

# Definirea semanticii propoziției în limbaj natural

Anatol Popescu, U.T.M.

Sergiu Crețu, A.S.E.M.

**Abstract.** This study represents a research concerning the principles and methods of elaboration of an interface in natural language (NL) for informational systems. When we say interpretation of texts in a NL, we mean the process of analyzing and extracting information contained in NL texts, realized according to a precise semantic context.

## 1. Definirea cadrului conceptual.

În această lucrare vom elabora un mecanism de definire a semanticii propoziției în limbaj natural reieşind din semantica lexicală fixată de vocabular și din regulilor sintactice de combinare a valorilor lexicale în unități sintactice.

Pentru interfețele sistemelor informatice este caracteristică reprezentarea informației sub formă de texte în limbaj natural. Implementarea unor astfel de interfețe pentru sistemele informatice cere elaborarea unor mecanisme, care asigură interpretarea sintactico-semantică a textelor remise interfeței.

Baza acestor mecanisme este asigurată de:

1. Componenta lingvistică reprezentată de un vocabular. Vocabularul definește în fond semantica lexicală a textului remis interfeței, adică sensul unităților lexicale ale limbajului.
2. Componenta sintactică care definește sintaxa frazelor (propozițiilor) limbajului utilizat pentru interacțiunea utilizatorului cu sistemul informatic.
3. Componenta semantică ce definește așa numita semantică propozițională a textului în limbaj natural folosit.

Practic, fiecare dintre componentele enunțate anterior reprezintă un cadru conceptual relativ independent. Ele pot fi reunite într-un singur sistem doar utilizând anumite tehnici integrative elaborate cu acest scop. De remarcat, aceste tehnici integrative devin posibile pe baza Principiului de Compoziționalitate al lui Frege. Acest principiu presupune pentru articulația sintaxă – semantică a textului existența unui homomorfism între unitățile sintactice și semantice ale limbajului natural.

Toate acțiunile, care pot fi inițiate prin intermediul procesului de consultare a vocabularului în vederea stabilirii sensului fragmentului textual interpretat, pot fi descrise prin intermediul unei algebre multi-sort [1,2]. Această teză are un caracter fundamental pentru toată expunerea ulterioară. Vom concretiza pe scurt conceptele ce țin de algebrele multi – sort.

**Definiția 1.** Un sort este un element  $s$  al unei mulțimi finite  $S$ , care servește drept indice pentru o familie de mulțimi  $\langle P_s \rangle$  numite mulțimi portante .

Familia de mulțimi  $\langle P_s \rangle$  este o mulțime de grafuri reprezentând sintaxa abstractă a fragmentului de text [3]. Amintim, că sintaxa abstractă în accepțiunea Bacus-Naur este o colecție finită de producții, ale căror parte dreaptă începe cu un simbol special. El identifică univoc producția (numele clasei operaționale) și este urmată de unul sau mai multe sorturi . De exemplu :

$\langle \text{predicat} \rangle ::= \text{Verb} \langle \text{verb} \rangle \langle \text{nume} \rangle$

**Definiția 2.** Signatura  $S$ -sortată  $\Sigma$  reprezintă o mulțime  $\Sigma_{w,s}$ , unde  $w \in S^*$  (șiruri finite din elementele mulțimii  $S$ ), iar  $s \in S$ . Elementele sunt considerate simboluri operaționale de orientarea  $w$  și de sortul  $s$ .

**Definiția 3.** Fie  $\Sigma$  o signatură  $S$ -sortată, atunci  $\Sigma$ - algebră este o mulțime  $P_s$  de mulțimi portante cu o funcție  $f$  pentru fiecare element  $\sigma \in \Sigma_{w,s}$  și  $w = s_1 \dots s_n$  de forma:

$f\sigma: P_{s_1} \times \dots \times P_{s_n} \rightarrow P_s$ .

O importanță deosebită în atribuirea de sens unui text în LN o are noțiunea de homomorfism între  $\Sigma$ -algebre, deoarece vocabularul, practic, reprezintă o  $\Sigma$ -algebră  $S$ -sortată.

**Definiția 4.** Fie date două  $\Sigma$  - algebre S-sortate A și B. Prin aplicație homomorfică a algebrei A pe algebra B se subînțelege o familie de funcții de forma:

$$g_s: P_S^A \rightarrow P_S^B, \text{ unde } P_S^A \text{ și } P_S^B$$

sunt mulțimile portante ale algebrelor A și B, care satisfac condiția:

$$g_s(f\sigma_A (p_1^A, \dots, p_n^A)) = f\sigma_B(g_s(p_1^A), \dots, g_s(p_n^A)).$$

Prin aceasta se semnifică faptul că la o interpretare se păstrează structura semantică a textului interpretat. Dacă homomorfismul din definiția 3.4 este unic, atunci o astfel de  $\Sigma$  - algebră se numește inițială. Pentru astfel de algebre nu există posibilitatea unei ambiguități semantice.

Pentru algebrele inițiale se asigură îndeplinirea condiției: dacă  $\sigma \in \Sigma_{w,s}$ , atunci  $\sigma(p_1, \dots, p_n) \in P_s$  unde  $w = s_1, \dots, s_n$ .

Încă o noțiune algebrică, extrem de utilă în contextul interpretării textelor în limbaj natural, este noțiunea de derivator.

**Definiția 5.** Fie  $\Sigma$  o semnătură S1-sortată și  $\Omega$  o semnătură S2-sortată. Atunci un derivator din  $\Sigma$  în  $\Omega$  este:

- funcția  $f: S1 \rightarrow S2$ ;
- familia de funcții  $d: \langle \Sigma_{w,s} \rightarrow (T_\Omega)_{f(w),f(s)} \rangle$ , unde  $w = s_1, \dots, s_n$ , iar  $f(s_1, \dots, s_n) = f(s_1), \dots, f(s_n)$ .

Partea dreaptă a aplicației este formată de mulțimea tuturor  $\Omega$  - termenilor de sortul  $f(s)$ . Simbolurile folosite la scrierea termenilor au sorturile  $f(s_i)$ . Pentru simplitate derivatorul se va nota  $\langle f, d \rangle$  precizând ambele componente ale derivatorului.

**Definiția 6.** Fie un derivator  $\langle f, d \rangle$  pentru semnăturile  $\Sigma$  și  $\Omega$ , atunci algebra B de semnatura  $\Omega$  va fi algebră d-derivată, dacă ea va avea portantele  $dB = B_{f(s)}$ , unde  $s \in S$ , iar simbolurile de operații  $f\sigma_{dB}$  vor fi simbolurile definite de  $d(\sigma)$ .

Cele expuse mai sus ne permit să formulăm rezultatul principal al paragrafului dat: atribuirea de sens unor unități sintactice poate fi tratată ca o consecutivitate de acțiuni, care se înscrie în cadrul homomorfismelor multi – sort definite de derivatorii algebrici.

Textul în limbaj natural, supus interpretării sintactico - semantice, formează limbajul sursă – domeniul de definiție al derivatorului, iar limbajul de interpretare – domeniul de variație al aceluiași derivator.

Limbajul de interpretare reprezintă un limbaj, clauzele căruia pot fi interpretate nemijlocit prin executarea pe calculator, de exemplu.

Derivatorul pune în corespondență fiecărui term (graf sintactic abstract) din limbajul sursă un fragment al limbajului de interpretare. Limbajul de interpretare, în caz particular, poate fi un sublimbaj al limbajului sursă. În acest caz procesarea poartă un caracter recursiv.

## 2. Structura limbajului semantic

În paragraful precedent a fost determinat locul limbajului semantic în definirea semanticii propoziției. Este necesar ca acest limbaj să fie reprezentat de o sintaxă și o semantică bine determinată, prin urmare el trebuie să fie formal.

Semantica limbajului semantic (LS) trebuie să fie consistentă și simplă pentru a fi exprimată formal. În același timp, S trebuie să fie în măsură să exprime semantica textelor în limbaj natural. Drept punct de plecare pentru elaborarea unui astfel de limbaj au fost luate acele acțiuni, pe care le întreprinde agentul, când determină semnificația unor părți de vorbire pe baza consultării vocabularului morfo – semantic din baza de cunoștințe [4].

Limbajele formale sunt definite prin specificarea tipurilor și structurilor de date asigurate de limbaj, prin multitudinea de operații cu tipurile de date precum și prin operatori, care reprezintă proceduri de manipulare a datelor.

Limbajul semantic LS are trei tipuri de date elementare: numeric (NUM), textual (TXT), logic (LOG). El posedă, de asemenea, mijloace de identificare a datelor. O noțiune importantă a limbajului semantic S o constituie noțiunea de sintagmă. Sintagmele în lingvistică sunt, în general, grupuri de cuvinte, sensul cărora nu poate fi principial redus la valorile semantice ale părților constituente. În limbajul semantic S, sintagmele denumesc grupuri de valori cu sens precis, dar, spre deosebire de lingvistică, sensul sintagmelor limbajului LS

poate fi sintetizat pe baza sensurilor părţilor componente.

Limbajul semantic LS conţine următoarele tipuri de sintagme:

- a) sintagme executabile, care prestează un rezultat material în urma executării unor şiruri de entităţi dinamice, separate prin delimitatorul ";";
- b) sintagme reţele semantice, elementele cărora au o structură specifică.

Se va specifica fiecare tip de sintagme, enumerat mai sus:

a) <sintagmă executabilă> ::= <nume sintagmă>:(<şir sintagme executabile>)

<şir sintagme executabile> ::= <identificator date> | <şir sintagme > ;<identificator

date >  
<identificator date > ::= <identificator date fără parametri> | <identificator date cu parametri>

<identificator date fără parametri> ::= <nume sintagmă >.{\*| <valoare> |< identificator date>}

<identificator date cu parametri> ::= <identificator>( <şir parametri>)

< şir parametri> ::= <parametru> | < şir parametri>, <parametru>

<parametru> ::= <identificator date >

b) < sintagmă reţea semantică> ::= <nume sintagmă>:(<nume sintagmă executabilă> IN <şir sintagme>)

<şir sintagme> ::= <sintagmă executabilă> | <sintagmă reţea semantică> | <şir sintagme>; <sintagmă executabilă> | <şir sintagme>; <sintagmă reţea semantică>

<nume sintagmă executabilă> ::= <nume sintagmă>

<nume sintagmă> ::= <identificator>

Exemplificăm specificările de mai sus:

(1) Identificator de date:  
VOCABULAR.VERB.\*

Trebuie interpretat astfel: se identifică toate verbele (sintagma VERB) din sintagma VOCABULAR. Semnul "\*" joacă rolul cuantificatorului –  $\forall$  din logica matematică.

(2) Sintagmă executabilă:  
EXECUT:(VOCABULAR.VERB.\*)  
EXECUT este numele sintagmei executabile. Rezultatul executării sintagmei constă în regăsirea tuturor verbelor din vocabular (specificate de identificatorul de date VOCABULAR.VERB.\* scris în paranteze).

(3) Sintagma reţea semantică:  
CONTEXT:( SUBST IN  
TRANZ:VOCABULAR.VERB.  
TRANZITIV.\*;SUBST;  
SUBST:  
VOCABULAR.SUBSTANTIV.  
om;ADJ;

ADJ:

VOCABULAR.ADJECT  
IV.frumos;)

Semantic executarea sintagmei  
CONTEXT are loc astfel:

**Pasul 1.** Se execută şirul de sintagme şi identificatori de date situat după cuvântul cheie IN. Pentru a avea acces la sintagmele interioare trebuie să existe un nume de sintagmă interioară identic cu numele de sintagmă situat anterior lui IN. Dacă această cerinţă nu este satisfăcută, atunci executarea sintagmei e terminată.

**Pasul 2.** În caz contrar se execută acea sintagmă, care deţine numele specificat anterior lui IN.

**Pasul 3.** Executarea continuă până există nume de sintagme interioare ce pot fi accesate.

Trebuie de remarcat că sintagma CONTEXT poate conţine în interior alte sintagme CONTEXT. În acest caz e valabilă regula blocurilor imbricate din limbajele de programare de tip PASCAL etc.

Elementele sintagmelor reţea semantică pot fi modelate prin intermediul producţiilor, care este una din interpretările posibile ale sintagmelor. Sistemele producţionale modelează eficient luarea deciziilor, de aceea îşi găsesc aplicaţia ca parte componentă în sistemele expert, sistemele cu inteligenţă artificială etc.

Producţia este o aplicaţie (în caz particular – o funcţie), cu domeniul de definiţie şi domeniul de variaţie bine definit:

$$D \rightarrow V$$

Aceeaşi situaţie are loc când construcţiilor sintactice li se pun în corespondenţă valorile semantice (semantica denotativă). În acest sens producţia de mai sus poate fi concretizată astfel:

$$D \rightarrow V^*, \text{ unde asteriscul „}^* \text{” denotă şiruri finite de valori.}$$

Ultimă concretizare conduce la forma producţională ("P" este folosit pentru desemnarea producţiilor):

$$P: T \rightarrow (D \rightarrow V^*)$$

T din formula de mai sus reprezintă mulţimea textelor în limbaj natural,  $V^*$  reprezintă rezultatul interpretării sintactico-semantice a textelor.

Dacă notăm conţinutul din paranteze prin S, atunci obţinem:

$$P: T \rightarrow S$$

Expresia este o aplicaţie din mulţimea textelor (entităţi sintactice) în S. Mulţimea S este vocabularul morfo - semantic (D sunt cuvintele cheie ale vocabularului, iar  $V^*$  este glosa) a priori dat.

Prin urmare, se poate conchide că multitudinea de producţii din T în S reprezintă un derivator din signatura algebrei iniţiale T în signatura algebrei iniţiale S

Revenim la limbajul semantic LS. Producţia în acest limbaj poate fi interpretată astfel:

$$\langle \text{nume sintagmă} \rangle: (\langle \text{şir sintagme} \rangle)$$

Ea mai poate fi interpretată şi ca o proto – procedură. Partea stângă (faţă de simbolul "":") a producţiei este antetul proto – procedurii şi este reprezentat de numele sintagmei, iar partea dreaptă - corpul proto – procedurii – şirul de sintagme.

**Observaţie.** Variabila  $\langle \text{nume sintagmă} \rangle$  înlocuieşte sintagma (sintagmele) de unde s – a efectuat transferul. Partea dreaptă – corpul proto-procedurii – conţine indicatori de date, sintagme, reprezentate sub formă de listă.

Procesul de atribuire a valorilor semantice (a sensului) pentru un fragment de text prin intermediul sintagmelor este un proces recursiv.

Reieşind din aceasta, vom reuni sintagmele acestui proces recursiv în cadrul unei construcţii speciale cu numele de vocabular virtual. El reprezintă o cale de acces la cunoştinţele înmagazinate în vocabular concret şi defineşte de fapt semantica propoziţională a fragmentului de text analizat.

Vocabularul virtual acţionează ca un compilator în sensul larg al cuvântului. El asigură interpretarea sintagmelor, reprezentate de frazele limbajului natural, în termenii semanticii propoziţionale. Această interpretare se bazează pe concepţia de flux al instrucţiunilor elementare.

### Concluzii

1. Sintetizarea sensului fragmentului de text analizat se bazează pe informaţia morfologică, sintactică şi semantică a textului interpretat.
2. Pentru definirea semanticii lexicale şi a semanticii propoziţiei textului în LN a fost elaborat un limbaj semantic, în cadrul căruia semantica se defineşte prin intermediul – vocabularului şi a vocabularului virtual.

### Bibliografie

1. Creţu S., Popescu A., Interpretarea sintactico – semantică a textelor în LN, Proceedings of the 5<sup>th</sup> International Conference on „Microelectronics and Computer Science”, v.II, Chisinau, 2007, p. 181-184.
2. Creţu S., Elaborarea unui mecanism formal de interpretare sintactico – semantica a textelor în limbaj natural, in Proceedings of the 30th Annual Congress of the American – Romanian Academy of the Arts and Sciences(ARA), Chişinău, 2005, p. 131 – 133.
3. Creţu S., A system for natural language text syntactic – semantic interpretation (SSI), in the 2<sup>nd</sup> supplement of the review Informatica Economică, International Conference Knowledge Management: Projects, Systems and Technologies, Bucharest, vol. 1, November, 2006, p. 171 – 174.
4. Creţu S., Interpretarea semantică a textelor în limbaj natural, EduSoft, Bacău, 2007, 208p.