele de mai sus, vom obține următoarele valori pentru cele patru măsuri de performanță:

Testul 1	Testul 2
$Precision = \frac{72}{72+27} = 0.73$	$Precision = \frac{72}{72+27} = 0.73$

$$Recall = \frac{72}{72+40} = 0.64$$
$$Accuracy = \frac{72+987}{72+987+27+40} = 0.94$$
$$F1\text{-}score = \frac{2*72}{2*72+27+40} = 0.69$$

$$Recall = \frac{72}{72+16} = 0.82$$
$$Accuracy = \frac{72+1010}{72+1010+27+16} = 0.96$$
$$F1\text{-}score = \frac{2*72}{2*72+27+16} = 0.77$$

6.3 Analiza erorilor de clasificare

Vom urmări să analizăm cazurile notabile pentru fiecare gamă de erori, pentru a determina cauzele producerii lor. Fiecare tip de eroare va fi însoțit de o serie de exemple reprezentative.

Cazurile fals pozitive

❖ Din punct de vedere semantic anumite fraze se situează la limita la care aproape pot fi considerate definiții. Pentru unele dintre ele am putea chiar argumenta că ar fi putut fi marcate drept definiții în faza de adnotare manuală.

Ex. 6.1
a. Forța râului este un curent mai puternic .
b. Majoritatea antenelor sunt tije , baghete sau discuri metalice .
c. Lumea naturală este un loc amestecat în care se combină tot felul de materiale .

❖ O serie de fraze, deși sintactic au forma unei definiții, nu ar putea fi identificate corect decât folosind informații semantice. Această dificultate este introdusă, în general, prin folosirea verbul definițional ”***a fi***”, un verb larg utilizat în contexte semantice din cele mai diverse.

Ex. 6.2
a. Punctul de sprijin este piesa din mijloc , pe care stă balansoarul .
b. Iarna este momentul ideal pentru a identifica steaua Sirius .

❖ Anumite potențiale definiții sunt incomplete, deoarece le lipsește parțial sau total conceptul definit (*definiendum*). Ele definesc un termen aflat într-una din frazele anterioare, la care se face referire de exemplu prin intermediul unui pronume.

Ex. 6.3
a. Această rată este frecvența lor .
b. Această pădure vastă este numită pădure boreală sau taiga .

Cazurile fals negative

❖ Un număr foarte mare de cazuri, poate chiar majoritar, apar datorită faptului că anumiți termeni sunt identificați în mod eronat de către instrumentul de etichetare și li se atribuie o parte de vorbire greșită.

Ex. 6.4
a. ***Pana*** este o mașină simplă folosită pentru tăiat .
   – corect: **substantiv** - pană — incorect: **prepoziție** - până
b. Undele radio sunt ***unde*** lungi , invizibile , cu energie joasă .
   – corect: **substantiv** - unde — incorect: **adverb** - unde
c. Un aerogel este un ***solid*** cu un gaz dizolvat în el .
   – corect: **substantiv** - solid — incorect: **adjectiv** - solid

❖ O altă categorie de erori apar datorită faptului că tiparul definiției nu există, pentru că nu au existat definiții cu format similar în corpusul de antrenare. Aceasta se datorează faptului că

multe definiții din corpusul de testare se distanțează chiar foarte mult de la tiparul clasic aristotelian ce caracterizează setul de antrenare.

Ex. 6.5

a. Prima este *lungimea de undă - distanța dintre un vârf al undei și altul .*
b. *ANTIBIOTIC : medicament care ucide bacteriile* sau le împiedică să se reproducă .
c. *Locurile unde masele de aer diferite se întâlnesc se numesc fronturi* .

6.4 Studierea corelației dintre concepte definiționale și frecvență (TF-IDF)

După cum am menționat anterior, ipoteza de lucru legată de conceptele definiționale importante este aceea că au în documentul din care provin o frecvență relativă mai mare decât majoritatea celorlalte concepte.

De asemenea suntem interesați în definirea unei metode de identificare automată a acestor concepte definiționale importante.

Pentru aceasta vom studia rezultatele obținute folosind frecvență relativă (*TF*) comparativ cu măsura TF-IDF a termenilor respectivi, pentru a determina dacă una dintre ele are o putere mai mare de discriminare.

Definiția măsurii TF-IDF este următoarea:

> Dacă se dă o colecție de documente *D*, TF-IDF este o măsură statistică ce permite evaluarea importanței unui termen pentru un anume document *d* din colecția *D*.

Pentru a determina în mod practic procentul conceptelor definiționale importante din document propunem doi algoritmi ale căror rezultate vor fi analizate comparativ pentru a determina care este cel mai potrivit scopului urmărit.

Trebuie să precizăm că multe etape sunt comune ambilor algoritmi, diferența înregistrându-se în etapele în care se realizează efectiv selectarea conceptelor definiționale importante. Din acest motiv vom prezenta complet doar primul algoritm, precizând doar diferențele pentru cel de-al doilea.

❖ *Algoritmul 1*

1. înregistrările din tabel vor fi sortate după *docID*, *TF* (*TF-IDF*) și în ultimul rând, după *expresie*
2. pentru fiecare document
   2.1. se vor contoriza înregistrările de tip TP, ce corespund conceptelor din definițiile reale. Vom face referire la această valoare prin notația nr_{DC}
   2.2. se vor contoriza câte din primele nr_{DC} înregistrări din listă sunt de fapt de tip TP. Această valoare, ce reprezintă numărul de concepte definiționale, va fi notată prin nr_{ImpDC}
   2.3. se va calcula raportul dintre nr_{ImpDC} și nr_{DC}, care ne dă procentul conceptelor definiționale
3. la final, pentru a calcula procentajul global al conceptelor definiționale considerate suficient de importante, vom calcula media aritmetică a valorilor parțiale obținute pentru fiecare document

Obs. Dacă un document nu conține definiții, ceea ce înseamnă că nr_{DC} este 0, rezultatul pasului 2 va fi automat 0.

❖ *Algoritmul* 2

Pasul 1 este ușor modificat deoarece pentru al doilea criteriu de sortare se va folosi doar TF.

Pasul 2 care se ocupă cu selecția conceptelor importante este diferit, după cum se poate vedea în continuare:

2. pentru fiecare document
   2.1. se vor contoriza înregistrările de tip TP notate prin nr_{DC}
   2.2. se va calcula valoarea medie pentru frecvențele relative ale tuturor conceptelor, care va fi notată cu nr_{avg}
   2.3. se vor contoriza câte dintre conceptele definiționale au frecvența relativă mai mare decât nr_{avg}, valoare ce va fi desemnată drept nr_{ImpDC}
   2.4. se va calcula procentul conceptelor definiționale importante nr_{ImpDC} / nr_{DC}

Pentru algoritmul 1, raționamentul din spatele alegerii numărului total de concepte definiționale nr_{DC} ca factor de limitare pentru fiecare document se bazează pe două considerente:

1. numărul nu poate fi o constantă, ci este necesar să fie adaptabil în funcție de numărul de definiții existente în document
2. dacă toate conceptele definiționale ar fi foarte importante, ele s-ar plasa în totalitate pe primele poziții în tabel, fapt ce ne face să concluzionăm că nr_{DC} este un interval rezonabil pentru a le alege pe cele importante.

Valorile procentajelor obținute pentru cele 17 documente (raporturile nr_{ImpDC} / nr_{DC}), precum și scorul final (media aritmetică) sunt afișate în tabelul 6-2.

Tabelul 6-2 Procentul conceptelor definiționale în documentele corpusului de testare

doc	TF(alg1)	TF-IDF	TF(alg2)	doc	TF(alg1)	TF-IDF	TF(alg2)
1	0.4	0.4	0.45	10	0.1	0.2	0.6
2	0.57	0.57	1	11	0.14	0.29	0.57
3	0.33	0.33	1	12	0.22	0.33	0.56
4	0.36	0.36	0.45	13	0.22	0.22	0.33
5	0.33	0.33	0.67	14	0.14	0.29	0.71
6	0.22	0.22	0.61	15	0.18	0.18	0.47
7	0.36	0.45	0.45	16	0	0	0.43
8	0.25	0.32	0.39	17	0.2	0.2	0.6
9	0.6	0.6	1	**scor**	**0.27**	**0.31**	**0.6**

În cazul *algoritmului 1* considerăm că varianta care folosește TF-IDF este mai bună deoarece este capabilă să plaseze mai multe concepte definiționale în intervalul desemnat celor importante. Rezultatele globale finale sunt obținute făcând media aritmetică și arată că în general 27% (respectiv 31%) din conceptele definiționale pot fi considerate importante. Deși *algoritmul 2* este capabil să selecteze mai multe concepte definiționale, el poate fi perceput ca fiind prea permisiv deoarece acceptă concepte cu frecvență relativă destul de mică comparativ cu cele de la vârful ierarhiei, multe dintre ele nefiind concepte definiționale.

Pe aceste considerente, am ales utilizarea primului algoritm, ale cărui rezultate le vom evalua în cele ce urmează.

6.5 Validarea rezultatelor

Deoarece nu avem o modalitate automată de a determina importanța conceptelor care este în

esență o operație de analiză semantică, vom utiliza o metodă de evaluare umană intuitivă pentru a determina cât de reprezentative sunt conceptele definiționale selectate, pentru documentul de proveniență.

Pentru a oferi un exemplu, vom considera conceptele definiționale importante extrase din documentele #1 și #3, care sunt prezentate în tabelul 6-3.

Tabelul 6-3 Concepte definiționale importante

docID	concept	nr	procent
1	forță	29	8.71%
	mașină	19	5.71%
	mașină simplu	8	2.40%
	rampă	7	2.10%
	forță activ	6	1.80%
	pană	6	1.80%
	șurub	5	1.50%
	tip	4	1.20%
3	undă	29	13.94%
	undă sonor	4	1.92%

Parcurgând lista ne putem da imediat seama că în documentul #1 se vorbește despre o serie de dispozitive simple (”*mașină simplă*”, ”*rampă*”, ”*pană*”, ”*șurub*”), precum și modul în care ele interacționează cu mediul (”*forță*”, ”*forță activă*”).

Pentru documentul #3 observăm că procentul de apariții ale conceptului ”*undă*” raportat la numărul total de apariții ale conceptelor este foarte mare (aproape 14%) ceea ce ne ajută să intuim cu mare grad de încredere tema generală a acestui document.

Un caz interesant este situația documentului #16 care aparent nu are concepte definiționale importante, după cum se poate observa și în datele prezentate în tabelul 6-2. De fapt și acest document conține concepte definiționale importante cum ar fi ”*stea*” (37 apariții - 10%), ”*soare*” (6 apariții - 1.62%), ”*stea neutronică*” (5 apariții – 1.35%), ”*fuziune*” (4 apariții – 1.08%) care însă apar în definiții cu format atipic, fapt pentru care nu au fost recunoscute ca atare de către clasificatorul de definiții.

Plecând de la premisa că pot exista definiții atipice nedetectate, pe baza conceptelor cu TF-IDF mare, se pot face sugestii de posibile contexte definiționale ce includ aceste concepte, în scopul validării supervizate ulterioare.

În concluzie, considerăm că aceste concepte definiționale au o capacitate suficient de ridicată de a descrie subiectele documentelor de proveniență ceea ce le conferă, în opinia noastră, un grad de importanță suficient, validând astfel metoda de selectare prezentată în cadrul *algoritmului 1*.

6.6 Concluzii

Lista de contribuții din capitolul 6:

- implementarea algoritmului de testare a definițiilor
- studierea legăturii între conceptele definiționale și frecvență (respectiv TF-IDF) în document
- elaborarea unei metode de a determina în mod automat conceptele definiționale importante

În acest capitol este prezentat modul în care se face testarea unei fraze în scopul detecției de definiții. O identificare reușită presupune găsirea unei combinații valide de sub-tipare RGET, VERB, GENUS și parțial REST care să se potrivească cu termenii respectivei fraze. O caracteristică importantă a acestei operații este faptul că sub-tiparele pot proveni de la definiții diferite din corpusul de antrenare.

În ultima parte a capitolului este realizat un studiu de caz pentru testarea clasificatorului, folosindu-se o colecție de 1128 de fraze, acoperind un total de 17 subiecte de discuție (documente) și conținând 112 definiții. În paralel se urmărește să se determine dacă supra-conceptele și conceptele importante din definiții au o frecvență relativă și/sau măsură TF-IDF mai mare decât majoritatea celorlalte concepte din documentele în care se găsesc.

În urma testării clasificatorului pentru toate definițiile s-a obținut o precizie de 0.73, reamintire de 0.64, acurateţe de 0.94 și o măsură F1 de 0.69, iar pentru setul de definiții cu format clasic, o precizie de 0.73, reamintire de 0.84 acurateţe de 0.96 și o măsură F1 de 0.77.

De asemenea, sunt analizate cauzele erorilor de clasificare pentru cazurile fals pozitive și fals negative, discutând pe baza unor exemple concrete.

Pentru validarea ipotezei de lucru care plasează conceptele definiționale importante în fruntea ierarhiei tuturor conceptelor dintr-un anume document, au fost folosite ca și criterii de ordonare alternative, atât frecvența relativă cât și măsura TF-IDF. Determinarea conceptelor importante se face pe baza observației că dacă toate conceptele definiționale ar fi foarte importante, ele s-ar plasa pe primele N poziții. Pentru a obține o valoare reprezentativă pentru fiecare document, N este ales ca fiind numărul total de concepte definiționale extrase de clasificator. În continuare, calculând numărul real al celor care se află pe primele N poziții, putem determina procentul de concepte definiționale importante. În final, făcând media aritmetică a tuturor valorilor, se înregistrează ca procentaje generale ale conceptelor definiționale selectate, 27% pentru frecvența relativă și 31% pentru TF-IDF.

7 Concluzii generale, contribuții originale și perspective

Prin cercetarea realizată în această lucrare am urmărit să dezvoltăm pentru limba română o metodă de extracție a definițiilor din text, în special din textele structurate și semi-structurate. Fiind concepută special pentru o compatibilitate maximă cu limba română, metoda ține cont de particularitățile acesteia pentru optimizarea performanțelor.

În vederea dobândirii unei imagini de ansamblu în domeniul detecției și extragerii definițiilor am analizat studiile anterioare în acest domeniu pentru a descoperi avantajele și dezavantajele fiecărei metode, astfel încât să putem decide în cunoștință de cauză care dintre acestea ar fi cele mai potrivite și dacă pot fi adaptate pentru scopul propus.

În această direcție, capitolul 1 face o trecere în revistă a metodelor de identificare/extragere a definițiilor care pot fi plasate în două categorii principale.

În prima categorie avem metodele care au la bază identificarea definițiilor folosind **tipare lexico-sintactice**. Aceste tipare pot fi obținute prin crearea iterativă manuală în urma analizei directe a textului sau într-un mod automatizat ce presupune folosirea unui corpus de antrenare adnotat în mod specific sau, alternativ, tehnica de *bootstrapping*.

În a doua categorie se clasează metodele care implementează **tehnici de machine learning** folosind pentru antrenare diferite caracteristici create în general pe baza proprietăților lexico-sintactice ale frazelor definiționale.

În cadrul capitolului sunt prezentate pe scurt tehnicile de preprocesare a colecțiilor de texte folosite pentru antrenare și testare (secțiunea 1.3) și metodele importante de abordare a analizei definițiilor (secțiunea 1.4).

Contribuțiile tezei sunt prezentate în capitolele 2-6, și descriu procesul prin care o metodă de extragere a definițiilor pentru limba engleză, bazată pe tipare lexico-sintactice, a fost preluată și modificată pentru limba română, împreună cu corpusul său de antrenare format în exclusivitate din definiții. În final motorul principal al acestei tehnici a fost complet schimbat deoarece s-a dovedit ineficient în cazul limbii române. Corpusul de antrenare a fost tradus în română păstrându-se esența adnotării manuale inițiale, dar adaptată după cum a fost necesar.

Clasificatorul descris în această lucrare a fost scris în limbajul de programare Java și implementează toate ideile descrise în cadrul acestor capitole.

În capitolul 2 sunt prezentate modificările aduse metodei originale de extragere a definițiilor pentru a o adapta la specificul limbii române.

După cum am precizat, mecanismul principal de creare automată a tiparelor a fost aproape complet schimbat. Metoda în limba engleză utilizează așa numitele **tipare stea** create pe baza claselor de cuvinte. Clasa poate fi chiar cuvântul respectiv, dacă acest are o frecvență ridicată (ex. prepozițiile) sau partea sa de vorbire în caz contrar. Prin această operație tiparul capătă un anumit grad de generalizare. Pentru limba română unde folosirea prepozițiilor în cadrul grupurilor substantivale (ex. obiect ascuțit *în* formă *de* pin) este foarte ridicată, tiparele obținute astfel ar avea o capacitate foarte redusă de generalizare. Ca alternativă se realizează simplificarea grupurilor substantivale după anumite criterii, astfel încât unitățile formale ale definiției, în special porțiunea ce constituie explicația, să prezinte raportul considerat optim între importanța informațională a secvenței de cuvinte și numărul acestora.

Acest raport nu este efectiv calculat ci mai curând estimat prin introducerea conceptului de **cuvânt de legătură** care în funcție de caracteristicile proprii tinde să introducă informații cu un anumit grad specific de importanță. Aceste cuvinte de legătură sunt descrise pe larg în secțiunea 2.4 și sunt grupate în patru categorii care reflectă importanța informațională (STRONG, MEDIUM, WEAK, STOP).

Categoriile de cuvinte de legătură sunt guvernate de reguli implicite pe baza cărora s-a realizat re-adnotarea manuală a unităților formale ale definițiilor din corpusul de antrenare. Anumite situații particulare legate de anumite cuvinte de legătură au determinat crearea unor reguli speciale.

Pentru a obține o înțelegere a tiparelor de utilizare a cuvintelor de legătură au fost definite o serie de măsuri ce caracterizează precis fiecare asemenea cuvânt: **gradul de potrivire** (fitness), **mobilitatea** (mobility), **eroarea de predicție** (predError) care sunt calculate pe baza unor date numerice statistice extrase din corpusul de antrenare: *numărul de apariții*, *numărul de tranziții*, *numărul de excepții*. Acest studiu este prezentat în secțiunea 2.5.

În prima parte a capitolului 3 este prezentat un studiu de caz asupra caracteristicilor a cinci instrumente de etichetare a părților de vorbire care sunt disponibile pentru limba română, în scopul alegerii celui mai potrivit.

Secțiunea 3.2 prezintă detaliat instrumentul ales, UAIC POS Tagger, care având o natură hibridă permite ajustarea rezultatului de etichetare prin intermediul unor reguli definite sub formă de grafuri ordonate. Aceste reguli sunt de fapt tipare care în momentul potrivirii cu o anumită secvență de termeni din fraza analizată pot determina ajustarea unei etichete de parte de vorbire. De asemenea, același sistem de reguli este folosit în cazul marcării grupurilor substantivale pentru a se decide limitele acestora.

Contribuția noastră în cadrul acestui proces, constă în îmbunătățirea și adăugarea unor noi reguli de etichetare a părților de vorbire corectând astfel o serie de erori frecvente, precum și îmbunătățirea regulilor de marcare a grupurilor substantivale.

Capitolul 4 descrie un limbaj special conceput pentru a reprezenta tiparele de simplificare, care constituie un mecanism extrem de important în cadrul metodei și care sunt folosite pentru a simplifica în mod secvențial fraza care se dorește a fi analizată. Sintaxa limbajului de reprezentare a fost concepută astfel încât să poată permite definirea facilă a regulilor de simplificare, înlesnind crearea și depanarea lor.

Aceste tipare sunt definite prin intermediul a trei fișiere de configurare fiecare având un specific aparte, iar secțiunea 4.3 prezintă structura acestor fișiere și sintaxa lor generală.

În partea a doua a capitolului, în secțiunea 4.5, sunt descrise elementele de sintaxă legate de componentele individuale de tipar. Această sintaxă este centrată pe componentă, deoarece operațiile specificate prin operatorii sintactici se aplică exclusiv termenului corespondent din fraza analizată.

Există componente de control care influențează structura tiparului și componente structurale care sunt efectiv folosite pentru a verifica potrivirea cu termenii frazei. Fiecare componentă structurală definește tipul de comparare (nivel cuvânt, lemă, parte de vorbire) ce se va efectua și i se pot atașa o serie de modificatori care stabilesc modul de comparare (ex. modificator de repetare, de opționalitate, etc.).

În capitolul 5 sunt descrise etapele de creare a tiparului simplu pentru o frază analizată.

Secțiunea 5.2 prezintă modul de reprezentare internă a acestui tipar. În scopul simplificării nedistructive a frazei a fost dezvoltat un mecanism care are drept suport un graf orientat pe care au fost definite o serie de operații atomice: **simplificarea** unei secvențe de noduri, **eliminarea** și **adăugarea** de noduri, **interschimbarea** nodurilor, **deplasarea** unui nod.

În cele ce urmează ne vom referi la tiparul corespunzător frazei analizate prin denumirea de **tipar în prelucrare** pentru a-l diferenția de tiparele de simplificare.

Pentru a ajuta instrumentul de etichetare să furnizeze etichetele corecte pentru termenii dificili din frază, se realizează o operațiune de preprocesare simplificatoare temporară care este descrisă în secțiunea 5.4.

Începând cu secțiunea 5.5 sunt prezentate etapele care permit obținerea tiparului simplu.

În secțiunea 5.14 este prezentată într-un mod general modalitatea de simplificare a grupurilor substantivale care este formată din două faze:

- etapa de identificare a enumerațiilor – în care se face analiza fiecărui grup substantival
- etapa de simplificare propriu-zisă

Etapa de identificare a enumerațiilor este prezentată în secțiunea 5.15. În această fază, se urmărește identificarea enumerațiilor de substantive, adjective și în particular adjective provenite din verb la participiu (cu rol special în crearea tiparului simplu). Totodată se urmărește reevaluarea relațiilor dintre substantive pe anumite considerente de similitudine. Determinarea celei mai bune configurații, se face pe baza unor scoruri de fiabilitate, după cum urmează: scor de **similitudine**, de **structură**, de **discrepanță** și de **distanță**.

Rezultatul obținut în această fază de analiză a enumerațiilor este utilizat pentru simplificarea efectivă a grupurilor substantivale care este descrisă în secțiunea 5.16. Procesul este unul recursiv, începându-se cu elementele interioare, operând progresiv fie simplificări de enumerații, fie de grupuri substantivale subordonate, după caz.

În final, în secțiunea 5.17, este prezentată modalitatea de stocare a tiparelor simple obținute. Pentru a maximiza gradul de acoperire a tiparelor, acestea sunt împărțite în sub-tipare corespunzătoare unităților formale ale definiției. În acest mod, la testare, se pot genera tipare complete noi, folosind combinații de sub-tipare, indiferent de proveniența acestora.

În prima parte a capitolului 6 este descrisă metoda de testare a frazelor pentru a se determina dacă sunt definiționale. Pentru aceasta este utilizat un corpus de testare compus din 17 documente, fiecare având în jur de 65 de fraze, tipul textului fiind semi-structurat. Totalul frazelor este de 1128 dintre care 112 sunt definiționale.

În secțiunea 6.2 sunt prezentate rezultatele clasificării, evaluându-se performanța prin calcularea indicatorilor obișnuiți: precizie (precision), rata de reamintire (recall), acuratețe (accuracy), scorul-F (F-measure). Tot aici sunt discutate cauzele cele mai frecvente pentru cazurile fals pozitive și fals negative.

A doua parte a capitolului se folosește de definițiile extrase anterior pentru a explora ipoteza potrivit căreia conceptele importante corespunzătoare termenului definit și super-conceptului acestuia (din cadrul definițiilor) au o frecvență crescută în cadrul documentului. În cadrul studiului aceste concepte sunt denumite **concepte definiționale**.

Vom face aici observația că prin concepte importante ne referim la acele concepte care au o mare capacitate descriptivă a textului din care fac parte, putând fi utilizate eventual, într-un proces de sumarizare.

În secțiunea 6.4 sunt prezentați doi algoritmi de selecție automată a conceptelor definiționale importante a căror rezultate sunt analizate comparativ. În secțiunea 6.5 este validată, prin

intermediul unor exemple reprezentative, alegerea algoritmului optim, prezentând rezultatele interesante care au apărut în urma aplicării sale.

7.1 Contribuții

În continuare sunt prezentate succint principalele contribuții din cadrul tezei:

Capitolul 2

- adaptarea modului de adnotare a unităților formale ale definițiilor pentru limba română
- definirea conceptului de cuvânt de legătură
    - stabilirea experimentală progresivă a caracteristicilor cuvintelor de legătură
    - stabilirea celor patru categorii: STRONG, MEDIUM, WEAK, STOP
    - stabilirea regulilor fundamentale de adnotare a unităților formale specifice fiecărei categorii
- elaborarea unui set extins de reguli de adnotare pentru tratarea excepțiilor de utilizare a cuvintelor de legătură
- elaborarea unor măsuri care descriu modul de utilizare a cuvintelor de legătură punând în evidență aceste excepții
- identificarea unei noi metode pentru generarea tiparelor de testare – *tiparele simple* – ce are la bază simplificarea grupurilor substantivale

Capitolul 3

- îmbunătățirea performanței instrumentului de analiză morfo-sintactică folosit
    - îmbunătățirea regulilor de corectare a etichetelor corespunzătoare părților de vorbire
    - îmbunătățirea regulilor de marcare a grupurilor substantivale (noun phrases) în text

Capitolul 4

- stabilirea structurii tiparelor de simplificare a frazelor
- determinarea structurii generale a fișierelor de configurare a tiparelor de simplificare
- crearea unui limbaj de reprezentare a tiparelor de simplificare
    - stabilirea sintaxei generale a fișierelor
    - stabilirea sintaxei componentelor tiparelor de simplificare
    - definirea a două tipuri de componente – de control și structurale
    - elaborarea unui set de modificatori pentru componentele structurale

Capitolul 5

- crearea unui mecanism pentru reprezentarea internă a tiparelor corespondente frazelor analizate
    - implementarea operațiilor atomice necesare prelucrării tiparelor
- dezvoltarea unui motor de interpretare și comparare între tiparele de simplificare și tiparele frazelor
- simplificarea frazelor pentru obținerea tiparelor simple
    - elaborarea algoritmilor specifici fiecărei faze de simplificare folosind tiparele definite în fișierele de configurare
- dezvoltarea unei modalități de identificare și de simplificare a enumerațiilor
    - elaborarea unor scoruri pentru evaluarea structurilor-enumerație candidat

Capitolul 6

- implementarea algoritmului de testare a frazelor în scopul identificării definițiilor
- testarea instrumentului de clasificare dezvoltat
- realizarea unui studiu privind legătura între conceptele definiționale și frecvența lor (respectiv TF-IDF) în document
- elaborarea unui algoritm de selectare automată a conceptelor definiționale importante

7.2 Direcții viitoare

Vom prezenta în cele ce urmează ideile ce se referă la posibile direcții de cercetare în scopul creșterii performanței instrumentului de clasificare dezvoltat în cadrul acestei teze de doctorat.

O metodă relativ ușor de realizat ce poate îmbunătăți performanța clasificatorului prin diminuarea cazurilor fals negative ar fi introducerea în corpusul de antrenarea a unor forme definiționale pentru care nu există tipare. Deși în prezent corpusul de antrenarea conține în exclusivitate definiții care se conformează tiparului clasic aristotelian, este posibilă și analiza unor forme care se abat ușor de la acest tipar, spre exemplu cele de tip glosar în care termenul definit și explicația sunt delimitate printr-un caracter, de obicei ”-” sau ”:”.

O altă posibilitate de îmbunătățire este tratarea cazurilor în care cuvântul definit se află la sfârșitul frazei. Acest tip de definiție ar trebui să se bazeze în primul rând pe verbele care sunt folosite cu precădere în definiții, de tipul ”*a se numi*” pentru a elimina pe cât posibil introducerea unor cazuri fals pozitive prin verbe generale precum ”*a fi*”. Problema principală care apare în acest caz este determinarea corectă a începutului definiției, mai precis a porțiunii din definiție care conține explicația termenului definit. Fiind plasată înaintea verbului definițional nu mai avem luxul de a ști precis unde anume începe ea, deoarece fraza poate să conțină la început o secvență care să nu facă parte din definiție.

Cazurile în care definiția se întinde în două fraze pot fi abordate într-o anumită măsură, însă necesită schimbări majore în modul curent de analiză. De asemenea, dată fiind incertitudinea identificării corecte a părților definițiilor, pentru ele ar trebui să se introducă utilizarea unui grad de încredere.

În acest moment rezultatul clasificării nu poate fi decât *definiție* sau *non-definiție*. Ar fi utilă rafinarea mecanismului de clasificare care să permită evaluarea tuturor definițiilor în termenii unui grad de încredere.

În forma actuală clasificatorul nu poate extrage dintr-o frază definițională decât o singură definiție. Deși adnotarea unităților formale în corpusul de antrenare tratează aceste cazuri, motorul actual de analiză nu integrează detectarea cazurilor în care un anumit termen are mai multe explicații în aceeași frază definițională. Problema în acest caz este să se determine unde se termină o explicație și unde începe următoarea.

Procedura de extragere a definiției poate fi de asemenea rafinată, deoarece în acest moment sunt extrase integral grupurile substantivale ce corespund porțiunii de explicație a termenului definit. În această direcție s-ar putea integra măsurile ce descriu tiparele de utilizare ale cuvintelor de legătură. În acest fel s-ar putea calcula un scor care să ajute la automatizarea procesului decizional de a include sau nu informațiile furnizate prin intermediul acestor cuvinte. Principala utilitate a acestui mecanism ar fi facilitarea unei modalități flexibile de a determina limitele cele mai potrivite pentru unitatea formală ce conține explicația, unde dorim maximizarea cantității de informații printr-o secvență de termeni cât mai scurtă.

Pentru reducerea cazurilor fals pozitive, ar fi foarte utilă implementarea unui modul care să se folosească de relațiile de hiperonimie din RoWordNet (www.racai.ro/tools/text/rowordnet) pentru a determina perechi valide hiponim – hiperonim. În acest caz rolul hiponimului ar fi ocupat de conceptul definit și astfel s-ar putea determina un grad de încredere al definiției verificând dacă supra-conceptul se află sau nu în lista de hiperonime obținute din RoWordNet.

O altă categorie de modificări vizează limbajul de reprezentare a tiparelor de simplificare, căruia i se pot aduce o serie de îmbunătățiri pentru creșterea flexibilității. Una dintre acestea ar fi posibilitatea de a defini secvențe de componente repetabile, deoarece în acest moment poate fi repetabilă doar o singură componentă. De asemenea, componentele cu tip nedefinit nu permit actualmente decât utilizarea în scopul testării adnotațiilor și nu pot avea modificatori deoarece ar deveni mult prea generale. Un posibil remediu pentru această problemă o constituie introducerea modificatorilor cu număr limitat de aplicări și aici avem în primul rând în vedere modificatorul de recurență.

O limitare majoră este lipsa unui mecanism adecvat de asigurare a persistenței datelor. În acest moment la fiecare utilizare este necesară executarea unei faze premergătoare de antrenare. Acest aspect ar permite de asemenea introducerea facilității de editare manuală a tiparelor folosind limbajul definit pentru reprezentarea tiparelor de simplificare, ceea ce ar duce la creșterea pozitivă a generalității acestora. Problema care este necesar să fie rezolvată în acest caz este recunoașterea acestor modificări manuale și integrarea lor în cadrul tiparului, având în vedere faptul că pentru ele nu vor exista informațiile lexico-sintactice adiționale care sunt disponibile pentru componentele de tipar create automat. Aici, ne confruntăm cu două situații adăugarea unor componente suplimentare în tipar și modificarea unora deja existente.

8 Listă lucrări publicate și prezentate

Niculiţă, C. and Dumitriu, L., 2020, October, An Experiment on Text Summarization: Frequent Terms and Concept Definition Extraction (accepted). *In 2020 24th International Conference on System Theory, Control and Computing, Sinaia, Romania*

Niculiţă, C. and Dumitriu, L., 2019, October. The Relational Parts of Speech in Text Analysis for Definition Detection, for Romanian Language. *In 2019 18th RoEduNet Conference: Networking in Education and Research (RoEduNet)* (pp. 135-140). IEEE.

Șușnea, I., Panait, O., **Niculiță, C.** and Munteanu, D., 2018. Machine Vision for Autonomous Vehicles-Potential and Limitations. A Literature Review. *The Annals of "Dunărea de Jos" University of Galati. Fascicle III, Electrotechnics, Electronics, Automatic Control, Informatics*, *41*(2), pp.24-30.

Niculiță, C., 2018. Survey on knowledge representation approaches for natural language processing. *The Annals of "Dunărea de Jos" University of Galati. Fascicle III, Electrotechnics, Electronics, Automatic Control, Informatics*, *41*(1), pp.5-12.

Niculiță C., Istrate A., Vlase M., Jâșcanu N., 2004. The Business Level Structure of WeBLE Platform. Load Tests. *Proceedings of The 12th International Symposium on Modeling, Simulation and Systems' Identification – SIMSIS 12*, Galați, Romania, ISBN 973-627-156-0

Bibliografie

[1] Poon, H. and Domingos, P., 2010, July. Unsupervised ontology induction from text. In *Proceedings of the 48th annual meeting of the Association for Computational Linguistics* (pp. 296-305).

[2] Borg, C., Rosner, M. and Pace, G., 2009, September. Evolutionary algorithms for definition extraction. In *Proceedings of the 1st Workshop on Definition Extraction* (pp. 26-32).

[3] Veyseh, A.P.B., Dernoncourt, F., Dou, D. and Nguyen, T.H., 2020. A Joint Model for Definition Extraction with Syntactic Connection and Semantic Consistency. In *AAAI* (pp. 9098-9105).

[4] Navigli, R. and Velardi, P., 2010, July. Learning word-class lattices for definition and hypernym extraction. In *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 1318-1327).

[5] Storrer, A. and Wellinghoff, S., 2006, May. Automated detection and annotation of term definitions in German text corpora. In *LREC* (pp. 2373-2376).

[6] Malaisé, V., Zweigenbaum, P. and Bachimont, B., 2004. Detecting semantic relations between terms in definitions. In *Proceedings of CompuTerm 2004: 3rd International Workshop on Computational Terminology* (pp. 55-62).

[7] Trimble, L., 1985. *English for science and technology: A discourse approach*. Cambridge University Press.

[8] Flowerdew, J., 1992. Definitions in science lectures. *Applied linguistics*, *13*(2), pp.202-221.

[9] Del Gaudio, R., Batista, G. and Branco, A., 2014. Coping with highly imbalanced datasets: A case study with definition extraction in a multilingual setting. *Natural Language Engineering*, *20*(3), pp.327-359.

[10] Meyer, I., 2001. Extracting knowledge-rich contexts for terminography. *Recent advances in computational terminology*, *2*, p.279.

[11] Auger, A., 1997. *Repérage des énoncés d'intérêt définitoire dans les bases de données textuelles* (Doctoral dissertation, Université de Neuchatel).

[12] Pearson, J., 1998. *Terms in context* (Vol. 1). John Benjamins Publishing.

[13] Alarcón, R., Sierra, G. and Bach, C., 2007. Developing a Definitional Knowledge Extraction System. In *Conference Proceedings of Third Language & Technology Conference LTC'07*.

[14] Reiplinger, M., Schäfer, U. and Wolska, M., 2012, July. Extracting glossary sentences from scholarly articles: A comparative evaluation of pattern bootstrapping and deep analysis. In *Proceedings of the ACL-2012 special workshop on rediscovering 50 years of discoveries* (pp. 55-65).

[15] Durán Muñoz, I., 2010. Specialized lexicographical resources: a survey of translators' needs. *Granger, Sylviane & Magali Paquot (eds.)*, pp.55-66.

[16] Weiten, W., Deguara, D., Rehmke, E. and Sewell, L., 1999. University, community college, and high school students' evaluations of textbook pedagogical aids. *Teaching of psychology*, *26*(1), pp.19-21.

[17] Cui, H., Kan, M.Y. and Chua, T.S., 2007. Soft pattern matching models for definitional question answering. *ACM Transactions on Information Systems (TOIS)*, *25*(2), pp.8-es.

[18] Duan, W. and Yates, A., 2010, June. Extracting glosses to disambiguate word senses. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 627-635).

[19] Navigli, R., Velardi, P. and Faralli, S., 2011, June. A graph-based algorithm for inducing lexical taxonomies from scratch. In *Twenty-Second International Joint Conference on Artificial Intelligence*.

[20] Pantel, P. and Ravichandran, D., 2004. Automatically labeling semantic classes. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004* (pp. 321-328).

[21] Yang, H. and Callan, J., 2009, August. A metric-based framework for automatic taxonomy induction. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP* (pp. 271-279).

[22] Kozareva, Z. and Hovy, E., 2010, October. A semi-supervised method to learn and construct taxonomies using the web. In *Proceedings of the 2010 conference on empirical methods in natural language processing* (pp. 1110-1118).

[23] De Benedictis, F., Faralli, S. and Navigli, R., 2013, August. Glossboot: Bootstrapping multilingual domain glossaries from the web. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 528-538).

[24] Tuan, L.A., Kim, J.J. and Ng, S.K., 2014, October. Taxonomy construction using syntactic contextual evidence. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 810-819).

[25] Seitner, J., Bizer, C., Eckert, K., Faralli, S., Meusel, R., Paulheim, H. and Ponzetto, S.P., 2016, May. A Large DataBase of Hypernymy Relations Extracted from the Web. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 360-367).

[26] Klavans, J.L. and Muresan, S., 2001. Evaluation of the DEFINDER system for fully automatic glossary construction. In *Proceedings of the AMIA Symposium* (p. 324). American Medical Informatics Association.

[27] Espinosa-Anke, L., Saggion, H. and Ronzano, F., 2015, June. Taln-upf: Taxonomy learning exploiting crf-based hypernym extraction on encyclopedic definitions. In

Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015) (pp. 949-954).

[28] Degórski, L., Marcinczuk, M. and Przepiórkowski, A., 2008, May. Definition Extraction Using a Sequential Combination of Baseline Grammars and Machine Learning Classifiers. In *LREC*.

[29] Ide, N., Bonhomme, P. and Romary, L., 2000. XCES: An XML-based encoding standard for linguistic corpora. In *Proceedings of the Second International Language Resources and Evaluation Conference. Paris: European Language Resources Association*.

[30] Schmid, H., 2013, November. Probabilistic part-ofispeech tagging using decision trees. In *New methods in language processing* (p. 154).

[31] Jurafsky, D., 2000. Speech & language processing. Pearson Education India.

[32] Brants, Thorsten., 2000. TnT: a statistical part-of-speech tagger. In *Proceedings of the sixth conference on Applied natural language processing*. Association for Computational Linguistics.

[33] Faralli, S. and Navigli, R., 2013, August. A java framework for multilingual definition and hypernym extraction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 103-108).

[34] Westerhout, E. and Monachesi, P., 2007. Extraction of Dutch definitory contexts for elearning purposes. *LOT Occasional Series*, *7*, pp.219-234.

[35] Toutanova, K. and Manning, C., 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT Conference EMNLP/VLC, 63-71, 2000*.

[36] Branco, A. and Silva, J., 2006. A suite of shallow processing tools for portuguese: Lx-suite. In *Demonstrations*.

[37] Espinosa-Anke, L., Ronzano, F. and Saggion, H., 2015. Weakly supervised definition extraction. In *Angelova G, Bontcheva K, Mitkov R, editors. International Conference on Recent Advances in Natural Language Processing 2015 (RANLP 2015); 2015 Sept 7-9; Hissar, Bulgaria. Stroudsburg: ACL (Association for Computational Linguistics); 2015. p. 176-85.*. ACL (Association for Computational Linguistics).

[38] Spala, S., Miller, N.A., Yang, Y., Dernoncourt, F. and Dockhorn, C., 2019, August. DEFT: A corpus for definition extraction in free-and semi-structured text. In *Proceedings of the 13th Linguistic Annotation Workshop* (pp. 124-131).

[39] Boella, G., Di Caro, L., Ruggeri, A. and Robaldo, L., 2014. Learning from syntax generalizations for automatic semantic annotation. *Journal of Intelligent Information Systems*, *43*(2), pp.231-246.

[40] Esteche, J., Romero, R., Chiruzzo, L. and Rosá, A., 2017. Automatic Definition Extraction and Crossword Generation From Spanish News Text. *CLEI ELECTRONIC JOURNAL*, *20*(2).

[41] Ravichandran, D. and Hovy, E., 2002, July. Learning surface text patterns for a question answering system. In *Proceedings of the 40th Annual meeting of the association for Computational Linguistics* (pp. 41-47).

[42] Saggion, H., 2004, May. Identifying Definitions in Text Collections for Question Answering. In *LREC*.

[43] Montes-y-Gómez, M., Pineda, L.V., Pérez-Coutiño, M.A., Soriano, J.M.G., Arnal, E.S. and Rosso, P., 2005, September. INAOE-UPV Joint Participation in CLEF 2005: Experiments in Monolingual Question Answering. In *CLEF (Working Notes)*.

[44] Westerhout, E., 2009, September. Definition extraction using linguistic and structural features. In *Proceedings of the 1st Workshop on Definition Extraction* (pp. 61-67).

[45] Cafarella, M.J., Downey, D., Soderland, S. and Etzioni, O., 2005, October. Knowitnow: Fast, scalable information extraction from the web. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing* (pp. 563-570).

[46] Velardi, P., Faralli, S. and Navigli, R., 2013. Ontolearn reloaded: A graph-based algorithm for taxonomy induction. *Computational Linguistics*, *39*(3), pp.665-707.

[47] Fahmi, I. and Bouma, G., 2006. Learning to identify definitions using syntactic features. In *Proceedings of the Workshop on Learning Structured Information in Natural Language Applications*.

[48] Monachesi, P. and Westerhout, E., 2008. What can NLP techniques do for eLearning. *INFOS2008 proceedings*, pp.150-156.

[49] Espinosa-Anke, L. and Schockaert, S., 2018, June. Syntactically aware neural architectures for definition extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)* (pp. 378-385).

[50] Chen, C., Liaw, A. and Breiman, L., 2004. Using random forest to learn imbalanced data. *University of California, Berkeley*, *110*(1-12), p.24.

[51] Schroeder, J., Cohn, T. and Koehn, P., 2009, March. Word lattices for multi-source translation. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)* (pp. 719-727).

[52] Pustet, R., 2003. *Copulas: Universals in the Categorization of the Lexicon*. OUP Oxford.

[53] Marcus, M., Santorini, B. and Marcinkiewicz, M.A., 1993. Building a large annotated corpus of English: The Penn Treebank.

[54] Manning, C.D. and Schütze, H., 2002. *Foundations of Statistical Natural Language Processing*. 5th ed., MIT Press, Cambridge, MA.

[55] Manning, C.D., 2011, February. Part-of-speech tagging from 97% to 100%: is it time for some linguistics?. In *International conference on intelligent text processing and computational linguistics* (pp. 171-189). Springer, Berlin, Heidelberg.

[56] Tufiş, D., Barbu, A.M., Pătraşcu, V., Rotariu, G. and Popescu, C., 1997. Corpora and corpus-based morpho-lexical processing. *Recent Advances in Romanian Language Technology, Editura Academiei*, pp.35-56.

[57] Simionescu, R., 2011. Hybrid pos tagger. In *Proceedings of Language Resources and Tools with Industrial Applications Workshop (Eurolan 2011 Summer School), Cluj-Napoca, Romania* (pp. 21-28).

[58] Simionescu, R., 2012. Romanian deep noun phrase chunking using graphical grammar studio. In *Proceedings of the 8th International Conference "Linguistic Resources And Tools For Processing Of The Romanian Language* (pp. 135-143).

[59] Zafiu, A., Dumitrescu, S.D. and Boroş, T., 2015. Modular language processing framework for lightweight applications (MLPLA). In *7th Language & Technology Conference*.

[60] Straka, M., Hajic, J. and Straková, J., 2016, May. UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, pos tagging and parsing. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16) (pp. 4290-4297).

[61] Tufis, D., 2000, May. Using a Large Set of EAGLES-compliant Morpho-syntactic Descriptors as a Tagset for Probabilistic Tagging. In *LREC*.

[62] Science Encyclopedia, Litera Ed., 2018, 304 pg., Bucharest, ISBN 978-606-33-3101-5

Anexa I

```
public enum AnnotationType {
    LINKWORD("LINK"),
    PARTICIPLE_FORM("PART"),
    PARTICIPLE_FORM_COLLAPSED("PARTCOL"),
    GERUND_FORM("GER"),
    PREC_ATTRIBUTE("PATTR"),
    ADVERB_BREAK("ADVBR"),
    NPLIMIT("NLIM"),
    CONJUNCTION("CNJBR"),
    CONJUNCTION_ANDOR("CNJAO"),
    ADPOSITION("ADPBR"),
    PUNCTUATION("PUNCT"),
    NP_UNDEF("NPUNDEF"),
    NP_DEF("NPDEF"),
    NP_IGNORE("NP_IGN"),
    NP_TARGET("NP_TARG"),
    MULTIPLE_WORD_NOUN("MULTNOUN"),
    OVERRIDE("OVRD"),
    ONEOF("ONEOF"),
    COMPGRADE("CMPGR"),
    ADJ_LIKE("ADJ_LIKE"),
    ENUMERATION("ENUM"),
    OBLIQUE("OBLQ"),
    LW_CAREO("LW_CAREO"),
    LW_ALCAREO("LW_ALCAREO"),
    PP("PP");
}
```

Listingul A-1 Etichetele de adnotare predefinite

Anexa II

```
    //enum section
    private static final double ST_COMMA = 0;
    private static final double ST_AND_OR = 50;
    private static final double ST_MULTIPLE_CONJ = -100;
    private static final double ST_BOTH_AND_OR = -2000;
    private static final double ST_BOTH_COMMA_CONJ = -30;
    private static final double ST_SINGLE_NOUN_WITH_ENUM = -500;

    private static final double ST_UNDEFINED = -10000;


    //enum only
    //private static final double PREV_BONUS = 25;
    private static final double ST_COMMA_CONJ_HANDICAP = -25;
    private static final double ST_SIMILAR_HANDICAP_MULTIPL = -60;
    //noun only
    private static final double ST_ADJ_NOUN_MIX = -50;

    //noun section
    private static final int ST_IDENTICAL = 100;
    private static final int ST_VERY_SIMILAR = 80;
    private static final int ST_RGET_VERY_SIMILAR = 90;
    private static final int ST_SOMEWHAT_SIMILAR = 50;
    private static final int ST_TOTALLY_DIFFERENT = 0;
    private static final int ST_UNKNOWN = -1;

    private static final double ST_COMMON_PROPER_FRACTION = 0.8;

    private static final int ST_LW_MAX_SIMILARITY = 50;

    private static final int ST_MAX_SIMILARITY = 150;

    private static final int ST_NOUNTYP_COMMON = 35;
    private static final int ST_NOUNTYP_PROPER = 85;

    private static final boolean ST_GROUP_COMMON_AND_PROPER_NOUNS = true;
    private static final boolean ST_RGET_GROUP_COMMON_AND_PROPER_NOUNS = false;

    //used only for common nouns
    //the type order is DEFINED (0), UNDEFINED (1), NONE(2), UNKNOWN(3) article
on both columns and lines
    private static final int[][] artTypePairs_st = {
            {50, 40, 15, 0},
            {40, 50, 45, 0},
            {20, 30, 50, 0},
            {0, 0, 0, 0}
    };

    private static final int[][] artTypePairs_st_rget = {
            {50, 0, 0, 0},
            {0, 50, 45, 0},
```

```
            {0, 0, 50, 0},
            {0, 0, 0, 0}
    };

    //the case order is N_AC(0), G_D(1), UNKNOWN(2)
    //UNKNOWN type should not appear in practice
    private static final int[][] nounCasePairs_st = {
            {50, 0, 0},
            {0, 50, 0},
            {0, 0, 0}
    };

    private static final int ST_LW_SAME_LINKWORD = 10;

    private static final int LWM = ST_LW_MAX_SIMILARITY;

    //the type order is STRONG(0), MEDIUM(1), WEAK(2), STOP(3), NONE(4),
UNKNOWN(5)
    private static final int[][] lwTypePairs_st = {
            {LWM, 0, 15, 0, LWM, 0},
            {0, LWM, 0, 0, LWM, 0},
            {15, 0, LWM, 0, LWM, 0},
            {0, 0, 0, LWM, LWM, 0},
            {15, 15, 15, 15, LWM, 0},
            {0, 0, 0, 0, 0, 0}
    };

    private static final int[][] lwTypePairs_st_rget = {
            {LWM, 0, 15, 0, LWM, 0},
            {0, LWM, 0, 0, LWM, 0},
            {15, 0, LWM, 0, LWM, 0},
            {0, 0, 0, LWM, LWM, 0},
            {0, 0, 0, 0, LWM, 0},
            {0, 0, 0, 0, 0, 0}
    };
```

Listingul A-2 Constantele numerice folosite la detecția enumerațiilor