V. S. Yakovishin

FORMAL PROTO-LANGUAGE IN A TEXT PROCESSING SYSTEM

Introduction. The presented text processing system is based on the use of a formal language that represents a dibasic ring-like algebra with a set of words over an alphabet and a set of sentences on which one unary operation and two binary operations are defined. So in the formal language, text sentences are expressed as semantic formulas using both words and special operation signs: the words can represent all lexical meanings; unary operation symbols can represent all general-sentence grammatical meanings (such as modality, negation, question, etc.); and symbols of binary operations can represent functional meaning of words [1].

Thus, in the proposed formal description, every natural language sentence assumes the shape of an algebraic expression – *semantic formulas*. A set of such algebraic expressions, representing "reconstructed" deep structures of the given natural language, can be considered as a *formal proto-language* (FpL) or, specifically, a *formal proto-English, formal proto-Russian*, etc. Then one can say that any natural language (NL) differs basically from various artificial formal languages (FL) only in having its natural (non-formal) surface manifestations.

It is obvious that a full description of the natural language must include not only sets of *grammatical* rules (syntax and semantics), that generate a formal proto-language, but also sets of *algorithmic* rules realizing an interpretation of generated formal language, i.e., its conversion into surface manifestation. Or, in other words: a realization of the full description of natural language can be considered as creating a text processing system in which the formal proto-language is used as a semantic representation of text information. As will be shown below, the presented formal proto-language can be used for the knowledge representation in the text processing system.

Initial syntax. We assume that any language description contains a set of rules of *initial* (algebraic) syntax. The initial syntax of text sentences can represent one unary operation, denoting the syntactically independent words, and two binary operations, denoting known syntactic connections – the determination ("grammatical multiplication") and coordination ("grammatical addition"). The determination is non-associative, non-commutative, and distributive over the coordination; and the coordination is associative and commutative [1; 2].

According to the said operation properties, the initial syntax of natural language can be specified as the following set of rules:

$$S\rightarrow\surd S|(S\Delta X)|(X\Delta S)|(X\nabla X)|X$$
$$X\rightarrow X\nabla X,$$

where $\surd$ is an unary operation; Δ (“determination”) and ∇ (“coordination”) are binary operation. That is, the associativity and commutativity are specified as parentheses-free rules in which the same symbol is used for both members: $S\nabla S$, $X\nabla X$; the parentheses in $(X\nabla X)$ are necessary to express the distributivity of determination over summand strings like $(X\nabla X\nabla\ldots\nabla X)$; the non-commutative property of determination is specified as a possibility to use combinations of two different elements: $S\Delta X$ and $X\Delta S$; the parentheses must be used to indicate the non-associative property of determination: $(S\Delta X)$, $(X\Delta S)$.

Generated by the initial syntax sentences have the form of abstract parenthesis expressions indicating various *syntagmatic structures* (“dependency structure”). The elementary syntagmatic structure is called the *syntagme* (aka the "syntagma” or "syntagm”). The syntagm can now represented as a two-part string in the form of $(\alpha\Delta\beta)$, where $\alpha\in A^*$, $\beta\in A^+$ (it can be an *elliptical* syntagme, if α is the empty word). The first part α is a *head* (independent member) and the second part $\Delta\beta$ is its *modifier* (dependent member); the determination symbol Δ represents a dependency meaning and plays the role of a relation feature in syntagmatic oppositions between the marked modifier and the unmarked head.

Note that sentence structures can be expressed not only in the standard algebraic form, using the usual means of syntactic connections, i.e., parenthesis with the fixed order. For the purposes of representation of sentences in computing systems, it is convenient to use a parenthesis-free form. In the proposed here formal language notation, syntactic structures are expressed by the set-theoretical form in which all given syntagmes are expressed as certain set’s elements [3]. So the known types of syntactic structures can be given by the following both forms:

(*a*) $((X_1\Delta_1X_2)\Delta_2X_3)\Leftrightarrow\{X_1, X_1\Delta_1X_2, X_1\Delta_2X_3\}$,
(*b*) $(X_1\Delta_1(X_2\Delta_2X_3))\Leftrightarrow\{X_1, X_1\Delta_1X_2, X_2\Delta_2X_3\}$,
(*c*) $(X_1\Delta(X_2\nabla X_3))\Leftrightarrow\{X_1, X_1\Delta X_2, X_1\Delta X_3\}$,
(*d*) $((X_1\Delta_1(X_2\Delta_2X_3))\Delta_3X_4)\Leftrightarrow\{X_1, \Delta_3X_4, X_1\Delta_1X_3, X_1\Delta_3X_4\}$,

where (*a*) is the *collateral subordination,* or the parallel connection, in which all modifiers Δ_1X_2 and Δ_2X_3 are subordinated to a single head X_1; (*b*) is the *consecutive subordination,* or the stepwise connection, in which the modifier of preceding syntagme X_2 serves as the head of the succeeding syntagme $X_2\Delta_2X_3$; (*c*) is a sentence structure with *homogeneous* (homofunctional) parts ΔX_2 and ΔX_3; (*d*) is a sentence structure with an extreme modifier Δ_3X_4, i.e., emphasized part.

Thus, in the generated structures, two extreme parts of the sentence can be expressed. These are the extreme head and the extreme modifier. The extreme head, known as “subject”, is a part of the sentence that occupies the absolutely independence position (the root node) of the syntagmatic structure. As an absolutely independent member the extreme head is expressed in languages by a zero form, i.e., a syntactically neutral (unmarked) noun form (known as the "nominative", "indefinite case", "casus rectus"). The extreme modifier, known as "emphasized part", "determinant", "theme", 'topic', is a part of the sentence that relates as a modifier to the whole sentence and not associates with any of its individual word-head.

Secondary syntax. Semantics. Combinatory capabilities of real words are specified by a *secondary* syntax. The secondary syntax can be induced on a postulated algebraic syntax using special symbols to denote various word categories, *parts of speech*, and functional word forms, *parts of the sentence.*

The parts of speech will be specified in terms of the existing dependency relation. That is, the belonging of various words to one and the same part of speech is grounded on possibility of each of these words as a head to attach the proper modifier as the agreed inflectional word-form. So the *noun* X_{n} can have the adjective modifier $a_{\mathrm{n}}S$ (*the big boy*); the *verb* X_{v} can have the adverbial modifier $a_{\mathrm{v}}S$ (*to run quickly*); and the *quality* X_{q} can have the ad-quality modifier $a_{\mathrm{q}}S$, i.e., the "adverb of degree" (*unusually big*). It seems that only these three syntactic parts of speech exist in languages. The other word classes (such as the pronoun and the numeral) will be considered as *semantic* categories.

Thus, we can say that the given word represents a part of speech X_{k}, if it has (as a head word) the special *agreed* modifier $a_{\mathrm{k}}S$. It means that on the basis of the initial algebraic syntagme $(X\Delta S)$ we can obtain an induced syntagme $(X_{\mathrm{k}}\ a_{\mathrm{k}}S)$ for $k\in\{n,\ v,\ q\}$. Similarly, we can get the categories of words that play the role of heads in constructions with *governed* modifiers. Such syntactic categories are the valence words, in particular the univalent $X_{\mathrm{v}1}$ and bivalent $X_{\mathrm{v}2}$ transitive verbs.

Using the rules of algebraic syntax together with the obtained secondary rules, we can construct various secondary syntactic systems. So we have received a multilevel hierarchy of syntagmes with the following types of modification.

Free modification: realized by $S\to(S\Delta X)$, where ΔX (modifying a whole sentence) can be considered as an emphasized part (determinant).

Predicativity: realized by $X\to(X_{\mathrm{n}}\ \Delta_{\mathrm{p}}S)$, where $\Delta_{\mathrm{p}}S$ (predicates) modify a noun phrase representing a subject.

Attributivity: realized by rules in the form $X_{\mathrm{k}}\to(X_{\mathrm{k}}a_{\mathrm{k}}S)|(X_{\mathrm{k}}a_{\mathrm{k}}\Delta S)$, where $a_{\mathrm{k}}S$ or $a_{\mathrm{k}}\Delta S$ (attributes) can modify the noun (adjective attributivity), verb (adverbial attributivity), or quality (attributivity of the quality).

Valence modification: realized by $X_{\mathrm{v}}\to(X_{\mathrm{v}1}\ o_1X_{\mathrm{n}})$ and $X_{\mathrm{v}1}\to(X_{\mathrm{v}2}\ o_2X_{\mathrm{n}})$, where the modifiers are distinguished as the (direct) *object* (o_1X_{n}) and *indirect object* (o_2X_{n}).

In the stage of semantic interpretation, the operation symbols are replaced by special designations denoting the various semantic elements, known as "grammatical meanings". The unary operation can be interpreted by means of the usual logical modal operators ("possibly", "necessarily", etc.) and usual punctuation symbols (e.g., question and exclamation marks); it can also be represented by usual independent words or their abbreviations. The coordination symbol can be replaced using the special words (grammatical codes), for example: $\nabla\to\nabla_c|\nabla_{dj}\ |\nabla_{adv}\ |\ldots$, where the special words (∇ with subscripts) denote the conjunctive (*and*), disjunction (*or*), adversative (*but*), etc. Similarly, the determination symbol is replaced by special words used to link the subordinate adverbial clauses (formed from ΔS): $\Delta \rightarrow \Delta_{cd}|\Delta_{cs}|\Delta_{pl}|\ldots$, where the special words denote the condition (*if*), cause (*why?*), place (*where?*), etc.

The initial determination used to link the subordinate words can be replaced by meanings of the free (adverbial) modifiers (formed from ΔX). In the text, these meanings are expressed by the case endings and prepositions; and the lexical part (a word stem) of these subordinate words (as adverbial modifiers) can be expressed either by nouns or modifying root words ("adverbs"), i.e.

$$\Delta X \Rightarrow \Delta_{\mathrm{m}} X_{\mathrm{n}} \mid mX$$

where $\Delta_{\mathrm{m}} X_{\mathrm{n}}$ is a modifying noun (*in the evening*, *on the table*); mX is a modifying root word (*yesterday*, *here*). The further concretization of Δ_{m} can be realized using the signs (prefixes) denoting the various case meanings of nouns. For example:

$$\Delta_{\mathrm{m}} \rightarrow abs.|\ in.|\ inl.|md.|\ sb.|\ sp.|\ trs.\ |\ldots$$

where *abs.* is a meaning of the abessive case (*without whom/what?*); *in.* is a meaning of the inessive case (*inside whom/what?*); *inl.*is a meaning of the illative case (*into whom/what?*); *md.* is a meaning of the mediative/instrumental case (*by whom/what?*); *sb.* is a meaning of the subessive case (*under whom/what?*); *sp.* is a meaning of the superessive case (*on whom/what?*); *trs.* is a meaning of the transitive case (*during whom/what?*).

The predicativity is concretized as combinations of tense, aspect, and mood:

$$\Delta_{\mathrm{p}} \rightarrow p.|\ pt.\ |\ pf.|\ pCt.|\ p_{\mathrm{p}}Ct.|\ p_{\mathrm{f}}Ct.|\ldots$$

where *p.*, *pt.*, and *pf.* denote the present, past, and future indefinite tense; *pCt.*, $p_{\mathrm{p}}Ct.$, and $p_{\mathrm{f}}Ct.$ denote the present, past, and future continuous tense, etc.

All abbreviations denoting the grammatical (relational) meanings of the modifying words can be represented as prefixes attached to usual word stems:

*abs.*glass (*without glasses*); *in.*garden (*in the garden*); *inl.*house (*into the house*); *md.*train (*by train*); *sb.*table (*under the table*); *sp.* (*on the table*); *trs.*forest (*through the forest*); *pt.*write (*wrote*), *pCt.*write (*to be writing*); a_{v}quick (*quickly*); $o_1$book ([to read] *a book*).

The lexical meanings can be represented by the following word-formative rules:

$$X_{\mathrm{n}} \rightarrow R_{\mathrm{n}}\ |RD_{\mathrm{n}}\ |R_{\mathrm{c}}\ |R_{\mathrm{p}},$$
$$X_{\mathrm{v}} \rightarrow R_{\mathrm{v}}\ |RD_{\mathrm{v}}\ |R_{\mathrm{c}},$$
$$X_{\mathrm{q}} \rightarrow R_{\mathrm{q}}\ |R_{\mathrm{q}}D_{\mathrm{q}},$$
$$R \rightarrow R_{\mathrm{n}}\ |R_{\mathrm{v}}\ |R_{\mathrm{q}}\ |R_{\mathrm{c}},$$

where R_{n} ("noun"), R_{v} ("verb"), R_{q} ("quality"), R_{c} ("numeral"), and R_{p} ("pronoun") are the categories of primary meanings (semantic parts of speech); D_{n}, D_{v}, and D_{q} are the categories of grammatical (derivational) meanings of the respective word categories.

The primary meanings are denoted by usual root morphemes: $R_{\mathrm{n}} \rightarrow$book$|\ldots$, $R_{\mathrm{v}} \rightarrow$go$|R_{\mathrm{v}1}|\ldots$, $R_{\mathrm{v}1} \rightarrow$read$|R_{\mathrm{v}2}|\ \ldots$, etc. The derivational meanings are denoted by special postfixes: $D_{\mathrm{n}} \rightarrow .n|.d|.pl|\ldots$, $D_{\mathrm{v}} \rightarrow .v|.rv\ |\ \ldots$, $D_{\mathrm{q}} \rightarrow .q\ |.cp\ |\ .sp|\ldots$, where *.n* (“noun”), *.d* (“definiteness”), *.pl* (“plural”), etc. are derivational meanings of the noun; *.v* (“verb”), *.rp* (“repeating”), *.rv* (“reverse”), etc. are derivational meanings of the verb; *.q* (“quality”), *.cp* (“comparative”), *.sp* (“superlative”), etc. are derivational meanings of the quality. For example:

big.*n* (*bigness*); big.*cp* (*bigger*); read.*n* (*reading*); slow.*n* (*slowness*); five.*n* (*quintuple*); **modern.v (modernize);** glory.*v* (*glorify*); drink.*q* (*drinkable*); man.*pl* (*men*); turn.*rv* (*return*).

Note that the known two types of “grammatical meanings” become now the purely formal definition: the *relational* meaning is a sign of the word dependency (and hence it cannot be used in the position of the absolute independent word); the *derivational* meaning is a word modifier (i.e., it can be used in any syntactic position, including the absolute independent part of the sentence). It is known that the distinctions between the relational and the **derivational** meanings are observed at the surface level as distinctions of morphological means: the relational meanings are expressed by case endings or prepositional cases; the derivational meanings are expressed by word-formative affixes or particles and articles. But these distinctions are not always necessarily. For example, the word-formative affix **-ly** (that changes adjectives into adverbs) must be considered as a sign of the relational meaning: $a_\text{v}$slow “**slowly**”.

Using the above syntactic and semantic rules, we can represent sentence descriptions on syntactic and semantic levels. For example:

$S \Rightarrow_1 ((X_\text{n} p \Delta_\text{n}((X_{\text{v}2}\, o_2 X)\ o_1(X_\text{n} \Delta_\text{n} X_q)\Delta_\text{n} X))\Delta X_\text{n})$
$S \Rightarrow_1$((author.*d* p_p((give $o_2$we) o_1(book a_n new) a_n he)) *in*.meet.*d*)
In the meeting, the author gave us his new book.

where $\Rightarrow_1$ is the deducibility on the syntactic level; $\Rightarrow_2$ is the deducibility on the semantic level.

Knowledge representation. The recognizable subject (as a noun phrase denoting “the something or someone that the sentence is about”) can be served as a basis for integration of input sentences into their common knowledge representation. In the text processing system, input text sentences can be transformed into set-theoretical form, and then the resulting formal sentence structures are selected and united into growing knowledge representations. The integration of the sentences that have one and the same subject (a noun phrase contained in user’s request) is considered as a subject knowledge representation. And any collection of the subject knowledge representations produced in the knowledge formation process can be considered as a user-oriented knowledge base [4].

So all sentences that contain one and the same subject N_0, indicated in the user’s request, can be integrated into a user-oriented *subject knowledge representation* $\sigma(N_\text{o})$, i.e.,

$$\sigma(N_\text{o}) = \{S_\text{i}\,|N_\text{i} \supseteq N_0\},$$

where $S_\text{i} \supseteq N_\text{i}$ ($i \geq 1$) is a sentence of the incoming text. Then any collection of the subject knowledge representations $\{\sigma(N_\text{o})_1, \sigma(N_\text{o})_2, \ldots\}$, produced in the formation process, is a *field of subject knowledge representation*, i. e., a user-oriented knowledge base. The integration of input sentences into a common knowledge representation is based on the use of set-theoretical relations and operations. That is, the knowledge forming process can be realized by the following integration rules:

$$(1)\ S_0 \times S_i \rightarrow S_i,\ \text{if}\ S_0 \subseteq S_i;$$
$$(2)\ S_i \times S_j \rightarrow S_i \cup S_j,\ \text{if}\ N_i = N_j,$$

where × is the integration sign; S_0 is a user's request; S_i and S_j are any sentences obtained as a result of use of (1).

Let us suppose that the user's request is S_0={man, man *p*.read}. Then, as a result of the use of (1), we can obtain a setof sentences like

S_1 = {man, man *p*.read, read *o*.book}
The man reads a book;
S_2 = {man, man *p*.read, read *o*.book, read *in*.library}
The man reads a book in the library;
S_3 = {man, man *a*.young, man *p.read*, read *in*.park}
The young man reads in the park,

And as a result of the use of (2), we get the subject knowledge representation:

σ(man)={man, man *p*.read, read *o*.book, read *in*.library; man *a*.young, man *p.read*, read *in*.park}
The man reads a book in the library; the young man reads in the park.

Thus, it can be obtained an on-line (user-oriented) knowledge base. The forming of the knowledge base is a two-stage process: in the first stage (data search), the usual information retrieval is realized (by drawing text information from all sources that contain the user's request); in the second stage (knowledge extraction), the obtained result is processed to extract the more specific information.

REFERENCES

1. *Yakovishin, V. S.* Algebraic representation of syntagmatic structures [Electronic resource] / V. S. Yakovishin // Web journal of formal, computational & cognitive linguistics, Issue 11. – Mode of access : fccl.ksu.ru/issue11. – Date of access : 12.04.2016.

2. *Yakovishin, V. S.* An algebraic basis of formal language description / V. S. Yakovishin // Problems of modern applied linguistics. – Minsk : MSLU, 2014. – P. 59–64.

3. *Yakovishin, V. S.* Formation of object descriptions on the basis of discrete syntagmatic representation / V. S. Yakovishin // Interactive systems and technologies. The problem of human-computer interaction : Proceedings of the 4th International conference, Ulyanovsk, 26–30 Sept. 2005. – Ulyanovsk : UISTU, 2005. – P. 16–18.

4. *Shibut, M. S.* Selection and Aggregation of Sentences in the Knowledge Formation Process / M. S. Shibut, V. S. Yakovishin // Proceedings of the 6th International Conference on Intelligent Data Acquisition and Advanced Computing Systems (IEEE'11). – Prague, Czech Republic, 2011. – Vol. 2. – P. 647–650.