MEANING OF THE SENTENCE IN THE NATURAL LANGUAGE: SEMANTIC INSIGHTS

[1]S. Crețu, assoc.prof., dr., [2]A. Popescu, prof. dr. hab.
[1]E.S. Academy of Moldova,
[2]Technical University of Moldova

INTRODUCTION: THEORETICAL FRAMEWORK

The current study focuses on the development of certain techniques to accurately determine the meaning of phrases, be they written or spoken, in the natural language (NL). Complex phrases can be easily broken down into simpler sentences (syntactic units) and words (lexical units). Hence, the meaning of phrases could be extracted/derived from theirs four underlying components: 1. the lexical component; 2. the syntactic component; 3. the semantic component; 4. the pragmatic component.

The lexical component of a phrase refers to a vocabulary. A typical vocabulary comprises a set of lexical units, each having defined meanings (the meaning of the lexical unit).The syntactic component defines the order of the lexical and syntactic units within a sentence and phrase, respectively. Accordingly, the semantic component specifies the relation of the syntactic units (sentences) to a set of facts amenable to interpretation. The meaning of the lexical and syntactic units depends on additional factors: 1. timing (i.e., when the unit was written or pronounced); 2. location (i.e., where the unit was written or spoken); 3. modality (i.e., how the unit was written or spoken). These factors taken together constitute the pragmatic component of a phrase.

Although individual components (lexical, syntactic, semantic, pragmatic) can be readily determined, theirs relation to the overall meaning of the phrase is not straightforward. A common approximation (Frege's Principle of Compositionality) assumes a homomorphism between the syntactic and the semantic components of the phrase in the NL. However, this classical approach has several drawbacks [1]: 1. the way complex phrases in the NL are divided into syntactic units for analysis influences theirs overall meaning; 2. neglecting the pragmatic component of phrases in the NL leads to erroneous estimation of theirs meaning, especially in the case of non-assertive phrases (e.g., orders, directives); 3. the true emitter (author) of phrases in the NL cannot be accurately clarified (this is referred to as the "*game of the subjects*" conundrum). Here we propose an integrated theoretical framework aimed at defining intrinsic relations between both the lexical, syntactic, semantic components and the illusive pragmatic component.

1. COMPETENCE MODELS

To simply further analyses, the lexical, syntactic and pragmatic components of phrases in the NL were redefined as competences (Ch. Morris, N. Chomsky):

Definition 1.1: The syntactic competence refers to the ability of the speaker (emitter, author) to generate correct linguistic phrases with or without meaning**.**

Definition 1.2: The semantic competence refers to the ability of the speaker (emitter, author) to establish semantic relations between the lexical and syntactic units of phrases in the NL. There are several types of semantic relations: 1. inclusion relations; 2. reference relations; 3. consistence relations; 4. coherence relations.

Definition 1.3: The pragmatic competence refers to the ability of the speaker (emitter, author) to apply correct linguistic phrases in a proper syntactic-semantic context.

The relations between the competences are unclear and poorly defined. As a simplification, a hierarchical model has been postulated: 1. the syntactic competence constitutes a prerequisite for the other two competences; 2. the semantic competence with its underlying connections "charges" the emitted linguistic phrases with meanings; 3. the pragmatic competence results from the integration of the syntactic and semantic competences with an *a priori* "experience" (prior to comprehension), presumably stored in a knowledge base. The semantic competence

has a static-dynamic character. Thus, certain semantic relations (e.g., synonymy) can be stored in a vocabulary. Conversely, the pragmatic competence has a dynamic character and cannot be compiled in a vocabulary. This issue can be alleviated in two ways: 1. explicit definition of pragmatics; 2. construction of semantic networks.

Therefore, to model the ability of the speaker (emitter, author) to generate complex phrases in the NL bearing meanings, three competence models were generated: 1. a syntactic model, describing the employed syntax; 2. a semantic model, storing the semantic component; 3. a pragmatic model, defining and describing the pragmatic component. Each model functions independently, has its own formal language and relates hierarchically to the other two models using interpretation rules.

2. THE SYNTACTIC COMPETENCE MODEL

The syntactic competence model was developed using a categorial grammar [2], comprising the following categories: 1. N – proper nouns (singular); 2.CN – common nouns (singular, nominative case); 3. IV – intransitive verbs (infinitive); 4. TV – transitive verbs (infinitive); 5. VP – verbal phrases; 6. S – sentences. Having a clear lexical meaning, N, CN, IV and TV represent primary categories. For these categories, B was defined as the set of names of the basic categories:

$B=\{B_N, B_{CN}, B_{IV}, B_{TV}\}$

, where B_N, B_{CN}, B_{IV}, B_{TV} represent labels of the sets of the used proper nouns, common nouns, intransitive verbs and transitive verbs, respectively.

Conversely, the values of VP and S were derived only from the interpretation rules. The syncategorematic entities cannot be defined as categories (e.g., conjunctions). Thus, they were included in the syntactic rules, used to assemble linguistic phrases.

The employed syntactic rules were generated using the following formalism:

<construct> ::= <condition>, then <conclusion>

, where <condition> is a logical expression.

Thus, for a given sentence S (i.e., N-TV-CN) its corresponding syntactic rule can be constructed as:

S_n. $\alpha \in TV \wedge \beta \in CN, then\ \alpha'\beta' \in VP$

S_m. $\chi \in N \wedge \delta \in VP, then\ \chi\delta' \in S$

, where n, m $\in Z^+$, α, β, χ, δ – categories and the apostrophe " ' " – an inflection.

Example: Let "John hoists the flag" be a phrase to be modeled. In this case, its primary categories are specified as: $B_N=\{John\}$, $B_{CN}=\{flag\}$, $B_{TV}=\{hoist\}$, and the corresponding syntactic rule is:

S1. "hoist" $\in$ TV and "flag" $\in$ CN, then "hoist flag'" $\in$ VP.

S2. "John" $\in$ N and "hoist flag'" $\in$ VP, then "John hoists the flag" $\in$ S**.**

The described above formalism can be conveniently simplified by the introduction of two additional operators: 1. the /(A,B) operator – specifying the rightward location of a given A category with respect to a given B category; 2. the \(A,B) operator – specifying the leftward location of a given A category with respect to a given B category. For sentence S (3) the TV category can be rewritten as \(N,/(CN,S)). This expression exactly posits the TV category within the S sentence: TV is located to the right of the N category (N is placed to the left of TV) and to the left of the CN category (CN is positioned to the right of TV). Therefore, the TV and, analogically, the VP categories can be precisely expressed using the N, CN and S categories and hence excluded from the defined basic categories. Thus we can conclude with a definition:

Definition 2.1. The categorial grammar **G**, defined for the **V** vocabulary, is a finite relation as follows:

$$G \subseteq V \times Cat(B)$$

, where the **V** is the vocabulary – a finite set with its elements representing the words of a NL, **B** – a countable set of categories, including a special S category (the set of the basic categories), Cat(**B**) – the algebra of the terms generated with the "/" and the "\" operators and containing the B set. **G** defines a single category for each element of the **V** vocabulary is considered to be a classical categorial rigid grammar [2].

Definition 2.2. I. For every **V** vocabulary of terminal elements two reduction rules can be applied to the definition 2.1:

1) FA (forward application) - /(A,B) A $\rightarrow$ B.
2) BA (backward application) – A \(A,B) $\rightarrow$ B.

II. In general, a set of categories from Cat(B) should be attributed to every element of the **V** vocabulary with the help of the "/" and "\" operators.

III. I. and II are necessary and sufficient to generate the L language:

$$L=\left\{c_1..c_n\in V^*\middle|\forall i\{1,...,n\},\exists A_i\in Cat(B)\wedge A_1..A_n\xrightarrow[FABA]{}S\right\}$$

Example: Let "John expertly hoists the flag" be a phrase to be modeled. The rigid classical categorical grammar (CCG) models this phrase is composed of :

1. The basic categories of the **B** set: **B**={N, CN, S}.
2. The **V** vocabulary: **V**={John, flag, expertly, to hoist}.
3. The rigid CCG:

G={<John, N>, <flag, CN>, <to hoist, \(N, /(CN, S))>, <expertly, \(\(N, /(CN,S)), \(N, /(CN, S)))>}.

Another kind of theory implied Lambek grammars [3] may be done.

3. THE SEMANTIC COMPETENCE MODEL

To efficiently interpret a sentence the NL sentence should be converted to a logical object. Conversion of a sentence from the NL to logical object relies on a specific logical language. The used logical language should be a typed one. That is, for each logical object we are to assign his type. In general, the type is a label refers to a subset of elements belonging to a set containing all the elements in use for interpretation. This universal set, usually, is named as Universe. For example, the vocabulary **V** containing all the NL words may be considered as Universe set.

Definition 3.1. The **Type** set is a minimal set which includes the following elements:

1. $e\in Type$. Element e denotes the individuals – the elements belong to Universe.
2. $t\in Type$. Element t denotes just only two values: true and false, also belonging to Universe.
3. If $a\in Type$ and $b\in Type$, then $\langle a,b\rangle\in Type$, where <a, b> - a function with its definition domain D_a (a set of the type a) and variation domain D_b (a set of the type b).

For example, the type expression <e, t> refers to a set of Universe's individuals and <<e, t>, t> is an expression denotes a second degree predicate.

The proposed logical language has two components: 1. the syntactic component; 2. the semantic component. The syntactic component comprises:

A. A set containing all the types for a given vocabulary as Universe (definition 3.1);
B. A set of all non-logical constants - Con (e.g., Con_a - the set of the constants of the type a);
C. A set of all the variables - Var (e.g., Var_a – the set of the variables of the type a).
D. A set of all the expressions of the type a ME_a
E. The following syntactic rules are available:
1. If a is a variable of the type a, then $v_a\in ME_a$.
2. If a is a constant of the type a, then $c_a\in ME_a$.
3. If $\alpha\in ME_b$ and $v\in Var_a$, then $\lambda v\alpha\in ME_{<a,\,b>}$.
4. If $\alpha\in ME_{<a,\,b>}$ and $\beta\in ME_a$, then $\alpha(\beta)\in ME_b$.
5. If $\alpha,\beta\in ME_a$, then $\alpha=\beta\in ME_t$.
6. If $\varphi\in ME_t$ and $\psi\in ME_t$, then $\neg\varphi,[\varphi\wedge\psi],[\varphi\vee\psi],[\varphi\to\psi],[\varphi\leftrightarrow\psi]\in ME_t$.
7. If $\varphi\in ME_t$ and $u\in Var$, then $\forall u\varphi\in ME_t$.
8. If $\varphi\in ME_t$ and $u\in Var$, then $\exists u\varphi\in ME_t$.

The semantic component embodies:

A. A model M interpreting the syntactic rules:
M=<I, F, g>, where I – a non-null set of elements form the Universe, F – a function attributing values of the type a to every single constant from the Con_a set, g - a function attributing values of the type a to every single variable from the Var_a set
B. The following semantic rules:

1. If α is a constant, then $|\alpha|^{M,g}=F(\alpha)$.
2. If α is a variable, then $|\alpha|^{M,g}=g(\alpha)$.
3. If $\alpha\in ME_{<b,\ a>}$, and $\beta\in ME_b$, then $|\alpha(\beta)|^{M,g}=|\alpha|^{M,g}(|\beta|^{M,g})$.
4. If $\varphi\in ME_t$, then $|\neg\varphi|^{M,g}=1$, if and only if $|\varphi|^{M,g}=0$ or the other way around.
5. If $\varphi\in ME_t$ and $\psi\in ME_t$, then $|\varphi\wedge\psi|^{M,g}=1$ if and only if $|\varphi|^{M,g}=1$ and $|\psi|^{M,g}=1$.
6. For $\vee,\to,\leftrightarrow$ similar to 5.
7. If $\varphi\in ME_t$ and $v\in Var_a$, then $|\forall v\varphi|^{M,g}=1$, if and only if $\forall e\in D_a$, where D_a- a domain of the type a, $|\varphi|^{M,g,v/e}=1$ and v/e - substitution.
8. If $\varphi\in ME_t$ and $v\in Var_a$, then $|\exists v\varphi|^{M,g}=1$, if and only if $\exists e\in D_a\ |\varphi|^{M,g,v/e}=1$.

Comment: The λ - sign represents the λ - operator from λ - calculus. For example, the expression λ p[$\forall$ xp(x)] is of the <<e, t>, t> type and denotes the set of characteristics (of second degree predicate) of the elements from the adopted Universe. Conversely, the λ x[p(x)] expression is of the <e, t> type and specifies the elements of the Universe having the p as predicate.

4. INTERPRETATION OF THE LINGUISTIC PHRASES.

To accurately interpret the linguistic phrases generated with the rigid CCG approach in the context of the proposed logical language a correspondence between the syntactic categories and the semantic has to be defined.

Definition 4.1. I. For the basic grammar categories (definition 2.1) a morphism f should be parsed as follows:

1. N $\rightarrow e$, proper nouns are associated with the elements of the V vocabulary;
2. S $\rightarrow t$, sentences are associated with the t element (true, false) from Universe;
3. CN $\rightarrow \langle e,t\rangle$, common nouns are associated with the first degree predicates;

II. For the other categories of the Cat(**B**) set the following relation should be defined: f(\(A, B))= $\langle f(A), f(B)\rangle$ and f(/(A,B))=$\langle f(A), f(B)\rangle$, where A and B $\in$ Cat(B).

Example: Let "John expertly hoists the flag" be a phrase to be interpreted. Using the morphism f, described above, it can be easily derived that:

<John, N> $\rightarrow e \rightarrow John''$

<flag, CN> $\rightarrow \langle e,t\rangle \rightarrow \lambda x[\, flag''(x)]$

<hoist,\(N,/(CN, S))>

$\rightarrow \langle e,\langle\langle e,t\rangle,t\rangle\rangle \rightarrow \exists x[\, flag''(x) \wedge hoist''(John'', x)]$

<expertly, \(\(N, /(CN,S)), \(N, /(CN, S)))> $\rightarrow \langle\langle e,\langle\langle e,t\rangle,t\rangle\rangle,\langle e,\langle\langle e,t\rangle,t\rangle\rangle\rangle \rightarrow$

$\exp ertly''(\exists x[\, flag''(x) \wedge hoist''(John'', x)])$

Comments: 1. The transitive verb "to hoist" has been extensionally interpreted.

2. In the latter expression the following semantic rule was used: $|\alpha(\beta)|^{M,g} = |\alpha|^{M,g}(|\beta|^{M,g})$.

3. The elements followed by a double apostrophe represent translations of the words into the logical language.

The formulas in the logical language generated with the morphism f are limited by the analyzed sentence. However, they can be further generalized using the λ - operator. For example, the transitive verb "to hoist" can be described as:

$$\lambda N[\lambda A[\lambda D \exists x[D(x) \wedge A(N,x)]]]$$

5. NATURAL LANGUAGE: INTENSIONAL ASPECTS

The described model follows an extensional approach: it is assumed that the semantics of phrases in the NL can be derived from the interpretation of theirs components (words, sentences). However, this is only a simplification. In reality, a plethora of factors influence the semantics of phrases in the NL are: 1. contexts; 2. modal contexts; 3. temporal contexts; 4. intensional contexts. These factors taken together constitute some extra-linguistic objects. To support an intensional approach the proposed formalism [4] was extended as follows:

A. Definition 3.1 (extended). The Type set is a minimal set which includes the following elements:

1. $e \in Type$.
2. $t \in Type$.
3. If $a \in Type$ and $b \in Type$, then $\langle a,b\rangle \in Type$.
4. If a $\in$ Type, then <s, a> $\in$ Type.

, where <a, b> - a function with its definition domain D_a (a set of the type a) and variation domain D_b (a set of the type b**)**, s – the third object added to model the contexts. B. The syntactic rules of the logical language (extended):

1. If a is a variable of the type a, then $v_a \in ME_a$.
2. If a is a constant of the type a, then $c_a \in ME_a$.
3. If $\alpha \in ME_b$ and u $\in Var_a$, then $\lambda u\alpha \in ME_{<a, b>}$.
4. If $\alpha \in ME_{<a, b>}$ and $\beta \in ME_a$, then $\alpha(\beta) \in ME_b$.
5. If α , $\beta \in ME_a$, then $\alpha = \beta \in ME_t$.
6. If $\varphi \in ME_t$ and $\psi \in ME_t$, then $\neg\varphi,[\varphi \wedge \psi],[\varphi \vee \psi],[\varphi \rightarrow \psi],[\varphi \leftrightarrow \psi] \in ME_t$.
7. If $\varphi \in ME_t$ and u $\in$ Var, then $\forall u\varphi \in ME_t$.
8. If $\varphi \in ME_t$ and u $\in$ Var, then $\exists u\varphi \in ME_t$.
9. If $\alpha \in ME_a$, then $^\wedge\alpha \in ME_{<s, a>}$.
10. If $\alpha \in ME_{<s, a>}$, then $^\vee\alpha \in ME_a$.

C. The semantic rules of the logical language (extended):

1. If α is a constant, then $|\alpha|^{M,g} = F(\alpha)$.

2. If α is a variable, then $|\alpha|^{M,g} = g(\alpha)$.

3. If $\alpha \in ME_{<b,\ a>}$, and $\beta \in ME_b$, then $|\alpha(\beta)|^{M,g} = |\alpha|^{M,g}(|\beta|^{M,g})$.

4. If $\varphi \in ME_t$, then $|\neg\varphi|^{M,g} = 1$, if and only if $|\varphi|^{M,g} = 0$ or the other way around.

5. If $\varphi \in ME_t$ and $\psi \in ME_t$, then $|\varphi \wedge \psi|^{M,g} = 1$ if and only if $|\varphi|^{M,g} = 1$ and $|\psi|^{M,g} = 1$.

6. For $\vee, \rightarrow, \leftrightarrow$ similar to 5.

7. If $\varphi \in ME_t$ and $v \in Var_a$, then $|\forall v\varphi|^{M,g} = 1$, if and only if $\forall e \in D_a$, where D_a- a domain of the type a, $|\varphi|^{M,g,v/e} = 1$ and v/e - substitution.

8. If $\varphi \in ME_t$ and $v \in Var_a$, then $|\exists v\varphi|^{M,g} = 1$, if and only if $\exists e \in D_a\ |\varphi|^{M,g,v/e} = 1$.

9. If $\alpha \in ME_a$, then $|^{\wedge}\alpha|^{M,gw,t,g}$ is a function f with the domain $W \times T$, which satisfies the following : $\forall \langle gw', t' \rangle \in W \times T \rightarrow f(\langle gw', t' \rangle)\ is\ |\alpha|^{M,gw',t',g}$

10.If $\alpha \in ME_{<s,a>}$,then $|^{\vee}\alpha|^{M,gw,t,g}\ is\ |\alpha|^{M,gw,t,g}(\langle gw, t \rangle)$.

Comments: The s element permits to operate with extra-linguistic objects and to define the functions of the <s, a> type. The $D_{<s,a>}$ domain contains functions of this type and forms the domain of all possible senses (meanings of). These senses can be extracted with the $^{\wedge}$ operator and have the general form <w, t>, where w – the index of the meaning and t – its temporal component [5]. The intensional of a linguistic phrase $^{\wedge}\alpha$ is a function applied to <w, t>.

The extended model M interpreting the syntactic rules comprises:

1. I – a non-null set of elements form the Universe;
2. F – a function attributing the intensional to every constant the Con set;
3. T – the non-null set of temporal components with the defined < relation;
4. W – the non-null set of all possible worlds (meanings of all extra-linguistic objects).

We have considered necessarily modifying the structure of possible worlds, because the structure of all possible worlds is a syntactical one. Therefore, this extended formalism permits modeling of extra-linguistic objects using categorial grammars: the index <gw, t> includes the categorial grammars gw.

Example. The type expression <s, e> denotes the individual concept, the expression <s, <<s, e>, t>> refers to the properties of sets of concepts of individuals, but - <s, <e, t>> represents the properties of individuals. The properties of individuals may be express by the formula: $^{\wedge}\lambda x[P(x)]$ and, finally, the transitive verb "to hoist" from p.2 will be representing in logical language as:

$$hoist^{"}(John^{"}, ^{\wedge}\lambda A \exists x[flag^{"}(x) \wedge A(x)])$$

CONCLUSIONS AND PERSPECTIVES:

This study is aimed at developing systems for the interpretation of natural language texts. In fact, here was treated sentence interpretation - a particular case. Even at this level there are enough problems. The proposed approach allows to elaborate a system that would take into account the complexity of the problem. Many problems remain unresolved in theory. For example, it is important to investigate the structure of possible worlds, the relationship between the situation and the type of analyzed sentence: assertions, orders etc.

Bibliography

1. ***Searle, J. R**. The Nature of Intensional States.// In: Intensionality, Cambridge, 1983, pp. 1-29, Cambridge University Press, 1983.*
2. ***Casadio, C**. Categorial Grammars and Natural Language Structures. // D. Reidel Publishing Company, Dordrecht, 1988, Semantic Categories and the Development of Categorial Grammars, pp.95-123.*
3. ***Lambek, J**. The Mathematics of Sentence Structure. // American Mathematical Monthly, 65 (1958), pp.154-170.*
4. ***Montague, R**. Universal Grammar. // Reprinted in Formal Philosopfy; Selected Papers of Richard Montague, 1974, pp.222-246.*
5. ***Montague, R**. The Proper Treatment of Quantification on Ordinary English // Reprinted in Formal Philosopfy; Selected Papers of Richard Montague, 1974, pp. 247-270.*

Recommended for publication:22.06.2013.