TEHNICI SIMBOLICE FUNDAMENTALE ALE LINGVISTICII COMPUTAȚIONALE

V. Cotelea, dr.hab. conf.univ.
Academia de Studii Economice din Moldova

Procesarea limbajului natural (PLN) şi-a început dezvoltarea, în calitate de disciplină, imediat după cel de-al Doilea Război Mondial [9], ca un mecanism folosit pentru traducerea documentelor. Acesta a fost unul dintre primele obiective computaționale cele mai investigate. Dar eforturile premature depuse pentru analiza şi modelarea limbajului uman s-au caracterizat printr-o aplicare a unei metode fără cunoştințe lingvistice şi cu o performanță redusă a tehnicii de calcul din acele timpuri.

Potrivit lui Covington, "*Procesarea limbajului natural constă în utilizarea calculatoarelor pentru a înțelege limbile (naturale) umane, cum ar fi engleza, franceza sau japoneza. Prin înțelegere, nu se presupune că computerul are gânduri, sentimente sau cunoştințe asemenea omului, dar se subînțelege că computerul poate să recunoască şi să utilizeze informații exprimate în limbajul uman*" [4].

2. NIVELURI DE CUNOSTINȚE ÎN PROCESAREA LIMBAJULUI NATURAL

Manaris şi Slator definesc un sistem de PLN ca pe unul care încapsulează un model al limbajului natural în algoritmi adecvați şi eficienți. În acest caz, tehnicile de modelare sunt larg asociate cu evenimente din multe alte domenii, printre care [10]:

- Informatica, care prevede metode de reprezentare a modelelor, proiectarea şi implementarea algoritmilor pentru instrumentele software.
- Lingvistica, care contribuie cu noi modele şi procese lingvistice.
- Matematica, care identifică modele formale şi metodele.
- Neuroştiința, care explorează mecanismele mentale şi alte activități creierului.

Dintre aceste domenii, lingvistica a oferit cunoştințele lingvistice despre limbile naturale. Această cunoaştere, în cadrul unui sistem de PLN, poate fi împărțită în niveluri definite în termeni de caracteristici declarative (ce) şi procedurale (cum), precum se arată în Tabelul 1. [10]. După cum se poate observa în acest tabel, cunoştințele lingvistice pot fi aranjate la diferite niveluri sau componente, deoarece structura oricărui limbaj uman se poate diviza natural între aceste niveluri [4].

Fonologic. Nivelul studiază sunetele limbii din punctul de vedere al valorii lor funcționale, stabilind inventarul de foneme ale unei limbi şi caracterul diferitelor variante ale acestora. Fiecare limbă are un alfabet de sunete care se disting, acestea numindu-se foneme. Obiectul fonologiei îl constituie, aşadar, sunetele ca realitate lingvistică, iar nu fizică sau fiziologică. Sunetele reprezintă materia sonoră. Astfel, nivelul fonologic tratează realizările acustice, de aceea, apar doar în sistemele de recunoaştere a vorbirii. Din punct de vedere tehnologic, prelucrarea vorbirii de către calculator este oarecum separată de restul prelucrării limbajului natural, dat fiind faptul că acest tratament al vorbirii este condiționat de analiza formei de undă a sunetului şi de recunoaştere a formelor, în timp ce restul nivelurilor depind de o programare simbolică şi un raționament automat.

Morfologic. Morfologia este ramura lingvisticii, care studiază regulile privind structura internă a cuvântului, adică regulile de combinare a morfemelor lexicale şi gramaticale în cuvinte, stabilirea paradigmelor lor în dependență de categoriile de gen, număr, caz etc. sau cuprinde regulile privitoare la modificările formale ale lor în diferite întrebuințări. Ideea generală se bazează pe faptul că morfemele individuale pot fi combinate pentru a forma cuvinte.

Morfemele sunt unitățile minimale de sens. Există două tipuri de morfeme: forma liberă, care poate să apară ca şi cuvinte separate şi forma legată, care nu poate apărea ca şi cuvinte în sine. Acestea din urmă sunt, de obicei, numite afixe. De exemplu, cuvântul englezesc "*unselfish*" este compus din trei morfeme, "*un*", "*self*" şi "*ish*"."*Self*" este o formă liberă, în timp ce "*un*" şi "*ish*" morfeme legate. În special, "*un*" este aici un prefix, "*ish*" este un sufix şi "*self*" este rădăcina.

Există trei procese morfologice principale utilizate în formarea cuvintelor:

- **Flexiunea.** Morfologia flexionară este preocupată de crearea cuvintelor noi, lăsându-le în aceeaşi categorie sintactică, dar schimbând relaţiile gramaticale, cum ar fi, de exemplu, plural, timpul trecut şi posesia. Cuvintele se formează cu ajutorul afixelor de inflexiune, care sunt conectate la morfemul liber. De exemplu, atât „*pom*”, ca şi „*pomi*” (morfemul liber "*pom*" are afixul de plural "*i*") sunt substantive.
- **Derivarea.** Morfologia derivaţională descrie modul în care sunt create cuvinte noi cu ajutorul unor afixe, trecându-le dintr-o categorie sintactică în alta. De exemplu, adjectivul „*naţional*” se derivă din substantivul „*naţiune*”.
- **Compoziţia.** Compoziţia se ocupă cu construcţia cuvintelor noi prin combinarea morfemelor libere.

Graniţa dintre flexiune (care furnizează diferitele forme ale unui cuvânt) şi derivare (care produce cuvinte noi pornind de la cele existente) este, uneori, neclară. O diferenţă esenţială o constituie faptul că numai derivarea poate introduce o schimbare de sens (prin introducerea de cuvinte noi). O altă deosebire constă în faptul că formele derivative ar putea să nu existe, în timp ce formele flexionare nu lipsesc aproape niciodată.

Indiferent dacă avem de-a face cu morfologia flexionară sau cu cea derivaţională, putem spune că, din punct de vedere computaţional, nivelul morfologic al limbii se ocupă de modul în care sunt alcătuite cuvintele pornindu-se de la unităţile de bază numite *morfeme*.

Tabelul 1. Niveluri de cunoştinţe în prelucrarea limbajului natural.

	Caracteristici	
Nivel	**Declarativ (ce)**	**Procedural (cum)**
Fonologic	Sunete vorbite	Formarea morfemelor
Morfologic	Unităţi de cuvinte, cuvinte	Formarea cuvintelor, Derivarea unităţilor de sens.
Sintactic	Rolul structural al cuvintelor (sau colecţii de cuvinte)	Formarea frazelor
Semantică	Semnificaţia independentă de context	Derivarea semnificaţiei frazelor
Discurs	Rolul structural al frazelor (sau colecţii de fraze)	Formarea dialogurilor
Pragmatic	Semnificaţia dependentă de context	Derivarea semnificaţiei frazelor ţinând cont de discursul ambiant

Sintactic. Sintaxa, sau construcţia propoziţiilor, reprezintă nivelul cel mai de jos la care limbajul uman este, în mod constant, în proces creare. Vorbitorii unei limbi creează mult mai rar unităţi fonice şi lexicale. În schimb, sunt concepute în mod constant noi propoziţii şi fraze. Noam Chomsky (1957) a fost primul care a evidenţiat acest aspect. El a introdus "*gramatica generativă*", în care frazele sunt descrise de anumite reguli, şi nu prin listarea lor şi a structurilor în mod direct. Astfel de reguli au devenit standard nu numai în lingvistică, ci şi în informatică, cu precădere în proiectarea compilatoarelor.

Cunoştinţele sintactice reprezintă o componentă de bază a oricărui sistem de PLN, care este responsabilă pentru recunoaşterea frazelor gramaticale şi atribuirea unor structuri acestora. Procesul de recunoaştere a structurii unei fraze de către calculator se numeşte analiza sintactică computaţională sau „*parsare*”.

Semantic. Semantica este o ramură a lingvisticii, al cărei obiect de studiu este *sensul*, unitate greu de abordat dintr-o perspectivă unică şi unitară. La acest nivel, semantica computaţională face un studiu al *sensului independent de context*. Cu alte cuvinte, interesează sensul pe care o propoziţie îl are fără legătura cu contextul în care ea a fost utilizată. Semantica frazelor este o parte esenţială a oricărui sistem, pentru că fără ea nu am putea atribui semnificaţie structurilor analizate.

Discursul tratează aspectele de interpretare afectate de frazele pronunţate anterior. La acest nivel, se acumulează cunoştinţele care se referă la legarea sensului frazei izolate pentru a se integra în unităţi mai mari. În particular, această cunoaştere este folosită pentru a interpreta pronumele anaforic, a soluţiona *elipsele* şi a interpreta aspectele temporale.

Anafora, în limbajul natural, constă dintr-o expresie, care se referă la o expresie anterioară a unui discurs. În general, se foloseşte un prenume pentru a se referi la persoane, locuri sau lucruri menţionate mai sus. De exemplu: "*Pasărea a murit. Ea era foarte bătrână* ". În propoziţia a doua „*ea*” se refera la pasăre. Pe de altă parte, elipsa: se referă la situaţii ale căror fraze sunt scurtate sau este eliminat vreun constituent, lăsând o parte din ele să fie înţelese din context. De exemplu, când cineva

este întrebat "*Care este numele tău*?" şi se răspunde "*Ion Potcoavă*", ultima este o formă eliptică a propoziției "*Numele meu este Ion Potcoavă*".

Această componentă este necesară sistemelor, pentru ca acestea să posede cunoştințe din contextul comunicativ, în care sunt produse mesajele şi să țină cont de aspectele pragmatice, precum intențiile expeditorului şi destinatarului.

Pragmatic. Se referă la utilizarea limbii în context. În general, pragmatica include aspecte ale cunoştințelor conceptuale ale lumii, care merg mai departe de condițiile reale ale fiecărei propoziții. Această cunoaştere este considerată atunci când se comunică într-o singură limbă.

Sunt folosite, pentru a înțelege, o mulțime de informații, subînțelese, dar nu şi exprimate în mod explicit în propoziții. În timp ce sintaxa şi semantica studiază propozițiile, pragmatica studiază "*acțiunile discursului*" şi situațiile în care limba este utilizată.

Multe cuvinte şi fraze pot fi ambigue şi să aibă mai mult de un sens, semnificația lor poate fi falsă sau să producă implicații false. Semnificația depinde de principiile pe care oamenii le folosesc atunci când vorbesc (de exemplu, să fie relevante şi se pune accentul pe fraze adevărate [6]).

În acest sens, pragmatica are două concepte importante: implicarea şi presupoziția propozițiilor. Implicarea unei propoziții conține informații care nu fac parte din sensul acesteia, dar trebuie să fie deduse de către un ascultător rezonabil. Presupozițiile unei propoziții sunt lucrurile care trebuie să fie adevărate pentru ca propoziția să fie adevărată sau falsă. Adică, în baza presupozițiilor (ipotezelor) (cunoaşterea adevărată a unui domeniu), oamenii interpretează fraze şi derivă cunoştințe (implicații), care pot fi sau nu adevărate.

2. PROCESAREA LIMBAJULUI NATURAL ŞI LINGVISTICA COMPUTAȚIONALĂ

În procesarea limbajului natural nivelul pragmatic tratează folosirea propozițiilor în diverse situații (contexte), precum şi modul în care contextul influențează interpretarea unei propoziții.

Cunoştințele lingvistice sunt încorporate în sistemul de PLN începând cu anii şaizeci, şi a devenit una dintre componentele sale majore. Pornind din acel moment a fost definită aria de cunoştințe numit Lingvistica Computațională .

2.1. Lingvistica computațională

Potrivit cercetătorilor în domeniu lingvistica computațională este o disciplină care se bazează pe două lucruri: limbile naturale şi calculatoare. Multe direcții de cercetare împărtăşesc ambele obiective, dar din perspective diferite. Ca întotdeauna, orice obiect nou de studiu se confruntă cu definirea terminologiei ştiințifice. Termenul lingvistica computațională este echivalent cu PLN şi nu este echivalent cu Lingvistica informatică sau Ingineria lingvistică.

Lingvistica informatică: o disciplină care se referă la utilizarea calculatoarelor în legătură cu limbaje şi limbi. Include toate tipurile de instrumente care asistă studierea limbilor străine şi a lingvisticii. Lingvistica computațională este o parte a lingvisticii informatice.

Ingineria lingvistică se referă la potențialele aplicații comerciale care implică utilizarea noilor tehnologii. Include ediții electronice (dicționare, cărți), produse multimedia etc.

Conform lui Grishman, lingvistica computațională poate fi definită ca "*studiul sistemelor informatice utilizate pentru înțelegerea şi generarea limbilor naturale*" [7]. Allen oferă o definiție echivalentă pentru procesarea limbajului natural "*Scopul acestei cercetări constă în crearea modelelor de calcul suficient de detaliate, care ar permite scrierea programelor care realizează diferite sarcini în ce priveşte limbajul natural*" [1]. Prin urmare, LC şi PLN tratează acelaşi lucru: dezvoltarea programelor de calculator care simulează capacitatea lingvistică umană.

Inteligența Artificială (IA) este responsabilă de codificarea într-un program a facultăților cognitive cum ar fi inferența, luarea deciziilor, achiziția cunoştințelor etc. În acest sens, LC este parte integrantă a IA, în acelaşi mod, cum, pentru mulți lingvişti, Lingvistica mai face parte din psihologie pentru tratarea unei dintre capacitățile cognitive, prin excelență, limba.

În continuare, termenii PLN şi LC sunt interşanjabili. Cu toate acestea, termenul PLN apare mai des, deoarece este mai bine înțeles.

Lingvistica computațională are mai multe aplicații practice, principalele fiind prezentate în clasificarea următoare:

1. Sisteme care încearcă să emuleze capacitatea omului de a procesa limbile naturale. În cadrul acestui grup, cele mai importante sunt: traducerea automată, recuperarea şi extragerea informațiilor, interfețe om-maşină.
2. Sisteme care ajută la îndeplinirea sarcinilor lingvistice. Acest grup este format din instrumente care pot fi utilizate de către lingvişti pentru facilitarea executării anumitor sarcini complexe. Unele aplicații, de acest tip, sunt: instrumentele de analiză textuală, bazele de date lexicografice, instrumentele de gestionare a corpusului. Corpusul

reprezintă o colecție de date lingvistice, de obicei, formată din mai multe texte. Această cantitate mare de texte, în limbaj natural, sunt utilizate pentru acumularea statisticilor necesare pentru analiza limbii.

3. Programe pentru ajutorarea la scrierea şi compoziția textului. Aplicațiile incluse în acest grup au fost foarte dezvoltate şi orice utilizator, obişnuit cu un procesor de text, este familiarizat cu ele: corectoare ortografice, corectoare sintactice şi de stil.

4. Instruirea asistată de calculator. Acesta este un câmp de aplicație în continuă expansiune şi are mai multe aspecte. Cel mai important este programul educațional pentru învățarea limbilor străine.

2.2. Limba din punct de vedere ştiințific

Definiția limbii, din punct de vedere al ştiinței, a determinat mulți lingvişti să convină asupra diferitelor puncte. Covington le prezintă pe cele mai importante [4]:

- **Limbajul este formă**, **nu substanță**. Acest lucru înseamnă că limba nu este un set de pronunții sau de comportament, ci este un sistem de reguli care determină comportamentul. Un alt mod de a exprima acest lucru este de a distinge între competența vorbitorului (sistemul) şi performanță (comportamentul observabil). Această distincție recunoaşte că declarațiile accidentale, frazele întrerupte etc. nu sunt, realmente, instanțe de limba vorbită de o persoana, ci sunt derivații de limbă.
- **Limba este arbitrară**. O limbă constituie un set de simboluri, pe care oamenii sunt de acord să le utilizeze într-o manieră specifică**.**
- **Toate limbile umane utilizează modele de dualitate,** în care cuvintele sunt şiruri de sunete, iar propozițiile sunt şiruri de cuvinte. Cuvintele au o semnificație, sunetele, în sine, nu.
- **Toate limbile sunt aproape la fel de complicate**, **cu excepția dimensiunii vocabularului.** Limbile se schimbă în mod constant, dar fiecare schimbare este lentă. O limbă evoluează într-o anumită direcție sute de ani.
- **Toată lumea vorbeşte despre propria lor limbă.** Limba română vorbită de o persoană nu este complet identică cu limba română vorbită de tatăl ei. Acest lucru se datorează modului în care limba este învățată. În procesul de învățare a limbii, apar mici diferențe între indivizi, şi deosebiri mari, inevitabile între grupuri sociale.

2.3. Probleme în utilizarea limbajului natural

Cunoaşterea lumii este un factor important în sistemele de PLN. Prin urmare, un sistem de PLN ar trebui să impună limite cu privire la necesitatea de cunoaştere externă şi experiența umană. Covington afirmă, în plus față de cele de mai sus, că PLN depinde de doi factori: primul se referă la puterea calculatoarelor. Apariția, în 1980, a microcalculatoarelor a schimbat situația. Anterior, PLN a fost atât de scumpă încât oamenii acceptau orice rezultat perfect, oricare ar fi fost atins. Această situație s-a schimbat şi sistemele PLN, încă imperfecte, au devenit mai ieftine, iar utilizatorii găsesc aplicări bune pentru ele.

Al doilea factor, şi poate cel mai important, constă în faptul că PLN depinde de cunoaşterea exactă a modului în care limbajul uman funcționează, lucru care, acum, nu se cunoaşte suficient. Până în ultimii ani, limba a fost studiată cvasi-exclusiv cu scopul de a o preda altei persoane. Principiul care stă la baza tuturor limbilor umane a era ignorat. Mai mult decât atât, ştiința lingvistică are doar câteva decenii vechime, şi nici nu există încă un consens cu privire la unele aspecte de bază. Sistemele PLN trebuie să abordeze o varietate de probleme în ce priveşte limbajul natural [10]:

- **Inexactitatea,** inclusiv erorile de ortografie, punctuația incorectă, cuvintele transpuse şi propozițiile negramaticale.
- **Incompletitudinea,** inclusiv în construcții eliptice, anafora etc.
- **Imprecizii,** inclusiv utilizarea de termeni relativi, fără un anumit punct de referință şi cu utilizarea de termeni calitativi.
- **Ambiguitatea,** deoarece multe interpretări pot apărea la orice nivel de cunoştințe lingvistice (a se vedea tabelul 1). Ambiguitatea poate fi rezolvată folosind cunoştințe de un nivel mai înalt.

2.4. Modele simbolice de procesare a limbajului natural

Modelele şi metodele PLN pot fi clasificate în: metode simbolice, empirice sau statistice, conexioniste şi abordări hibride. Primele două sunt numite modele matematice ale limbilor. Abordarea simbolică se bazează pe cunoaştere, utilizează reguli şi algoritmi care funcționează pe structuri de date simbolice şi reprezintă cunoaşterea limbajului natural. Abordarea empirică sau statistică implică colecții de mostre selective de limbă (corpus), care sunt etichetate şi folosite pentru a crea modele statistice pentru PLN. Tehnica conexionistă

utilizează rețelele neuronale pentru reprezentarea cunoştinţelor lingvistice. Pe de altă parte, tehnicile hibride combină una sau mai multe din modelele anterioare, pentru a completa beneficiile fiecăruia şi a rezolva problemele domeniilor şi aplicațiilor specifice.

Sistemele simbolice se bazează pe manipularea de simboluri. Ele au fost concepute de matematicieni pentru a capta într-o demonstrație riguroasă şi sistematică a teoremelor matematice şi logice. În lingvistică, Chomsky a fost primul care a introdus sistematic paradigma logicii formale.

De obicei, regulile de inferență, într-un sistem formal, permit să se concentreze pe sintaxa modelului, indiferent de interpretarea lui. Mulți lingvişti cred că limba are o natură bazată pe reguli sau logică şi că este ceea ce se încearcă să se reflecte în gramaticile formale. În general, aceste gramatici s-au dovedit eficiente în descrierea şi explicarea fenomenelor legate de competență. Competența se referă la cunoştinţele pe care fiecare vorbitor le are despre limba lui maternă.

3. GRAMATICI FORMALE

O gramatică formală este o specificație riguroasă şi explicită a structurii unei limbi. Acest lucru este redat cu ajutorul unui formalism gramatical, adică cu o limbă artificială creată pentru descrierea limbilor naturale. Utilizarea lor se datorează faptului că un limbaj bine definit, riguros, facilitează evaluarea ipotezelor şi permite elaborarea previziunilor.

Există diferite tipuri de gramatici, care sunt bine formalizate.

Acestea includ: gramatici generative, gramatici categoriale, gramaticilor de dependență, gramatici de lanţuri lingvistice Harris şi gramatici cu arbori adiacenți [7,14]. Cu toate acestea, cele mai răspândite sunt gramaticile generative, de asemenea, cunoscute sub numele de gramatici de structură a frazei sau sintagmatice propuse de Chomsky [3].

Gramaticile generative sunt constituite dintr-un set de reguli generative, care atribuie, în mod explicit, structura internă a propozițiilor. Aceste reguli, numite reguli de rescriere, operează pe mulțimi de elemente neterminale şi terminale. Atât gramaticile transformaționale, cât şi gramaticile de unificare sunt gramatici generative. Bach, în 1974, spunea că orice gramatică, care defineşte, în mod explicit şi precis, frazele unei limbi este o gramatică generativă [2]. Ele sunt cele mai răspândite în lingvistică computațională.

Chomsky a propus o clasificare a tipurilor de gramatici, aplicată la gramaticile generative sau sintagmatice, care a devenit celebră cu numele creatorului său, Ierarhia Chomsky. Această ierarhie este organizată în conformitate cu "puterea generativă slabă". Conceptul de putere generativă sau formală se referă la capacitatea de predicție a unei gramatici. În special, preocupările puterii generative se referă la tipul de fraze ale gramaticii, care pot fi recunoscute drept gramaticale.

Există patru tipuri de gramatici generative, fiecare definit de un anumită clasă de reguli, pe care le conține.

Tabelul 2 rezumă fiecare din cele patru tipuri de gramatici ale lui Chomsky, clasele de limbaje pe care le generează, tipul de automat, care le recunoaşte şi forma regulilor de producție. În tabel, A reprezintă un simbol neterminal şi α, β şi χ - şiruri de terminali şi neterminali, iar t - un simbol terminal. Şirurile α şi β pot fi vide, dar χ - nu.

Tabelul 2. Ierarhia Chomsky

Tip	Gramatici	Restricții asupra formei regulilor	Limbaje	Automate
0	Nerestricționate	Nicio restricție	Recursiv enumerabile	Maşinile Turing
1	Dependente de context	Partea dreaptă conține cel puțin simbolurile din partea stângă: $\alpha A\beta \to \alpha\chi\beta$	Dependente de context	Automatele linear limitate
2	Independente de context	Partea stângă poate avea doar un simbol neterminal: $A \to \chi$	Independente de context	Automatele cu stivă
3	Regulate sau cu stări finite	Regulile pot avea doar ultimele două forme: $A \to tB$ şi $A \to t$ sau $A \to Bt$ şi $A \to t$	Regulate	Automatele finite

Această clasificare este una teoretică, deoarece nu există tipuri pure de gramatici. În practică, gramaticile formale sunt modificate în funcție de nevoile specifice. În consecință, nu se poate uşor decide cărui tip aparține gramatica. Aici indică distanța dintre realizările teoretice şi practice, care este similar cu diferențele dintre lingviştii teoreticieni şi lingviştii computaționali.

Lingvistică teoretică: se axează pe analiza competenței vorbitorilor, utilizează, în principal, introspecția pentru a obține date şi are tendința de a ajunge la concluzii prin metode deductive. Obiectivele sale principale sunt de a obține o teorie gramaticală simplă, restrictivă, ținând cont de universaliile lingvistice.

În lingvistica computațională, s-au obținut anumite rezultate şi aplicații cu următoarele gramatici.

3.1. Gramatici regulate sau de stări finite

Gramaticile regulate sau, de asemenea, numite rețele de tranziție, sunt formate din noduri sau stări (reprezentate prin circumferințe) şi arce (reprezentate prin săgeți) etichetate. Fiecare arc reprezintă o tranziție între două stări. Există două tipuri speciale de stări: Stările inițiale (marcate cu o mică săgeată de intrare) sunt singurele care nu primesc alte săgeți de arcuri şi Stările finale (reprezentate cu o circumferință dublă), care sunt singurele, de la care nu pornesc tranziții la alte stări. Nu este posibilă tratarea recursivității cu o gramatică regulată, deoarece datele prelucrate sunt doar cele redate de starea în care se află.

Această gramatică de stări finite este aplicată la morfologie şi recunoaşterea vocabularului, lexicului, deoarece, în orice limbă, regulile de flexiune formează o mulțime aproape închisă şi mult mai mică decât regulile de sintaxă. În definitiv, au fost folosite în mai multe scopuri limitate ale limbajului, oferind o metodă foarte eficientă pentru calculator [11].

3.2. Gramatici independente de context

Termenul de Gramatică independentă de context (GIC) este un model specific propus pentru descrierea sintaxei unei limbi. Gramaticile pot fi folosite pentru descrierea oricărei limbi (naturale sau artificiale) şi acestea trebuie să respecte nişte constrângeri foarte simple. Constrângerile se referă la modul în care sunt generate clauzele. Strict vorbind, clauzele dependente trebuie să fie adiacente componentei de care depind. Majoritatea limbilor naturale par să urmeze acest comportament, cu o posibilă excepție idioma germană-elvețiană [5]. Mai multe teorii contemporane ale sintaxei limbajului natural sunt derivate ale schemei GIC.

Precum se vede în ierarhia Chomsky, GIC este o gramatică de tipul 2, care este formată dintr-un set de reguli (producții) şi un set de intrări lexicale (sau lexicul). Cu aceste reguli, se poate descrie structura sintactică a mai multor fenomene ale limbilor naturale. Aceste gramatici sunt capabile să recunoască şi să genereze propoziții.

Astfel, o GIC este o descriere formală a sintaxei unei limbi. În mod concret, descrierea este dată de setul de "reguli de producție", care definesc propozițiile bine formate ale limbii. Regulile, desigur, sunt ele însele, de asemenea, scrise într-un limbaj formal. Ca toate idiomele formale oficiale, acesta este definit de un vocabular şi sintaxă. Tabelul 3 relevă definiția mai strictă a GIC.

Tabelul 3. Gramatici independente de context.

Vocabularul	Regulile conțin trei tipuri de simboluri: • **Neterminale**, care corespund componentelor limbajului descris. Unul din simbolurile neterminale are o poziție specială. Acesta se numeşte simbol distinctiv. • **Terminale**, care corespund cuvintelor limbajului descris. • → - simbolul săgeata, care delimitează partea stângă a unei reguli de partea ei dreaptă.
Sintaxa	Propozițiile în limbajul GIC sunt reguli de producție (Sintaxa se referă aici la formatul regulilor înseşi, nu la limba pe care o descriu). O regulă de producție are următoarele proprietăți: 1. Acesta constă dintr-o parte stângă (LHS) şi o parte dreaptă (RHS), separate de o săgeată: LHS →RHS 2. LHS constă dintr-un singur simbol neterminal. 3. RHS constă din unul sau mai multe simboluri neterminale sau terminale.

Aceste gramatici oferă propozițiilor o structura ierarhică internă. Pot fi descrise exprimarea alternanței şi a opționalității. În afară de aceasta, gramaticile independente de context au proprietăți formale care facilitează proiectarea algoritmilor de parsare. Cu toate construcțiile recursive, care nu pot fi tratate cu gramaticile regulate. GIC permit, de asemenea, acestea, există probleme cu tratamentul anumitor fenomene lingvistice, precum ar fi constituenții discontinui, subcategorizarea şi concordanța.

Aici trebuie menționat că componentele discontinue sunt componentele care pot fi găsite în mai mult de o poziție structurală, iar subcategorizarea, în principiu, fiind un fenomen în care o structură a frazei lexico-semantice este prezisă în funcție de semantica verbală. Ea are importanță în sintaxă, deoarece se specifică combinațiile posibile de cuvinte. Verbele şi unele adjective admit o structură de complemente, iar subcategorizarea se referă la numărul şi categoriile complementelor fiecărui verb. În acelaşi timp, concordanța este un fenomen care are loc în mai multe limbi, în care cuvintele iau anumite flexiuni, în funcție de modul în care acestea se referă la alte cuvinte într-o propoziție. Această listă se referă la gen, număr. Un exemplu simplu are loc cu verbele la persoana a treia singular şi subiecții săi: "*ea dansează*", "*tu dansezi"*.

Din aceste motive, în practică, niciun sistem PLN, de o anumită acoperire, nu foloseşte versiunea pură a acestui tip de gramatici.

3.3. Gramatici de unificare cu caracteristici

Cele mai frecvente restricții care apar în gramaticile independente de context sunt reprezentate de fenomenele de concordanță şi subcategorizare. Gramaticile de unificare cu caracteristici tratează ambele cazuri [8], astfel, încât acestea sunt considerate ca un model de calcul mai cuprinzător şi restrictiv, în acelaşi timp, din cele cunoscute până în prezent.

Aceste gramatici sunt redate prin descrieri formale complexe cu utilizarea de caracteristici şi folosesc o operațiune generală de combinare şi verificare a informațiilor gramaticale, cunoscută sub numele de unificare [12].

Structura caracteristicilor este mecanismul de bază de reprezentare a informației despre unitățile lingvistice. Astfel, fiecare element de informații (unitate lingvistică) este asociat cu o structură caracteristică, unde o caracteristică reprezintă o pereche formată dintr-un atribut şi o valoare. Atributul poartă un nume care identifică caracteristica, de exemplu, "*număr* = *plural*", unde număr este atribut, iar plural - valoare. Valorile caracteristicilor pot fi valori complexe, şi care, la rândul lor, pot fi structuri caracteristice.

Informațiile conținute într-o structură caracteristică este combinată într-o nouă structură, prin operația de unificare. Pentru ca acest lucru să se întâmple, structurile de informații trebuie să fie compatibile, deoarece, în caz contrar, nu pot fi unificate. Compatibilitatea ține de natura caracteristicii şi valorii acesteia. Caracteristicile, care apar numai într-una din structurile unificate, sunt incluse în structura rezultatului unificării, reuşind să combine informații comune şi diferite. Acest fapt permite ca diferitele structuri informaționale să poată fi combinate coerent.

Astfel, tendința lingviştilor de a utiliza gramatici mai restrictive, a condus la faptul că gramaticile independente de context s-au extins cu caracteristici, folosind mecanismul de unificare [12].

3.4. Structura unui sistem PLN simbolic

Se consideră că orice software de PLN are două tipuri mari de cunoştințe stocate:

1. **Cunoştințele lingvistice** în formă de gramatică, vocabular şi model conceptual al lumii. Gramatica este, pur şi simplu, o definiție abstractă a unui set de elemente structurate şi bine formate. Ar fi echivalentă cu competența noastră lingvistică şi pragmatică.

2. **Programul** sau **parserul**, care conține instrucțiunile pentru procesarea datelor lingvistice. Parser-ul este un algoritm sau un set de instrucțiuni, care leagă secvențele de simboluri cu cunoştințele lingvistice stocate. Parserul este un mecanism de calcul, care a deduce structura secvențelor de cuvinte pornind de la cunoştințele stocate în gramatică şi dicționar şi stabileşte dacă şirurile de caractere sunt, din punct de vedere gramatical, corecte sau incorecte. Problema pe care trebuie să o rezolve parserul este una pur sintactică: recunoaşterea frazelor gramaticale şi atribuirea unei structuri. Alte componente sunt responsabile pentru interpretare. Figura 1 prezintă structura generală a procesului de analiză [13].

Algoritmii de parsare sunt responsabili pentru determinarea regulilor ce trebuie aplicate şi în ce ordine. Fiecare algoritm îmbină, de obicei, diferiți parametri şi diferite structuri de lucru. Există mulți algoritmi, dar toți se bazează pe o

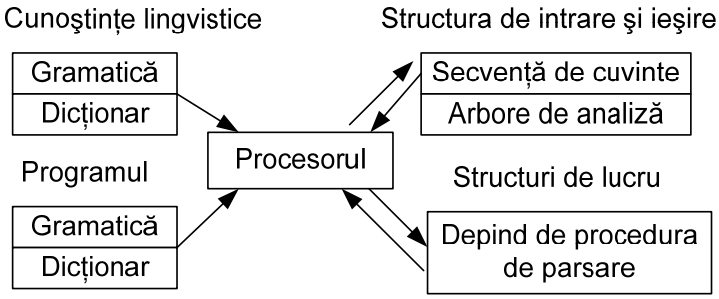

Figura 1. Structura generală a procesului de parsing

combinație de trei parametri esențiali care sunt luați în considerare: analiza descendentă (top-down) / analiza ascendentă (bottom-up), procesarea secvențială / procesarea paralelă, procesarea deterministă / procesarea nedeterministă. Având în vedere aceşti parametri, pot fi menționaşi câțiva algoritmi de parsare: (1) Algoritmul descendent în serie cu backtracking [7] şi (2) Algoritmii cu Chart [1,14]

Modelele simbolice reprezintă paradigma predominantă în LC, repertoriul lor de concepte şi metode este mai amplu şi au fost aplicate în multe probleme şi limbi. Dintre acestea, cele mai utilizate sunt automatele finite (pentru simplitate şi eficiență de procesare) şi gramaticile independente de context, completate cu gramatici de unificare cu caracteristici (pentru puterea expresivă de a ține cont de fenomenele lingvistice).

Bibliografie

1. Allen, J. *Natural Language Understanding. Addison-Wesley, 2 edition 654 pages, 1994.*

2. Bach, E. *Syntactic Theory. Univ Prof Amer, 310 pages, February 1982.*

3. Chomsky, N. *Syntactic structures. De Gruyter; 2nd edition, 117 pages, 2002.*

4. Covington, Michael A. *Natural Language Processing for Prolog Programmers. Englewood Cliffs, NJ: Prentice Hall, 348 pages, 1994.*

5. Gazdar, G. Mellish, C. *Natural Language Processing in PROLOG. An Introduction to Computational Linguistics, Addison Wesley, 504 pages, 1989.*

6. Grice, H., P. *Logic and conversation. In Cole, P. and J.L. Morgan, eds. Syntax and semantics. vol. 3, Speech acts. NY: Academic Press, p. 41-58, 1975.*

7. Grishman, R. *Computational Linguistics: an introduction. Cambridge, Cambridge University Press, 193 pages, 1986.*

8. Kay, M. *Parsing in functional unification grammar. In Dowty, D., Kartunnen, L. and Zwicky, A. pp 251-278, 1985.*

9. Locke W.N. and Booth A.D. *Machine Translation of Languages. Technology Press of MIT ans Wiley, Cambridge, Mass., p.15–23, 1955.*

10. Manaris, Bill Z. and Slator, Brian M. *Interactive Natural Language Processing: Building on Success. IEEE Computer, vol.29, Nr.7, p.28-32, 1996.*

11. Roche, E. and Schabes, Y. *Finite-State Language Processing. Cambridge, The M.I.T. Press, 482 pages, 1997.*

12. Shieber, S. *An Introduction to unification-based approaches to grammar. Chicago, Chicago University Press, 120 pages, 2001.*

13. Winograd, T. *Language as a Cognitive Process: Syntax. Reading, Addison-Wesley, 654 pages, 1983.*

14. Winograd, T. *Understanding Natural Language. New York: Academic Press, 191 pages, 1972.*

Recomandat spre publicare:16.05.2013.