The prospects revealed by the theory of K-representations for bioinformatics and Semantic Web

Vladimir A. Fomichov

Department of Innovations and Business in the Sphere of Informational Technologies
Faculty of Business Informatics, National Research University "Higher School of Economics"
Kirpichnaya str. 33, 105679 Moscow, Russia
vfomichov@hse.ru and vfomichov@gmail.com

Résumé

L'article décrit la structure et les applications possibles de la théorie des K-représentations (représentation des connaissances) dans la bioinformatique afin de développer un Réseau Sémantique d'une génération nouvelle. La théorie des K-répresentations est une théorie originale du développement des analyseurs sémantico–syntactiques avec l'utilisation large des moyens formels pour décrire les données d'entrée, intermédiaires et de sortie. Cette théorie est décrit dans la monographie de V. Fomichov (Springer, 2010). La première partie de la théorie est un modèle formel d'un système qui est composé de dix opérations sur les structures conceptuelles. Ce modèle définit une classe nouvelle des langages formels – la classe des SK-langages. Les possibilités larges de construire des répresentations sémantiques des discours compliqués en rapport à la biologie sont manifestes. Une approche formelle nouvelle de l'élaboration des analysateurs multilinguistiques sémantico-syntactiques est décrite. Cet approche a été implémentée sous la forme d'un programme en langage PYTHON.

Abstract

The paper describes the structure and possible applications of the theory of K-representations (knowledge representations) in bioinformatics and in the development of a Semantic Web of a new generation. It is an original theory of designing semantic-syntactic analyzers of natural language (NL) texts with the broad use of formal means for representing input, intermediary, and output data. The current version of the theory is set forth in a monograph by V. Fomichov (Springer, 2010). The first part of the theory is a formal model describing a system consisting of ten operations on conceptual structures. This model defines a new class of formal languages – the class of SK-languages. The broad possibilities of constructing semantic representations of complex discourses pertaining to biology are shown. A new formal approach to developing multilingual algorithms of semantic-syntactic analysis of NL-texts is outlined. This approach is realized by means of a program in the language PYTHON.

Mots-clés : dialogue homme-machine en langage naturel, algorithme de l'analyse sémantico-syntactique, sémantique intégrale formelle, théorie des K-représentations, SK-langues, représentation sémantique, bases de données linguistiques, réseau sémantique d'une génération nouvelle, réseau sémantique multilingue, bioinformatique

Keywords: man-machine natural language dialogue, algorithm of semantic-syntactic analysis, integral formal semantics, theory of K-representations, SK-languages, semantic representation, text meaning representation, linguistic database, Semantic Web of a new generation, multilingual Semantic Web, bioinformatics

1 Introduction

Many years ago, before the birth of the World Wide Web and bioinformatics, the author of this paper came across the following idea : the progress of computers and informational technologies will reach the point when the continuation of this progress will require the applied computer systems with the well-developed abilities of processing natural language (NL): written texts and spoken speech. The analysis carried out at that time distinguished three significant problems : (a) NL-interaction with applied intelligent systems (AIS) ; (b) the construction of question-answering systems dealing with free texts ; (c) automatic extraction of information from NL-texts for updating knowledge bases of AIS (Fomitchov, 1984; Fomichov, 1992, 1993a).

An important precondition of solving these problems seemed to be the formal means allowing for representing the structured meanings (SMs), or semantic structures, of arbitrary NL-texts pertaining to economy, technology, medicine, and other fields of human professional activity. This idea underlay the birth of Integral Formal Semantics (IFS) of NL – an original branch of mathematical and computational linguistics (see, in particular, (Fomichov, 1992 - 1994) and Chapter 2 of (Fomichov, 2010a)).

The beginning of the XXIst century appears to be just the time point when the progress of computers and the Internet demands powerful and flexible technologies of NL processing for applying them in numerous thematic domains. One has been able to observe in the 2000s in different parts of the world the permanent growth of interest in designing NL-interfaces to applied intelligent systems and other kinds of natural language processing systems (NLPS), or linguistic processors. In particular, a number of projects being useful for practice is described in the publications (Cimiano et al., 2007; Duke, Glover, Davies, 2007; Frank et al., 2007; Popescu, Etzioni, Kautz, 2003).

On the one hand, the first version of a NLPS usually is not an ideal one. Additional work is necessary in order to expand the input language of the system (being a sublanguage of NL) and enhance the intelligent capabilities of the system. On the other hand, when a useful for practice system has been designed, the proposals to adapt this system to the utilisation in a different domain may be received. That is why it is important to have a collection of formal tools enabling the designers to fix the assumptions about semantic structures, linguistic databases, input and output data structures, and about intermediate data structures being outputs of some subsystems and the inputs of other subsystems.

One of the most acute and large-scale problems is to endow the existing Web with the ability of extracting information from numerous sources in various natural languages (of cross-language information retrieval) and of constructing NL-interfaces to a number of knowledge repositories recently developed under the framework of the Semantic Web project (Wilks, Brewster, 2006 ; Fomichov, 2005, 2009a – 2010b).

As far as in the middle of the 1960s, the researchers had practically the only formal approach to describing structured meanings (SMs) of NL-texts : the first-order predicates logic (FOPL). Due to numerous restrictions of FOPL, the search for more powerful and flexible formal means for decribing SMs of NL-texts was started in the second half of the 1960s. As a result, a number of new theories have been developed, first of all, the Theory of Generalized Quantifiers (TGQ), Discourse Representation Theory (DRT), Theory of Semantic Nets (TSN), Theory of Conceptual Graphs (TCG), Episodic Logic (EL), and Theory of K-representations. The latter theory being now the central component of IFS is an original theory of designing semantic-syntactic analyzers of NL-texts with the broad use of formal means for representing input, intermediary, and output data.

In order to understand the principal distinction of the theory of K-representations from other mentioned approaches to formalizing semantics of NL, let's consider an analogy. Bionics studies the peculiarities of the structure and functionning of the living beings in order to discover the new ways of solving certain technical problems. Such theories as TGQ, DRT, TSN, TCG, EL and several other theories were elaborated on the way of expanding the expressive mechanisms of FOPL. To the contrary, the theory of K-represenations was developed as a consequence of analysing the basic expressive mechanisms of NL and putting forward a conjecture about a system of partial operations on conceptual structures underpinning these expressive mechanisms. Of course, the idea was to develop a formal model of this system being compatible with FOPL.

This paper aims at attracting the attention of the researchers to new prospects revealed by the theory of K-representations for the design of semantics-oriented NLPSs (first of all, in the field of bioinformatics) and for developing a Semantic Web of a new generation (SW-2), where its principal distinguished feature will be the well-developed mechanisms for conceptual processing of texts and spoken speech in many natural languages. So

SW-2 can be also called a Meanings Understanding Web or a Multilingual Semantic Web (Fomichov, 2009a – 2010b).

The first subject of this paper is the demonstration (on the examples of complex biological discourse and definition) of some precious features of a mathematical model introduced in (Fomichov, 2010a) and describing a system of 10 partial operations on conceptual structures for building semantic representations (in other terms, text – meaning representations) of, most likely, arbitrary sentences and discourses in French, English, German, Russian, and other natural languages (texts pertaining to arbitrary spheres of professional activity). This model is the kernel of the theory of K-representations.

The second subject is the arguments in favour of employing the theory of K-representations as a foundation of an original strategy of transforming the existing Web into a Semantic Web of a new generation.

The third subject of this paper is the description of a new method of designing multilingual semantics-oriented NLPS with the help of formal means for representing intermediary, output, and a part of input data. A multilingual algorithm of semantic-syntactic analysis of NL-texts called *SemSynt1* and introduced in the second part of the monograph (Fomichov, 2010a) was developed in accordance with this new approach and was implemented by means of the language of Web programming PYTHON.

2 Two basic principles of designing linguistic processors

Most often, semantics-oriented natural language processing systems, or linguistic processors, are complex computer systems, their design requires a considerable time, and its cost is rather high. That is why usually, as it was mentioned above, it is necessary to elaborate a series of NLPSs, step by step expanding the input sublanguage of NL and satisfying the requirements of the end users. On the other hand, the same regularities of NL are manifested in the texts pertaining to various thematic domains.

That is why, in order to diminish the total expenses of designing a family of NLPSs by one research centre or group during a certain several-year time interval and in order to minimize the duration of designing each particular system from this family of NLPSs, it seems to be reasonable to pay more attention to: (a) the search for best typical design solutions concerning the key subsystems of NLPSs with the aim to use these solutions in different domains of employing NLPSs; (b) the elaboration of formal means for describing the main data structures and principal procedures of algorithms implemented in semantic-syntactic analyzers of NL-texts or in the synthesizers of NL-texts.

That is why it appears that the adherence to the following two principles in the design of semantics-oriented NLPSs by one research centre or a group will contribute, in the middle-term perspective, to reducing the total cost of designing a family of NLPS and to minimizing the duration of constructing each particular system from this family:

the ***Principle of Stability*** of the used language of semantic representations in the context of various tasks, various domains and various software environments (stability is understood as the employment of a unified collection of rules for building the semantic structures as well as domain- and task-specific variable set of primitive informational units);

the ***Principle of Succession*** of the algorithms of NLPS based on using one or more compatible formal models of a linguistic database and unified formal means for representing the intermediate and final results of semantic-syntactic analysis of natural-language texts in the context of various tasks, various domains and various software environments (the succession means that the algorithms implemented in basic subsystems of NLPS are repeatedly used by different linguistic processors).

3 Formalization of basic assumptions about primary items of conceptual level

The monograph (Fomichov, 2010a) sets forth a current version of the theory of K-representations (knowledge representations). It is an original theory of designing multilingual semantic-syntactic analyzers of NL-texts (sentences and discourses) with the broad use of formal means for representing input, intermediary, and output data. Let’s start to consider the structure of this theory.

The *first basic constituent* of the theory of K-representations is the theory of SK-languages (standard knowledge languages). The kernel of the theory of SK-languages is a mathematical model describing a system of such 10 partial operations on structured meanings (SMs) of natural language texts (NL-texts) that, using primitive conceptual items as "blocks", we are able to build SMs of arbitrary NL-texts (including articles, textbooks, etc.) and arbitrary pieces of knowledge about the world. The analysis of the scientific literature on artificial intelligence theory, mathematical and computational linguistics shows that today the class of SK-languages opens the broadest prospects for building semantic representations (SRs) of NL-texts (i.e., for representing meanings of NL-texts in a formal way).

The first part of the theory of SK-languages is a mathematical model describing a system of primary conceptual units used by an applied intelligent system, in particular, by a NL processing system. This model defines (with the help of a rather long sequence of auxiliary steps) a new class of formal objects called *conceptual bases (c.b.),* where each concrete c.b. is constructed for a certain group of application domains. Each c.b. *B* is equivalent to a system of the form ($c_1, \ldots , c_{15}$) with the components $c_1,\ldots, c_{15}$ being mainly finite or countable sets of symbols and distinguished elements of such sets. In particular, $c_1 = St$ is a finite set of symbols called *sorts* and designating the most general considered notions (concepts); $c_5 = X = X(B)$ is a countable set of strings used as elementary blocks for building knowledge modules and semantic representations (SRs) of texts; *X* is called a primary informational universe; $c_6 = V$ is a countable set of variables; $c_8 = F$ is a subset of *X* whose elements are called functional symbols.

The set of sorts *St* can include, in particular, the elements *spatial.object, physical.object, dynamic.physical.object, intelligent.system, organization, moment, situation, event,* etc. The set of sorts *St* is a subset of the set *X.* For instance, *X* may include the elements *book, ship, firm, 12, green, Height, Weight, Authors, Part-of, Cheeper, printing, uploading*. The elements of the set *V* are used either as the marks of the entities of various kinds or jointly with the universal and existential quantifiers. The set *F* consists of the designations of functions and is a subset of the set *X = X(B).* The set *F* may include, for instance, the elements *Height, Weight, Authors.*

4 About a model of a system consisting of ten operations on conceptual structures

Each c.b. *B* determines three classes of formulas, the first class *Ls(B)* being considered as the principal one and being called *the SK-language (standard knowledge language) in the basis B*. Its strings (they are called K-strings) are convenient for building SRs of NL-texts. We'll consider below only the formulas from the first class *Ls(B).* If *Expr* is an expression in natural language and a K-string *Semrepr* can be interpreted as a semantic representation of *Expr,* then *Semrepr* will be called a K-representation (KR) of the expression *Expr*.

For determining for arbitrary c.b. *B* three classes of formulas, a collection of inference rules P[0], P[1], ... , P[10] is defined. The rule P[0] provides an initial stock of formulas from the first class. E.g., there is such c.b. B_1 that, according to P[0], $Ls(B_1)$ includes the elements *car1, green, city1, fin-set, India, 14, 14/cm, all, any, Height, Distance, Staff, Suppliers, Quantity, x1, x5.*

For arbitrary c.b. *B*, let *Degr(B)* be the union of all Cartesian m-degrees of *Ls(B),* where $m \geq 1$. Then the meaning of the rules of constructing well-formed formulas P[1], ..., P[10] can be explained as follows: for each k from 1 to 10, the rule P[k] determines a partial unary operation *Op[k]* on the set *Degr(B)* with the value being an element of *Ls(B).*

Example. There is a conceptual basis B possessing the following properties. The primary informational universe *X = X(B)* includes the conceptual items *prophase, prometaphase, metaphase, nanaphase, telophase* describing five distinct stages of mitosis (the process of somatic cell division, during which the nucleus also divides) and the conceptual items *China, India, Sri_Lanka*. Hence the value of the partial operation *Op[7]* (it governs the use of logical connectives ∧ - AND and ∨ - OR) on the six-tuple

< ∧, *prophase, prometaphase, metaphase, nanaphase, telophase* >

is the string *Semexpr1* of the form

(prophase∧ prometaphase ∧ metaphase ∧ nanaphase ∧ telophase),

and the value on the four-tuple < ∨, *China, India, Sri-Lanka* > is the K-string *(China ∨ India ∨Sri-Lanka).*

Let *X(B)* also include the item *mitosis* and the designation of a binary relation *Stages-relation*. Then the K-string

Stages-relation (mitosis, Semexpr1)

is the result of applying the partial operation P[4] to the operands *Stages-relation, mitosis, and Semexpr1.*

Besides, let *X(B)* include the items *article1* (a paper), *article2* (a manufactured article), and *h1 = article2, h2 = Kind1(certn article2, ceramics), h3 = (Country1(certn article2) ≡ (China ∨ India ∨ Sri-Lanka)), h4 = article2 * (Kind1, ceramics) (Country1, (China ∨ India ∨ Sri_Lanka))* are the elements of *Ls(B).* Then the K-string *h4* is the result of applying the partial operation P[8] to the operands *h1, h2, h3*.

Ls(B) includes the string *h5* of the form *certn h4*, being the result of applying the operation P[1] to the operands *certn* and *h4.* The item *certn* denotes the meaning of the expression "a certain", and the string *h5* is interpreted as a designation of a manufactured article being a kind of ceramics and produced in China, India, or Sri-Lanka.

Let h6 be the string of the form *(Height(h5) ≡14/cm).* Then *h6* belongs to *Ls(B)* and is the result of applying the partial operation P[3] to the operands *Height(h5)* and *14/cm.*Thus, the essence of the basic model of the theory of SK-languages is as follows: this model determines a partial algebra of the form (*Degr(B), Operations(B)*) , where *Degr(B)* is the carrier of the partial algebra, *Operations(B)* is the set consisting of the partial unary operations *Op[1], ..., Op[10]* on *Degr(B).*

The volume of the complete description in (Fomichov, 2010a) of the mathematical model introducing, in essence, the operations *Op[1], ..., Op[10]* on *Degr(B)* and, as a consequence, determining the class of SK-languages considerably exceeds the volume of this paper. That is why, due to objective reasons, this model can't be included in this paper.

5 Building Semantic Representations of Complex Biomedical Discourses

The theoretical results stated in chapters 1 - 6 of the monograph (Fomichov, 2010a) provide a framework for following-up the principle of stability of the used language of semantic representations. According to the hypothesis formulated in Chapter 6, the definition of the class of SK-languages enables us to build semantic representations of NL-texts in arbitrary application domains.

During several last years, the significance of natural language processing (NLP) technologies for informatics dealing with the problems of biology and medicine has been broadly recognized. As a consequence, the term BioNLP interpreted as the abbreviation for Natural Language Processing in Biology and Medicine was born. The formalization of natural language semantics is a very acute problem of BioNLP. The attention of many researchers in this field is now attracted by the phenomena of the semantics of sentences and discourses (Prince, Roche, 2009). That is why let's illustrate the new expressive possibilities provided by SK-languages on the example of building a semantic representation of a rather complex discourse pertaining to genetics.

It is known that each individual possesses two genes being responsible for a particular characteristic (e.g., the height) in case of almost all characteristics (or traits). The genes responsible for the contrasting values of a characteristic (for instance, the values "tall" and "short" for the trait "height") are referred to as *allelomorphs,* or *alleles* for short. Some genes have more than two allelic forms, i.e. multiple alleles. In the case of the ABO blood group system, there are at least four alleles (A_1, A_2, B and O). An individual can possess any two of these alleles, which can be the same or different (AO, A_2B, OO, and so on).

With respect to this context, let's consider the discourse D1 = "Alleles are carried on homological chromosomes and therefore a person transmits only one allele for a certain trait to any particular offspring. For example, if a person has the genotype AB, he will transmit to any particular offspring either the A allele or the B allele, but never both or neither" (Turnpenny, Ellard, 2005, p. 198).

Let S1 = "Alleles are carried on homological chromosomes", S2 = "therefore a person transmits only one allele for a certain trait to any particular offspring.", S3 = S1 and S2, S4 = "For example, if a person has the genotype AB, he will transmit to any particular offspring either the A allele or the B allele, but never both or neither".

First of all, we'll construct a possible K-representation (KR) of the sentence S1 as the following string *Semrepr1*:

*(Entails((Alleles-relation (certn gene * (Part, certn person : y1) : x1, certn gene * (Part, y1) : x2) ∧ Location(x1, x3) ∧ Location(x2, x4) ∧ Semantic-descr((x3 ∧ x4), chromosome * (Part, y1))), Homologous(x3, x4)) : P1 ∧ Correspondent-situation(P1, e1)).*

The K-string *Semrepr1* illustrates the following new properties of the theory of SK-languages: the possibilities (a) to construct the compound designations of the notions and of the objects qualified by these notions, (b) to use the logical connective ∧ (AND) for joining not only the semantic representations of the statements but also the designations of the objects, as in case of the substring *(x3 ∧ x4),* (c) to associate the mark of a situation with the mark of the meaning of sentence describing this situation, as in case of the substring *Correspondent-situation(P1, e1).*

As for the sentence S2, its possible KR will be the string *Semrepr2* of the form

(Cause(e1, e2) ∧ Correspondent-situation(P2, e2) ∧ (P2 ≡ ∀ y2(person) ∀ y3(person * (Offspring-rel, y2)) ∀ x5(trait1 * (Possessed-by, y2)) ∃ x6 (gene * (Element, Alleles-function(x5))) Situation(e3, transmission1 * (Source1, y2)(Recipient1, y3)(Object-transmitted, x6)) ∧ ¬ ∃ x7 (gene * (Element, Alleles-function(x5))) (Situation(e4, transmission1 * (Source1, y2)(Recipient1, y3)(Object-transmitted, x7)) ∧ ¬ (x7 ≡ x6)))))).

The symbols ∀ and ∃ in the K-string *Semrepr2* are the universal quantifier and of the existential quantifier. We can see here that SK-languages allow for restricting the domain of a logical quantifier with the help of the expressions like _(person * (Offspring-rel, y2)), (trait1 * (Possessed-by, y2)), (gene * (Element, Alleles-function(x5)))_, and so on.

At this point of our analysis we have the appropriate building blocks *Semrepr1* and *Semrepr2* for constructing a possible KR of the sentence S3 as the string *Semrepr3* of the form

(Semrepr1∧ Semrepr2) : P3.

Now let's build a K-representation of the final sentence S4 in the context of the sentence S3. We see that the word combination "For example" from S4 encodes the reference to the meaning of the sentence S3. The system of ten partial operations on conceptual structures proposed by the theory of K-representations contains the operation Op[5] to be used just in such cases. This operation allows for constructing the formulas of the kind *form : var*, where the first operand *form* is a semantic description of an object (in particular, a SR of a statement), and *var* is a variable.

This operation was used for constructing the subformulas _certn gene * (Part, certn person : y1) : x1_ and _certn gene * (Part, y1) : x2_ of the formula *Semrepr1*; besides, for building the formula *Semrepr 3* from the operands *(Semrepr1∧ Semrepr2)* and *P3*.

Now we can use the variable *P3* as a mark of the meaning of the sentence S3 in the following K-representation *Semrepr4* of the sentence S4:

Example(P3, Entails(Situation(e4, posessing1 * (Owner1, arbitr person : y4)(Object1, certn genotype * (Designation, 'AB') : x7), Situation(e5, transmission1 * (Source1, y4)(Recipient1, arbitr person * (Offspring, y4) : y5)(Object-transmitted, (certn allele * (Designation, 'A") : x8 ∨ certn allele * (Designation, 'B") : x9)) ∧ Situation(e6, ¬ transmission1 * (Source1, y4)(Recipient1, y5))(Object-transmitted, (x8 ∧ x9))) ∧ Situation(e7, ¬ transmission1 * (Source1, y4)(Recipient1, y5))(Object-transmitted, NIL))))).

Here *NIL* is the constant reflecting the meaning of the word "nothing".

Actually, we build a K-representation of the discourse D1 as a string of the form *((Semrepr1 ∧ Semrepr2) : P3 ∧ Semrepr4).*

To sum up, SK-languages allow for describing semantic structure of the sentences with direct and indirect speech and of the discourses with the references to the meanings of phrases and larger parts of a discourse, for constructing compound designations of the notions, sets, and sequences. As far as one can judge on the available scientific literature, now only the theory of K-representations explains the regularities of structured meanings of, likely, arbitrary sentences and discourses pertaining to biomedicine and other fields of professional activity.

6 K-representations of complex biomedical definitions of notions

The analysis shows that the SK-languages possess a number of interrelated expressive mechanisms making them a convenient formal tool for building arbitrarily complex definitions of notions.

Example. Let T1 = "A flock is a large number of birds or mammals (e.g. sheep or goats), usually gathered together for a definite purpose, such as feeding, migration, or defence". T1 may have the K-representation *Expr1* of the form

*Definition1 (flock, dynamic-group * (Qualitative-composition, (bird $\vee$ mammal * (Examples, (sheep $\wedge$ goal)))), S1, (Estimation1(Quantity(S1), high) $\wedge$ Goal-of-forming (S1, certain purpose * (Examples, (feeding $\vee$ migration $\vee$ defence))))).*

The analysis of this formula enables us to conclude that it is convenient to use for constructing semantic representations (SRs) of NL-texts: (1) the designation of a 5-ary relationship *Definition1*, (2) compound designations of concepts (in this example the expressions *mammal * (Examples, (sheep $\wedge$ goal))* and *dynamic-group * (Qualitative-composition, (bird $\vee$ mammal * (Examples, (sheep $\wedge$ goal))))* were used), (3) the names of functions with the arguments and/or values being sets (in the example, the name of an unary function *Quantity* was used, its value is the quantity of elements in the set being an argument of this function), (4) compound designations of intentions, goals; in this example it is the expression *certain purpose * (Examples, (feeding $\vee$ migration $\vee$ defence))* . The structure of the constructed K-representation *Expr1* to a considerable extent reflects the structure of the definition T1.

7 Related approaches to describing semantic structure of NL-texts

The advantages of the theory of SK-languages in comparison with first-order predicates logic, Discourse Representation Theory (DRT) and Episodic Logic (EL) are, in particular, the possibilities: (1) to distinguish in a formal way objects (physical things, events, etc.) and concepts qualifying these objects; (2) to build compound representations of concepts; (3) to distinguish in a formal manner objects and sets of objects, concepts and sets of concepts; (4) to build complex representations of sets, sets of sets, etc.; (5) to describe set-theoretical relationships; (6) to effectively describe structured meanings (SMs) of discourses with references to the meanings of phrases and larger parts of discourses; (7) to describe SMs of sentences with the words "concept", "notion"; (8) to describe SMs of sentences where the logical connective "and" or "or" joins not the expressions-assertions but designations of things, sets, or concepts; (9) to build complex designations of objects and sets; (10) to consider non-traditional functions with arguments or/and values being sets of objects, of concepts, of texts' semantic representations, etc.; (11) to construct formal analogues of the meanings of infinitives with dependent words and, as a consequence, to represent proposals, goals, obligations, commitments.

The items (3) - (8), (10), (11) in the list above indicate the principal advantages of the theory of SK-languages in comparison with the Theory of Conceptual Graphs (TCG). Besides, the expressive possibilities of the new theory are much higher than the possibilities of TCG as concerns the items (1), (2), (9).

The global advantage of the theory of K-representations is that it puts forward a hypothesis about a system of partial operations on conceptual structures being sufficient and convenient for constructing semantic representations (or text meaning representations) of sentences and discourses in NL pertaining to arbitrary fields of humans' professional activity.

8 A strategy of developing a Semantic Web of a new generation

It seems that the Principle of Stability of the used language of semantic representations has much broader sphere of application than the professional activity of any concrete research group or research centre dealing with NLP. There are reasons to believe that following-up this principle can considerably speed-up the progress of the studies bridging a gap between the Semantic Web and NLP. The process of endowing the existing Web with the ability of understanding many natural languages is an objective ongoing process (Wilks, Brewster, 2006). It is a decentralized process, because the research centres in different countries mainly independently develop the translators from particular natural languages to semantic representations (or text meaning representations) and the applied computer systems extracting the meanings from texts in particular natural languages or producing summaries of the collections of texts in particular languages.

The analysis has shown that there is a way to increase the total successfulness, effectiveness of this global decentralized process. In particular, it would be important with respect to the need of cross-language conceptual information retrieval and question - answering. The proposed way is a possible new paradigm for the mainly decentralized process of endowing the existing Web with the ability of processing many natural languages.

The principal idea of a new paradigm is as follows. There is *a common thing* for the various texts in different natural languages. This common thing is the fact that *the NL-texts have the meanings.* The meanings are associated not only with NL-texts but also with the visual images (stored in multimedia databases) and with the pieces of knowledge from the ontologies.

That is why the great advantages are promised by the realization of the situation when a unified formal environment is being used in different projects throughout the world for reflecting structured meanings of the texts in various natural languages, for representing knowledge about application domains, for constructing semantic annotations of informational sources and for building high-level conceptual descriptions of visual images.

The analysis of the expressive power of SK-languages (see the chapters 3 – 6 of (Fomichov, 2010a)) shows that the SK-languages can be used as a unified formal environment of the kind. It is a direct consequence of the following hypothesis put forward by the author in (Fomichov, 2005, 2007, 2010a, 2010b): SK-languages are a convenient tool of building semantic representations of arbitrarily complex NL-texts (sentences and discourses) pertaining to arbitrary field of professional activity.

This central idea underlies an original strategy of transforming step by step the existing Web into a Semantic Web of a new generation, where its principal distinguished feature would be the well-developed ability of NL processing; it can be also qualified as a Meanings Understanding Web or as a Multilingual Semantic Web. The versions of this strategy are published in (Fomichov, 2009b – 2010b).

9 A new method of designing multilingual semantics-oriented natural language processing systems

The theory of K-representations proposes a collection of formal tools being useful for the design of arbitrarily complex NLPS. Let’s consider the basic steps of a new method of designing multilingual semantics-oriented NLPS with the help of formal means for representing intermediary, output, and a part of input data. A multilingual algorithm of semantic-syntactic analysis of NL-texts called *SemSynt1* and introduced in the second part of the monograph (Fomichov, 2010a) was developed in accordance with this new approach and was implemented by means of the language of Web programming PYTHON. The rationale for using PYTHON can be found in (Bird et al., 2009). An explicit description of this approach is given below for the first time.

9.1 Step 1: formalization of additional assumptions about primary items of conceptual level

The content of Step 1 is to introduce additional assumptions about some components of the considered conceptual basis *B*. For instance, NL-texts often include the compound designations of the sets. Hence it would be reasonable to introduce the following assumptions: (a) the component *X = X(B)* includes a subset *Nat* consisting of all strings of the form *d[1], ..., d[n],* where $n \geq 1$, for *k = 1,..., n, d[k]* is a digit from the set {‘0’, ‘1’, ‘2’, ..., ‘9’} ; (b) the subset *F(B)* includes the element *Quantity* interpreted as the name of the function «Quantity of the elements of a set » ; (c) the set X(B) includes the elements *Quality-composition* and *Thing-composition* in order to construct, for example, the formulas *Quality-composition(S3, container1 * (Weight, 3/tonna))* and *Thing-composition(S4, (c1* $\wedge$ *c2* $\wedge$ *c3)),* where *c1, c2, c3* are the marks of the concrete containers with ceramics from Indi*a.*

Chapter 5 of the monograph (Fomichov, 2010a) can be used as a good introduction to the ways of fixing the additional assumptions about the used system of primary conceptual items.

9.2 Step 2: selecting the form of text meaning representations

The expressive power of the class of SK-languages is very high. SK-languages enable us to build semantic representations of natural language texts in arbitrary application domains. That is why it is necessary to select such collection of the expressive mechanisms of SK-languages that it is useful and convenient to employ these expressive mechanisms for constructing semantic (or text meaning) representations of the input NL-texts. It is the content of the Step 2 of the proposed method of designing multilingual, semantics-oriented NLPS.

Let's consider the examples illustrating the correspondence between the sentences in English, Russian (in Latin transcription), and German and their semantic representations (SR) being the expressions of a certain SK-language, that is, being the K-representations of the input texts. In these examples, the SR of the input text T will be the value of the string variable *Semrepr* (Semantic representation). The considered examples illustrate the correspondence between the inputs and outputs of the developed algorithm *SemSynt1* , see Chapters 9 and 10 of (Fomichov, 2010a).

Example 1. Let T1eng = "Find a description of the programming language PYTHON on the Web-site http://docs.python.org", T1rus = “Naydite opisanie yazyka programmirovaniya PYTHON na veb-sayte http://docs.python.org”, T1germ = “Finden eine Beschreibung der Programmiersprache PYTHON auf dem Site http://docs.python.org”. Then

Semrepr = (Command (#Operator#, #Executor#, #now#, e1) ∧ Target (e1, finding1 * (Object-file, certn file1 * (Inf-content, certn description1t * (Focus-object, certn progr-lang * (Name1, “PYTHON”) : x3) : x2))(Web-source, http://docs.python.org))).

Example 2. Let T2eng = "The international scientific conference “DEXA-2009” took place in Linz, Austria, during August 31 – September 4, 2009”, T2rus = “Mezhdunarodnaya nauchnaya konferentsiya “DEXA-2009” prokhodila v gorode Linz, Avstriya s 31 avgusta po 4 sentyabrya 2009 goda”, T2germ = “Die internationale wissenschaftliche Konferenz “DEXA-2009” war in Linz, Oesterreich waehrend 31. August – 4. September 2009 stattgefunden”. Suppose that the used basic semantic items are constructed with respect to the spelling of English expressions corresponding to these items. For instance, the English words “city” and “town”, the Russian word “gorod”, and the German word group "die Stadt” will be associated with the semantic item *city1.* From the formal standpoint, it means that the elements of the used conceptual basis are built on the basis of English expressions. If this condition is satisfied, the algorithm builds the K-representation

Semrepr = Situation(e1, taking-place * (Event1, certn conference1 * (Kind-geogr, international) (Kind-focus, science) : x1)(Place1, certn city1 * (Name1, “Linz”)(Belongs-to-country, certn country1 * (Name1, “Austria”) : x3) : x2) (Time-interval, <31.08.2009, 04.09.2009>)) .

Example 3. Let T3eng = "Did the international scientific conference "DEXA" take place in Hungary?", T3rus = “Prokhodila li mezhdunarodnaya nauchnaya konferentsiya “DEXA” v Vengrii?”, T3germ = “War die internationale wissenschaftliche Konferenz “DEXA” in Ungarn stattgefunden?”. Then

Semrepr = Question (x1, (x1 ≡ Truth-value (Situation (e1, taking_place * (Time, certn moment * (Earlier, #now#) : t1) (Event1, certn conference * (Type1, international) (Type2, scientific) (Name1, “DEXA”) : x2) (Place, certn country1 * (Name1, “Hungary”) : x3))))).

Example 4. Let T4eng = "What English scientist discovered penicillin?", T3rus = “Kakoy angliysky uchony otkryl penicillin?”, T3germ = “Welcher English Wissenschaftler hat Penizillin entdeckt?”. Then

Semrepr = Question (x1, Situation (e1, discovering1 * (Time, certn moment * (Earlier, #now#) : t1) (Agent1, certn scientist * (Country1, England) : x1) (New-object, certn medicine1 * (Name1, “penicillin”) : x2))).

Example 5. Let T5eng = "What European companies the firm “Rainbow” is cooperating with?", T5rus = “S kakimi evropeyskimi kompaniyami sotrudnichaet firma “Rainbow”, T5germ = “Mit welchen europaeischen Kompanien die Firma “Rainbow” kooperiert?”. Then

Semrepr = Question (S1, (Qualitative-composition (S1, company1 * (Location, Europe)) ∧ Description(arbitrary company1* (Element, S1) : y1, Situation (e1, cooperation * (Time, #now#) (Agent2, certn company1 * (Name1, “Rainbow) : x1) (Cooper-partner, y1))))).

Example 6. Let T6 = "Who produces the medicine "Zinnat"?". Then

Semrepr = Question (x1, Situation (e1, production1 * (Time, #now#) (Agent2, x1) (Product2, certn medicine1 * (Name1, “Zinnat”) : x2))).

Example 7. Let T7eng = "When and where did Dr. Erik Stein arrive to Zuerich from?", T7rus = “Kogda i otkuda doktor Erik Stein priekhal v Zurikh?”, T7germ = “Wann und woher hat Dr. Erik Stein nach Zuerich gekommen?”. Then

Semrepr = Question ((x4 ∧ x1), (Situation (e1, arrival *(Time, certn moment * (Earlier, #now#) : t1)
(Start-location, x1)(Agent1, certn person *(Qualif, Ph.D.)(Name, “Erik”)(Surname, “Stein”) : x2) (Final-location, certn city1 * (Name1, “Zuerich”) : x3)) ∧ (x4 ≡ t1)).

Example 8. Let T8eng = "How many countries did participate in the Olympic Games - 2008?", T7rus = “Skolko stran uchastvovalo v Olimpiyskikh Egrakh – 2008”, T7germ = “Wieviel Laender haben an den Olympischen Spielen – 2008 teilgenommen?”. Then

Semrepr = Question (x1, ((x1 ≡ Numb(S1)) ∧ Qualitative-composition (S1, country1) ∧
*Description (certn country1 * (Element, S1) : y1, Situation (e1, participation1 **
*(Time, certn moment * (Earlier, #now#) : t1) (Agent1, y1)*
(Time, 2008/year)(Event1, certn olymp-game : x2)))).

Example 9. Let T9eng = "How many times did Professor Bill Jones visit France?”, T7rus = “Skolko raz professor Bill Jones posetil Frantsiu”, T7germ = “Wieviel Mal hat Herr Professor Bill Jones Frankreich besucht?”. Then

Semrepr = Question (x1, ((x1 ≡ Numb (S1)) ∧ Qualitative-composition (S1, sit) ∧
*Description (arbitrary sit * (Element, S1) : e1, Situation (e1, visiting * (Time, certn moment **
*(Earlier, #now#) : t1) (Agent1, certn person * (Qualif, professor)(Name, “Bill”)(Surname, “Jones”) : x2)*
*(Place2, certn country * (Name1, “France”) : x3))))).*

9.3 Step 3 of the new approach: formation of semantic-syntactic components of a linguistic database

Chapter 7 of the monograph (Fomichov, 2010a) contains an original, broadly applicable mathematical model of a linguistic database (LDB). This model formalizes the structure of a linguistic database allowing for setting up various conceptual relations, e.g. 'Verb + Preposition + Noun', 'Verb + Noun', 'Noun1 + Preposition + Noun2', 'Numeral + Noun', 'Adjective + Noun', 'Noun1 + Noun2', 'Participle + Noun', 'Participle + Preposition + Noun', 'Interrogative pronoun + Verb', 'Preposition + Interrogative pronoun + Verb', 'Interrogative Adverb + Verb', 'Verb + Numerical Value Representation' (a number representation + a unit of measurement representation). The model defines a class of formal objects called *linguistic bases* (l.b.). Each l.b. *Lingb* is a mathematical representation of a morphological datatabase, of some functions corresponding to the subsystems of a morphological analyzer, and of semantic-syntactic components of the LDB.

The content of the considered Step 3 is to form several semantic-syntactic components of a LDB. The first component *Lsdic* is the set of finite sequences of the form

(i, lec, pt, sem, st[1], ..., st[k], comment),

where $i \geq 1$ is the ordered number of the k+5-tuple (we need it to organize the loops in the algorithms of processing NL-texts), and the rest of the components are interpreted in the following way: *lec* is an element of the set of basic lexical units *Lecs* for the considered morphological basis; *pt* is a designation of the part of speech for the basic lexical unit *lec*; the component *sem* is a string that denotes one of the possible meanings of the basic lexical unit *lec*.

For instance, the verb "to enter" has, in particular, the following two meanings: (1) entering a learning institution (in the sense "becoming a student of this learning institution"); (2) entering a space object ("Yves has entered the room", etc.). So one system from a possible lexico-semantic dictionary will have, as the beginning, the sequence *i1, enter, verb, entering1,* and the other will have, as the beginning, the sequence *i2, enter, verb, entering2.*

The component *sem* can be a complex string being an expression of the SK-language Ls(B) for a certain conceptual basis.

Example. If *lec* = ‘France’, then *sem* can be the K-string *certn country * (Name1, ‘France’);* if *lec* = ‘green’, then *sem* can be the K-string *Colour (z, green),* where *z* is a variable denoting an entity with the property “green”.

The number *k* is the semantic dimension of the considered sort system, that is, *k* is the maximal number of the different "semantic axes" used to describe one entity in the considered application domain.

Example. Let us consider the concepts "a firm" and "a university". We can distinguish three semantic contexts of word usage associated with these concepts. Firstly, a firm or a university can develop a tool, a technology etc., so the sentences with these words can realize the semantic coordinate "intelligent system". Secondly, we can say: "This firm is situated near the metro station "Taganskaya," and then this phrase realizes the semantic coordinate "spatial object". Finally, the firms and institutes have the directors. We can say, for example: "The director of this firm is Alexander Semenov." This phrase realizes the semantic coordinate "organization". In the considered examples, we'll presume that semantic dimension of the considered sort systems equals is equal to four or three.

The elements *st[1], ..., st[k]* are the different semantic coordinates of the entities characterized by the concept *sem*. For example, if *sem = firm*, then *st[1] = ints* (intelligent system), *st[2] = space.ob* (space object), *st[3] = org* (organization), *k =3*. The compo-nent *comment* is either a natural language description of a meaning associated with the concept *sem* or an empty element *nil.*

The second semantic-syntactic component of a LDB is called the *dictionary of verbal-prepositional frames*, it contains such templates (in other terms, frames) that enable us to represent the necessary conditions of realizing a specific thematic role in the combination "Verbal form + Preposition + Dependent word group". An example in the subsection 9.4.3 illustrates the structure of such templates.

The third semantic-syntactic component of a LDB is called the *dictionary of prepositional frames*, it contains the templates allowing for representing the necessary conditions of realizing a specific relation in the combination of the form "Noun 1 + Preposition + Noun 2" or of the form "Noun 1 + Noun 2".

Example. Let us assume that *Expr* is the expression "an article by Professor Novikov", and a linguistic database includes a template representing the sequence of the form

(*k1, 'by', inf.ob, ints, 1, Authors, "a poem by H. Heine'*),

were *ints* is the sort "intelligent system", 1 is the code of common case in English. We may connect the sorts *ints* and *dyn.phys.ob* (dynamic physical object) with the basic lexical unit "professor". We see that the expression *Expr* is compatible with this template having the number *k1.*

9.4 Step 4: development of an algorithm transforming the input texts into their matrix semantic – syntactic representations

9.4.1 Step 4-1: Building morphological representation of an input text

Skipping mathematical details, we'll suppose that a morphological representation (MR) of a text T with the length *nt* is a two-dimensional array *Rm* with the names of columns *base* and *morph* (more exactly, *morph* is the designation of a group of colums), where the elements of the array rows are interpreted in the following way. Let *nmr* be the number of the rows in the array *Rm* that was constructed for the text T, and *k* be the number of a row from the array Rm, i.e. $1 \leq k \leq nmr$. Then *Rm[k, base]* is the basic lexical unit (the lexeme) corresponding to the word in the position p from the text T. Under the same assumptions, *Rm[k, morph]* is a sequence of the collections of the values of morphological characteristics (or features) corresponding to the word in the position *p*.

Example. Let T1 be the question "Did the management board of the firm "Rainbow" change in May?", and T1germ be the same question in German "Hat der Verwaltungsrat der Firma "Rainbow" in Mai veraendernt sich?". Then the morphological representation Rm1 of T1 consists of the rows *(change, md[1]), (management-board, md[2]), (of, md[3]), (firm, md[4]), (in, md[5]), (May, md[6])*, where *md[1], ..., md[6]* are the sequences of the values of morphological properties associated with the corresponding lexical units from T1. Similarly, the morphological representation Rm2 of T1germ consists of the rows *(sich-veraendern, mdg[1]), (Verwaltungsrat, mdg[2]), (Firma, mdg[3]), (in, mdg[4]), (Mai, mdg[5])*, where *mdg[1], ..., mdg[5]* are the sequences of the values of morphological properties associated with the corresponding lexical units from T1germ.

9.4.2 Step 4-2: Building classifying representation of an input text

Classifying representation. From informal point of view, we will say that a classifying representation (CR) of the text T coordinated with the morphological representation *Rm* of the text T is a two-dimensional array *Rc* with the number of the rows *nt* and the column with the indices *unit, tclass, subclass, mcoord*, in which its elements are interpreted in the following way. Let *k* be the number of any row in the array *Rc* i.e. $1 \leq k \leq nt$. Then *Rc[k, unit]* is one of elementary meaningful units of the text T, i.e. if $T = t_1 \ldots t_{nt}$, then such position *p*, where $1 \leq p \leq nt$, can be found that $Rc[k, unit] = t_p$. If *Rc[k, unit]* is a word, then *Rc[k, tclass], Rc[k, subclass], Rc[k, mcoord]* are correspondingly a part of speech, a subclass of the part of speech, the sequences of the values of morphological properties. If *Rc[k, unit]* is a construct (i.e. a value of a numeric parameter), then *Rc[k, tclass]* is the string *constr*, *Rc[k, subclass]* is the designation of the subclass of informational units corresponding to this construct, *Rc[k, mcoord] = 0*.

Example. Let T1 = "Did the management board of the firm “Rainbow” change in May?". Then a classifying representation *Rc1* of the text T1 coordinated with the morphological representation *Rm1* of T1 may be the following array:

unit	tclass	Subclass	mcoord
did-change	verb	verb-in-indic-mood	1
the management-board	noun	common-noun	2
of	prep	nil	3
the-firm	noun	common-noun	4
“Rainbow”	artif-name	nil	0
in	prep	nil	5
May	noun	proper-noun	6
?	marker	nil	0

If T1germ =“Hat der Verwaltungsrat der Firma “Rainbow” in Mai veraendernt sich?”, then a classifying representation *Rc2* of the text T1germ coordinated with the MR *Rm2* of T1 may have the following form:

unit	tclass	subclass	mcoord
hat-veraendernt-sich	verb	verb-in-indic-mood	1
der-Verwaltungsrat	noun	common-noun	2
der-Firma	noun	common-noun	3
“Rainbow”	artif-name	nil	0
in	prep	nil	4
Mai	noun	proper-noun	5
?	marker	nil	0

9.4.3 Step 4-3: Building the projections of the components of a linguistic basis on the input text

Let *Lingb* be a linguistic basis (see Chapter 7 of (Fomichov, 2010a)), and *Dic* be one of the following components of *Lingb*: the lexico-semantic dictionary *Lsdic*, the dictionary of verbal-prepositional semantic-syntactic frames *Vfr*, the dictionary of prepositional semantic-syntactic frames *Frp* (see Chapter 8 of (Fomichov, 2010a)). Then the projection of the dictionary *Dic* on the input text T is a two-dimensional array whose rows represent all data from *Dic* linked with the lexical units from T .

Let's introduce the following denotations: *Arls* is the projection of the lexico-semantic dictionary *Lsdic* on the input text T; *Arvfr* is the projection of the dictionary of verbal-prepositional frames Vfr on the input text T ; *Arfrp* is the projection of the dictionary of prepositional frames *Frp* on the input text T.

Example. Let T1 = "Did the management board of the firm “Rainbow” change in May?". Then the projection of the lexico-semantic dictionary *Lsdic* on the input text T1 may be the following two-dimensional array:

ord	sem	st1	st2	st3	comment
1	change1	event	nil	nil	Yves has changed 700 franks
1	change2	event	nil	nil	The city has changed very much in the 1990s - 2000s
2	manag-board	org	ints	phys.ob	Management board of a company
4	Company1	org	ints	phys.ob	The firm IBM

5	“Rainbow”	artif-name	nil	nil	nil
7	May	month-value	nil	nil	nil

Here the elements of the column *ord* are the numbers of the corresponding rows of the classifying representation *Rc1*; the sorts *org, ints, phys.ob* are interpreted as the designations of the notions “an organization”, “an intelligent system”, and “a physical object”. The sorts *ints* and *phys.ob* characterize from different standpoints the elements (people) of any firms and management boards of the firms.

The verb “to change” has more than two meanings. That is why for real computer applications this array will be a subarray of the projection of the lexico-semantic dictionary *Lsdic* on the input text T1.

Example. If T1 = "Did the management board of the firm “Rainbow” change in May?", the projection of the dictionary of verbal-prepositional semantic-syntactic frames *Vfr* on the input text T1 *Arvfr1* may include the following subarray *Arvfr1fragm*:

nb	semsit	lang	fm	refl	vc	trole	sprep	grc	str	expl
1	change1	eng	indic	nrf	actv	Money-sum	nil	1	money-value	ex1
1	change1	eng	indic	nrf	actv	Location	nil	1	space-ob	ex2
1	change1	eng	indic	nrf	actv	Time	on	0	moment	ex3
1	change2	eng	indic	nrf	actv	Focus-object	nil	0	phys.ob	ex4
1	change2	eng	indic	nrf	actv	Start-time	since	0	moment	ex5
1	change2	eng	indic	nrf	actv	Time-interval	during	0	moment	ex6

Here the elements *eng, indic, nrf, actv* are interpreted as the values *English, indicative-mood, non-reflexive, active-voice* of the properties *language, form-of-verb, reflexivity, voice*; the elements *Money-sum, Location, Time, Focus-object, Start-time, Time-interval* are the designations of thematic roles (or conceptual cases); ex1 = “(Yves) has changed 700 franks”, ex2 = “(Yves) has changed (700 franks) in the exchange office No. 14”, ex3 = “(Yves) has changed (700 franks in the exchange office No. 14) on the 4th of March”, ex4 = “Mary has changed (very much since last summer)”; ex5 = “(Mary) has changed (very much) since last summer”; ex6 = “The town has changed very much during the 2000s)”. The fragments outside the parentheses are just the fragments where the considered thematic role (in other terms, a conceptual case) is realized. The fragments inside the parentheses only complement the fragments of the first kind in order to form a sentence.

9.4.4 Step 4-4: Constructing matrix semantic-syntactic representation of the input text

Following (Fomichov, 2010a), let's consider a new data structure called *a matrix semantic-syntactic representation (MSSR)* of a natural language input text T. This data structure will be used for representing the intermediate results of semantic-syntactic analysis on a NL-text. A MSSR of a NL-text T is a string-numerical matrix *Matr* with the indices of columns or the groups of columns

locunit, nval, prep, posdir, reldir, mark, qt, nattr ,

it is used for discovering the conceptual (or semantic) relations between the meanings of the fragments of the text T, proceeding from the information about linguistically correct short word combinations. Besides, a MSSR of a NL-text allows for selecting one among several possible meanings of an elementary lexical unit. The number of the rows of the matrix *Matr* equals to *nt* - the number of the rows in the classifying representation *Rc*, i.e. it equals to the number of elementary meaningful text units in T.

Let's suppose that *k* is the number of arbitrary row from MSSR *Matr.* Then the element *Matr[k, locunit],* i.e. the element on the intersection of the row k and the column with the index *locunit* is the least number of a row from the array *Arls* (it is the projection of the lexico-semantic dictionary *Lsdic* on the input text T) corresponding to the elementary meaningful lexical unit *Rc[k, unit]*. It is possible to say that the value *Matr[k, locunit]* for the *k*-th elementary meaningful lexical unit from T is the coordinate of the entry into the array *Arls* corresponding to this lexical unit .

The column *nval* of *Matr* is used as follows. If *k* is the ordered number of arbitrary row in *Rc* and *Matr* corresponding to the elementary meaningful lexical unit, then the initial value of *Matr[k, nval]* is equal to the quantity of all rows from *Arls* corresponding to this lexical unit; that is, corresponding to different meanings of this lexical unit. When the construction of *Matr* is finished, the situation is to be different for all lexical units with several possible meanings: for each row of *Matr* with the ordered number *k* corresponding to a lexical unit, *Matr[k, nval]* = 1. because a certain meaning was selected for each elementary meaningful lexical unit.

For each row of *Matr* with the ordered number *k* associated with a noun or an adjective, the element in the column *prep* (preposition) specifies the preposition (possibly, the void, or empty, preposition *nil*) relating to the lexical unit corresponding to the k-th row.

Let's consider the purpose of introducing the column group

$$posdir\ (posdir_1,\ posdir_2,\ ...,\ posdir_n),$$

where *n* is a constant between 1 and 10 depending on the program implementation. Let $1 \le d \le n$. Then we will use the designation *Matr[k, posdir, d]* for an element located at the intersection of the *k*-th row and the d-th column in the group *posdir*. If $1 \le k \le nt$, $1 \le d \le n$, then *Matr[k, posdir, d] = m*, where *m* is either 0 or the ordered number of the *d*-th lexical unit *wd* from the input text T, where *wd* governs the text unit with the ordered number k.

There are no governing lexical units for the verbs in the principal clauses of the sentences, that is why for the row with the ordered number *m* associated with a verb, *Matr[m, posdir, d] = 0* for any *d* from 1 to *n*. Let's agree that the nouns govern the adjectives as well as govern the designations of the numbers (e.g. "5 scientific articles"), cardinal numerals, and ordinal numerals. The group of the columns *reldir* consists of semantic relations whose existence is reflected in the columns of the group *posdir*. For filling in these columns, the templates (or frames) from the arrays *Arls, Arvfr, Arfrp* are to be used; the method can be grasped from the analysis of the algorithm *BuildMatr1* constructing a matrix semantic-syntactic representation of an input NL-text stated in (Fomichov, 2010a).

The column with the index *mark* is to be used for storing the variables denoting the different entities mentioned in the input text (including the events indicated by verbs, participles, gerunds, verbal nouns). The column *qt* (quantity) equals either to 0 or to the designation of the number situated in the text before a noun and connected to a noun. The column *nattr* (number of attributes) equals either to 0 or to the quantity of adjectives related to a noun presented by the *k*-th row, if we suppose that *Rc[k, unit]* is a noun.

10 Step 5 of the new method: development of an algorithm of semantic assembly

The content of the Step 5 is to use a matrix semantic-syntactic representation of a NL-text T as an intermediary data structure for constructing a semantic representation of T being an expression of a certain SK-language (that is, being a K-representation of T). The algorithm of semantic assembly *BuildSem1* described in Chapter 10 of (Fomichov, 2010a) gives an example of realizing this step of designing a NLPS.

Example. Let T1 be the question "Did the management board of the firm “Rainbow” change in May?", and T1germ be the same question in German “Hat der Verwaltungsrat der Firma “Rainbow” in Mai veraendernt sich?”. Then it is possible to associate both with T1 and with T1germ the same K-representation *Semrepr* of the form

*Question (x1, (x1 ≡ Truth-value(Situation(e1, change2 * (Focus-object,*
*certn manag-board * (Assoc-company, certn company1 * (Name1, “Rainbow”) : x3) : x2)*
(Time, Last-month(May, current-year)))))) .

11 Conclusions

The theory of K-representations was developed as a tool for dealing with numerous questions of studying semantics of arbitrarily complex natural language texts: both sentences and discourses. Grasping the main ideas and methods of this theory requires considerably more time than it is necessary for starting to construct the formulas of the first-order predicates logic. However, the efforts aimed at studying the foundations of the theory of K-representations would be highly rewarded. Independently on an application domain, a designer of a NLPS will have a convenient tool for solving various problems.

A new method of developing the algorithms of semantic-syntactic analysis of NL-texts was described above. The method has a number of significant advantages in comparison with other known methods of developing the algorithms of the kind. *Firstly*, the method was used for developing the algorithm *SemSynt1* described in

(Fomichov, 2010a). The explicitness and fullness of the description of the algorithm *SemSynt1* is many times higher than it is typical for the scientific publications on this problem (see, e.g., the paper (Popescu et al., 2003)). *Secondly*, the approach doesn't foresee the construction of a pure syntactic representation of the analyzed NL-text: it is oriented at discovering the semantic relations between the elementary meaningful units of a text.

Thirdly, the algorithm *SemSynt1* is multilingual in the following sense. This algorithm allows for using the same semantic-syntactic part of a linguistic database for English, German, and Russian languages. The algorithm *SemSynt1* contains the fragments meaning the calls of language-dependent auxiliary procedures. These procedures find several parts of a compound verbal form and join them into one elementary meaningful text unit, associate a preposition with a noun, etc. However, the discovery of possible semantic relations between the elementary meaningful text units is language-independent, and this promises economic advantages in case when the significant information may be obtained from the sources in several natural languages.

It seems that the method stated above together with the algorithm *SemSynt1* (as a substantial example of using this method) can be used as a framework for designing multilingual semantics-oriented analyzers of NL-texts and for obtaining much more detailed documentation of the algorithms as it is usually done.

References

BIRD S., KLEIN E., LOPER E. (2009). *Natural Language Processing with Python.* O'Reily.

CIMIANO P., HAASE P., HEIZMANN J., MANTEL M. (2007). ORAKEL: A portable natural language interface to knowledge Bases. *Technical report,* Institute AIFB, University of Karlsruhe, Germany.

DUKE A., GLOVER T., DAVIES J. (2007). Squirrel: An advanced semantic search and browse facility. In: *Proc. of the 4th European Semantic Web Conference. Innsbruck, Austria.*

FOMITCHOV V. (1984). Formal systems for natural language man-machine interaction modeling. In: *Ponomaryov, V.M. (ed.) Artificial Intelligence. Proc. of the IFAC Symposium, Leningrad, USSR, 4-6 October 1983 (IFAC Proc. Series, 1984, No. 9)*. Oxford, UK; New York: Pergamon Press, 203-209.

FOMICHOV V.A. (1992). Mathematical models of natural-language-processing systems as cybernetic models of a new kind. *Cybernetica (Belgium),* Vol. XXXV, 63-91.

FOMICHOV V.A. (1993a). Towards a mathematical theory of natural-language communication. *Informatica. An Intern. J. of Computing and Informatics (Slovenia),* Vol.17, 21-34.

FOMICHOV V.A. (1993b). K-calculuses and K-languages as powerful formal means to design intelligent systems processing medical texts. *Cybernetica (Belgium),* Vol. XXXVI, 161-182.

FOMICHOV V.A. (1994). Integral Formal Semantics and the design of legal full-text databases. *Cybernetica (Belgium),* Vol. XXXVII, 145-177.

FOMICHOV V.A. (2005). *The Formalization of Designing Natural Language Processing Systems.* Moscow: MAX Press (in Russian).

FOMICHOV V.A. (2007). *Mathematical Foundations of Representing the Content of Messages Sent by Computer Intelligent Agents.* Moscow: State University – Higher School of Economics, Publishing House "TEIS" (in Russian).

FOMICHOV V.A. (2009a) Theory of K-representations as a Source of an Advanced Language Platform for Semantic Web of a New Generation. *Web Science Overlay J. On-line Proceedings of the First International Conference on Web Science, Athens, Greece, March 18-20, 2009*; available at http://journal.webscience.org/221/1/websci09_submission_128.pdf.

FOMICHOV V.A. (2009b). A Scheme and Formal Tools for Transforming the Existing Web into Semantic Web of a New Generation. In: *Pre-Conference Proceedings of the Focus Symposium on Knowledge Management Systems (August 4, 2009, Focus Symposia Chair: Jens Pohl) in conjunction with InterSymp-2009, 21st International Conference on Systems Research, Informatics and Cybernetics, August 3 – 7, 2009, Baden-Baden, Germany),* Collaborative Agent Design Research Center, California Polytechnic State University, San Luis Obispo, CA, USA, 39-50.

FOMICHOV V.A. (2010a). *Semantics-Oriented Natural Language Processing: Mathematical Models and Algorithms.* New York, Dordrecht, Heidelberg, London : Springer.

FOMICHOV V.A. (2010b). Theory of K-representations as a Comprehensive Formal Framework for Developing a Multilingual Semantic Web. *Informatica. An International Journal of Computing and Informatics (Slovenia),* Vol. 34, No. 3, 387-396.

FRANK A., KRIEGER H.-U., XU F., USZKOREIT H., CRYSMANN B., JRG B., SCHAEFER U. (2007). Question answering from structured knowledge sources. *J. of Applied Logic,* 5 (1), 20-48.

POPESCU A.-M., ETZIONI O., KAUTZ H. (2003). Towards a theory of natural language interfaces to databases. In: *Proc. of the 8th International Conference on Intelligent User Interfaces,* Miami, FL, 149-157.

PRINCE V., ROCHE M., eds (2009). *Information Retrieval in Biomedicine: Natural Language Processing for Knowledge Integration.* IGI Global.

TURNPENNY P.D., ELLARD S. (2005). *Emery's Elements of Medical Genetics. Twelfth Edition*. Edinburgh, London, New York, Oxford, Sydney, Toronto: Elsevier Limited.

WILKS Y., BREWSTER C. (2006). *Natural Language Processing as a Foundation of the Semantic Web. Foundations and Trends in Web Science,* Vol. 1, No. 3. Hanover, MA; Delft: now Publishers Inc.