Theory of K-representations as a Comprehensive Formal Framework for Developing a Multilingual Semantic Web

Vladimir A. Fomichov
Department of Innovations and Business in the Sphere of Informational Technologies
Faculty of Business Informatics State University – Higher School of Economics
Kirpichnaya str. 33, 105679 Moscow, Russia
E-mail: vfomichov@hse.ru, vfomichov@gmail.com

Keywords: semantic web of a new generation, multilingual semantic web, semantics-oriented natural language processing, semantic representation, text meaning representation, theory of K-representations, SK-languages, semantic annotation, algorithm of semantic-syntactic analysis, bioinformatics, bioNLP

Received: March 15, 2010

A comprehensive theoretical framework for the development of a Semantic Web of a new generation, or of a Multilingual Semantic Web, is outlined. Firstly, the paper grounds the possibility of using a mathematical model being the kernel of the theory of K-representations and describing a system of 10 partial operations on conceptual structures for building semantic representations (or text meaning representations) of, likely, arbitrary sentences and discourses in English, Russian, French, German, and other languages. The possibilities of using SK-languages defined by the theory of K-representations for building semantic annotations of informational sources and for constructing semantic representations of discourses pertaining to biology and medicine are illustrated. Secondly, an original strategy of transforming the existing Web into a Semantic Web of a new generation with the well-developed mechanisms of understanding natural language texts is described. The third subject of this paper is a description of the correspondence between the inputs and outputs of the elaborated algorithm of semantic-syntactic analysis and of its advantages; the semantic representations of the input texts are the expressions of SK-languages (standard knowledge languages). The input texts can be the statements, questions, and commands from the sublanguages of English, Russian, and German. The algorithm has been implemented by means of the programming language PYTHON.

Povzetek: Predstavljena je formalizacija multilingualnega semantičnega spleta.

1 Introduction

Due to the stormy growth of the Internet, a huge number of the projects realized in the 2000s in Life Sciences and Health Care, and due to several other factors the users of the Internet and of specialized computer networks have received the access to an immense variety of information sources in many natural languages and to a number of knowledge bases formed with the help of ontology languages, first of all, the language OWL.

With respect to this situation, many specialists in various countries suppose that the only real way of realizing an effective interaction of people throughout the world with these natural language (NL) based information sources and with knowledge bases is the development of appropriate NL-interfaces and semantics-oriented advanced search systems.

In favour of this conclusion says the successful experience of designing in the 2000s several NL-interfaces to databases (see, e.g., [19]) and NL-interfaces to Semantic Web (SW) data repositories (see, e.g., [5, 6, 16]).

Since Web-based informational sources are formed with the help of many natural languages, it is high time to intensively develop the theoretical foundations of multilingual, semantics-oriented information retrieval on the Web and to expand the foundations of designing (for many natural languages) the NL-interfaces to SW data repositories.

On one hand, it is one of the central tasks for Web science, defined in [3] as the science of decentralized information systems. On the other hand, it seems that this task is a part of more general, large-scale problem – the problem of developing a Semantic Web of a new generation.

During several last years, it has been possible to observe that the achieved state of Semantic Web and a state to be relatively soon achieved are considerably different from the state of affairs outlined as the goal in the starting publication on Semantic Web by T. Berners-Lee, J. Hendler, and O. Lassila [2].

The principal reason for this conclusion is the lack of large-scale applications implemented under the framework of Semantic Web project. This situation is implied by the lack of a sufficiently big amount (of "a critical mass") of formally represented content conveyed by numerous informational sources in many fields. This means the lack of a sufficiently big amount of Web-sources and Web-services with semantic annotations, of

the visual images stored in multimedia databases and linked with the high-level conceptual descriptions, rich ontologies, etc.

This situation is often characterized as *the lack of a critical mass of semantic content.* That is why it has been possible to observe the permanent expansion in the scientific literature of the following opinion: a Semantic Web satisfying the initial goal of this project will be created in an evolutionary way as a result of the efforts of many research groups in various fields. In particular, this opinion is expressed in [1].

It is important to underline that this point of view is also expressed in the article "Semantic Web Revisited" written by the pioneers of Web: N. Shadboldt, W. Hall, T. Berners-Lee [22]. In this paper, the e-science international community is indicated as a community playing now one of the most important roles in quick generation of semantic content in a number of fields. The activity of this community seems to give a sign of future success of Semantic Web project.

One of the brightest manifestations of the need of new, strong impulses to developing Semantic Web is the organization of the First International Symposium on Incentives for Semantic Web under the framework of the Semantic Web International Conference – 2008 (Germany, Karlsruhe, October 2008).

The content of this paper is to be considered in the context of the broadly recognized need of the incentives for Semantic Web. Continuing the line of the papers [12 - 15] and the monograph [9], this paper outlines a comprehensive theoretical framework for the development of a Semantic Web of a new generation; it may be also called a Meanings Understanding Web [13] or a Multilingual Semantic Web with respect to [17].

Firstly, the paper grounds the possibility of using a mathematical model introduced in the monograph [9] 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 English, Russian, French, German, and other natural languages (texts pertaining to arbitrary spheres of professional activity). This model is the kernel of the theory of K-representations (knowledge representations).

Secondly, the paper sets forth an original strategy of transforming the existing Web into a Semantic Web of a new generation with the well developed mechanisms of understanding natural language texts.

For the realization of this strategy, the theory of K-representations provides a number of broadly applicable formal tools. *The third subject* of this paper are the peculiarities and input-output characteristics of the elaborated algorithm of semantic-syntactic analysis forming one of the principal constituents of the theory of K-representations. The outputs of this algorithm are the semantic representations of the input NL-texts being the expressions of SK-languages (standard knowledge languages). The input texts of this algorithm belong to the sublanguages of English, Russian, and German languages. For the development of a program implementation of this algorithm, the programming language PYTHON has been used.

2 The need of an advanced language platform for semantic Web

In [22], N. Shadbolt, W. Hall, and T. Berners-Lee ground the use of RDF as the basic language of the Semantic Web project with the help of the principle of least power: "the less expressive the language, the more reusable the data". As it is well known, the basic data structure of RDF is the triplets of the form subject – predicate – object. However, it seems that the stormy progress of, first of all, e-science urges us to find a new interpretation of this principle in the context of the challenges faced nowadays by the Semantic Web project. E-science (in particular, bioinformatics) needs to store on the Web the semantic content of the definitions of numerous notions, the content of scientific articles, technical reports, etc. The similar requirements are associated with semantics-oriented computer processing of the documents pertaining to economy, technology, medicine, law, politics, sport. In particular, it is necessary to store the semantic content of the articles from newspapers, of TV-presentations, etc.

The substantial discussions of the role of semantics-oriented natural language processing mechanisms for constructing a Semantic Web satisfying the demands of numerous end users can be found in the papers [12 – 15] and in the monographs [7, 9].

That is why it can be conjectured (see also [14]) that, in the context of the Semantic Web project, the following new interpretation of the principle of least power is reasonable: an advanced language platform for Semantic Web is to allow for modeling a system of operations on conceptual structures enabling us to build semantic representations (SRs) of practically arbitrary texts in Natural Language (NL) pertaining to arbitrary field of professional activity.

3 Shortly about ten conceptual operations considered by the theory of SK-languages

The question immediately emerges what a system of operations on conceptual structures satisfying the mentioned requirement might look like. A possible answer to this question is given by the theory of K-representations (knowledge representations) stated in the monograph [9]. The basic mathematical model of this theory describes a system consisting of 10 partial operations on conceptual structures [7 - 9]. The model determines a new class of formal languages for building semantic representations (SRs) of sentences and complex discourses in NL – the class of SK-languages (standard knowledge languages). An early version of this model set forth in [10, 11] determines the class of RSK-languages (restricted standard knowledge languages).

Let’s consider the central ideas of determining the class of SK-languages At the first step (consisting of a

rather long sequence of auxiliary steps), a class of formal objects called *conceptual bases (c.b)* is defined. 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.

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)*.

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 *China, India, Sri_Lanka*. Hence the value of the partial operation *Op[7]* (it governs the use of logical connectives $\wedge$ - AND and $\vee$ - OR) on the four-tuple

$< \vee$, *China, India, Sri-Lanka* >

is the K-string *(China* $\vee$ *India* $\vee$ *Sri-Lanka).*

Besides, X(B) includes the items article1 (a paper), article2 (a manufactured article), and h1 = article2, h2 = Kind1(certn article2, ceramics), h3 = (Country1(certn article2) ≡ (China $\vee$ India $\vee$ Sri-Lanka)), h4 = article2 * (Kind1, ceramics) (Country1, (China $\vee$ India $\vee$ 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 [9] 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. The short characteristics of these partial operations on conceptual structures can be found, in particular, in [13].

4 The use of SK-languages for building semantic representations of complex biomedical discourses

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 [20]. The formalization of natural language semantics is a very acute problem of BioNLP. 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 biomedicine.

Let D1 = T1.T2, where T1 = "The scientists know that a sequence of three bases (triplet) contains the message to call for the attachment of a specific amino acid in the protein chain", and T2 = "For example, the mRNA base code sequence GUC (guanine, uracil, cytosine) on mRNA calls for the attachment of the amino acid valine, while the mRNA base code sequence AUG (adenine, uracil, guanine) calls for the attachment of the amino acid methionine".

Let Semrepr1 = Situation(e1, knowing * (Agent1, certn set * (Qual-compos, scientist) : S1)(Content1, Contain2(arbitr sequence * (Numb, 3)(Qual-compos, base1) : x1, certn info-piece * (Determinator1, certn attachment1 * (Dynamic-object, specific amino-acid : x3)(Goal-object, certn chain1 * (Qual-compos, protein1) : x4)) : x2) : P1)).

Let us interpret the formula *Semrepr1* as a possible K-representation of the first sentence T1, that is, as a semantic representation (SR) of T1 being an expression of the SK-language determined by a certain conceptual basis. In the formula *Semrepr1*, the variable *P1* plays the role of a mark of the meaning of the principal part of the first sentence T1.

Let *Semrepr2* be the formula

*Example(P1, 1, Call-for(arbitr sequence * (Numb, 3)(Qual-compos, base1)(Compos-seq, (<1, G>* $\wedge$ *<2, U>* $\wedge$ *<3, C>)) : x5, certn attachment1 * (Dynamic-object,*

*specific amino-acid * (Name1, "valine")(Location, certn mRNA : x6) : x7) (Goal-object, certn chain1 * (Qual-compos, protein1) : x8))).*

Let *Semrepr3* be the formula

*Example(P1, 2, Call-for(arbitr sequence * (Numb, 3)(Qual-compos, base1)(Compos-seq, (<1, A> ∧ <2, U> ∧ <3, G>)) : x9, certn attachment1 * (Dynamic-object, specific amino-acid * (Name1, "methionine") : x10) (Goal-object, x8))),*

and let *Semdisc1 = (Semrepr1 ∧ Semrepr2 ∧ Semrepr3).*

Then the formula *Semdisc1* can be interpreted as a possible K-representation of the discourse D1. This formula provides the possibility to indicate several important advantages of the K-representations theory in comparison with first-order predicate logic and the Theory of Conceptual Graphs.

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 of people.

5 A universal tool for constructing semantic annotations

The analysis of a number of publications studying the problem of transforming the existing Web into Semantic Web allows for drawing the following conclusion: an ideal configuration of Semantic Web would be a collection of interrelated resources, where each of them has both an annotation in natural language (NL) and a formal annotation reflecting the meaning or generalized meaning of this resource, i.e. a semantic annotation. NL-annotations would be very convenient for the end users, and semantic annotations would be used by question-answering systems and advanced search engines.

Most likely, the first idea concerning the formation of semantic annotations of Web data would be to use the formal means for building semantic representations (SRs), or text meaning representations, of NL-texts provided by mathematical and computational linguistics.

However, the analysis shows that the expressive power of the main popular approaches to building SRs of NL-texts, in particular, of Discourse Representation Theory, Theory of Conceptual Graphs, and Episodic Logic is insufficient for effective representing contents of arbitrary Web data, in particular, of arbitrary biological, medical, or business documents.

First of all, the restrictions concern describing semantic structure of: (a) infinitives with dependent words (e.g., representing the goals, commitments, and the intended manners of using things and procedures); (b) constructions formed from the infinitives with dependent words by means of the logical connectives "and", "or", "not"; (c) the complex designations of sets; (d) the fragments where the logical connectives "and", "or" join not the designations of assertions but the designations of objects ("the product A is distributed by the firms B1, B2, ..., BN"); (e) the explanations of the terms being unknown to an applied intelligent system; (f) the fragments containing the references to the meanings of phrases or larger fragments of a discourse ("this method", etc.); (g) the designations of the functions whose arguments and/or values may be the sets of objects ("the staff of the firm A", "the number of the suppliers of the firm A", etc.).

Taking into account this situation and the fact that the semantic annotations of Web-sources are to be compatible with the format of representing the pieces of knowledge in ontologies, a number of researchers undertook the efforts of constructing computer intelligent systems, using the languages RDF, RDFS or OWL for building semantic annotations of Web-sources [18, 21].

However, the expressive power of RDF, RDFS or OWL is insufficient for being an adequate formal tool of building semantic annotations of scientific papers, technical reports, etc.

Meanwhile, the formulated idea of where to get the formal means for building semantic annotations from is correct. The main purpose of this section is to illustrate some principal ideas of employing the SK-languages for building semantic annotations of informational sources, in particular, Web-based sources.

Example. Let's consider a possible way of employing SK-languages for building a semantic annotation of the famous paper "The Semantic Web" by T. Berners-Lee, J. Hendler, and O. Lassila published in "Scientific American" in May 2001 [2].

Suppose that there is a Web-source associating the following NL-annotation with this paper: "It is proposed to create such a net of Web-based computer intelligent agents (CIAs) being able to understand the content of almost every Web-page that a part of this net will be composed by CIAs being able to understand natural language".

A semantic annotation corresponding to this NL-annotation can be the K-string of the form

*certn inf.ob * (Kind1, sci_article)(Source1,*
*certn journal1 * (Name1, "Scientific_American") : x1)*
(Year, 2001)(Month, May))(Authors,
*certn group1 * (Numb, 3)(Elements1,*
*(< 1, certn scholar * (First_name, "Tim")(Surname,*
*"Berners-Lee") : x2 > ∧ < 2, certn scholar **
(First_name, "James")(Surname, "Hendler") : x3 >
*∧ < 3, certn scholar * (First_name, "Ora")(Surname,*
"Lassila") : x4 >)) : S1) (Central_ideas,
(< 1, Semrepr1 > ∧ < 2, Semrepr2 >)) : v ,

where the variable *S1* designates the group consisting of all authors of this article, v is a variable being a mark of the constructed semantic annotation as an informational

object, and *Semrepr1, Semrepr2* are the K-strings defined by the following relationships:

*Semrepr1 = Proposed(S1, creation1 * (Product1 , certn family1 * (Qual-compos, intel_comp_agent * (Property, web-based)(Ability, understanding1 * (Inf_object , Content(almost_ every web_page)))) : S2) (Time, certn time_interval * (Part1, Nearest_future(decade1, #now#))) ,*

*Semrepr2 = Proposed(S1, achieving_situation * (Description1, (Exists (S3, set) ∧ Subset(S3, S2) ∧ Qual-compos (S3, intel_comp_agent * (Property , web-based)(Ability, understanding1 * (Inf_object , almost_every text * (Language1, certn language * (Belong, NL_family))))))) .*

To sum up, a comprehensive formal tool for building semantic annotations of Web data is elaborated. This tool is the theory of SK-languages. A very important additional expressive mechanism of SK-languages in comparison with the mechanisms illustrated in the example above is the convenience of building semantic representations of discourses with the references to the meanings of phrases and larger parts of a discourse.

The analysis of expressive power of the class of SK-languages (see the chapters 5 and 6 of [9]) allows for conjecturing that it is both possible and convenient to construct semantic annotations of arbitrary Web data by means of SK-languages. That is why the theory of SK-languages can be interpreted as a powerful and flexible (likely, universal) formal metagrammar of semantic annotations of Web data.

6 The formal tools provided by the theory of K-representations

The monographs [7], [9], stating two versions of the theory of K-representations, propose one universal (most likely) and several broadly applicable formal tools for the realization of this strategy.

The *first basic constituent* of the theory of K-representations is the theory of SK-languages (standard knowledge languages), stated, in particular, in [7 - 9]. 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 expressions of SK-languages will be called below the K-strings. If T is an expression in natural language (NL) and a K-string *E* can be interpreted as a SR of T, then *E* is called a K-representation (KR) of the expression T.

The *second basic constituent* of the theory of K-representations is a widely applicable mathematical model of a linguistic database (LDB). The model describes the frames expressing the necessary conditions of the existence of semantic relations, in particular, in the word combinations of the following kinds: “Verbal form (verb, participle, gerund) + Preposition + Noun”, “Verbal form + Noun”, “Noun1 + Preposition + Noun2”, “Noun1 + Noun2”, “Number designation + Noun”, “Attribute + Noun”, “Interrogative word + Verb”. The expressive power of SK-languages enables us to associate the lexical units with the appropriate simple or compound semantic units. The model describes the logical structure of linguistic databases being the components of natural-language interfaces to intelligent databases as well as to other applied computer systems (see Chapter 7 of [9]).

The *third basic constituent* of the theory of K-representations is several complex, strongly structured algorithms carrying out semantic-syntactic analysis of texts from some practically interesting sublanguages of NL. More details about these algorithms can be found below (see also Chapters 8 - 10 of [9]).

7 The principles of designing natural language processing systems

Most often, semantics-oriented natural language processing systems, or linguistic processors (LPs), are complex computer systems, their design requires a considerable time, and its cost is rather high. Usually, it is necessary to construct a series of LPs, 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 LPs 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 LPs, it seems to be reasonable to pay more attention to: (a) the search for best typical design solutions concerning the key subsystems of LPs with the aim to use these solutions in different domains of employing LPs; (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 LPs by one research centre or a group will contribute, in the long-term perspective, to reducing the total cost of designing a family of LPs and to minimizing the duration of constructing each particular system from this family:

the ***Principle of Stability*** of the used language of semantic representations (LSR) 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 LP 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 LP are repeatedly used by different linguistic processors).

The theoretical results stated in chapters 1 - 6 of the monograph [9] provide a basis for following-up the principle of stability of the used language of semantic representations. Chapter 4 defines a class of SK-languages (standard knowledge languages) that enable us to build semantic representations of natural language texts in arbitrary application domains. The broad perspectives for following-up the principle of succession of the algorithms of semantic-syntactic analysis of NL-texts are opened by the content of chapters 7 – 10 of [9].

8 A possible strategy of developing a multilingual semantic web

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 [23]. 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 [9]) 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 [7 – 9, 13, 15]: SK-languages are a convenient tool of building semantic representations of arbitrarily complex natural language texts (sentences and discourses) pertaining to arbitrary field of professional activity.

This central idea underlies the strategy (described below) 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 previous versions of this strategy are published in [9, 15].

The proposed strategy is based on (a) the mathematical model constructed in [9] and describing a system of 10 partial operations on conceptual structures and (b) the analysis of the expressive mechanisms of SK-languages. The new strategy can be very shortly formulated as follows:

1. An XML-based format for representing the expressions of SK-languages (standard knowledge languages) will be elaborated. Let's agree that the term "a K-representation of a NL-text T" means below a semantic representation of T built in this format and that the term "a semantic K-annotation" will be interpreted below as a K-representation of a NL-annotation of an informational source. The similar interpretations will have the terms "a K-representation of a knowledge piece" and "a high-level conceptual K-description of a visual image".
2. The NL-interfaces for different sublanguages of NL (English, Russian, German, Chinese, Japan, etc.) helping the end users to build semantic K-annotations of Web-sources and Web-services are being designed.
3. The advanced ontologies being compatible with OWL and using K-representations of knowledge pieces are being elaborated.

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 ∨ mammal * (Examples, (sheep ∧ goal)))), S1,*

*(Estimation1(Quantity(S1), high) ∧ Goal-of-forming (S1, certain purpose * (Examples, (feeding ∨ migration ∨ 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 ∧ goal))* and *dynamic-group * (Qualitative-composition, (bird ∨ mammal * (Examples, (sheep ∧ 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 ∨ migration ∨ defence))* .

The structure of the constructed K-representation *Expr1* to a considerable extent reflects the structure of the definition T1.

4. The new content languages using K-representations of the content of messages sent by computer intelligent agents (CIAs) in multi-agent systems are being worked up. In particular, this class of languages is to include a subclass being convenient for building the contracts concluded by the CIAs as a result of successful commercial negotiations.
5. The visual images of the data stored in multimedia databases are being linked with high-level conceptual K-descriptions of these images (see Section 6.3 of [9]).
6. The NL-interfaces transforming the NL-requests of the end users of Web into the K-representations are being designed.
7. The advanced Web-based search and question-answering systems are being created being able (a) to transform (depending on the input request) the fragments of a discourse into the K-representations, (b) to analyze these K-representations of the discourse fragments, and (c) to analyze semantic K-annotations of Web-sources and Web-services.
8. The NL processing systems being able to automatically extract knowledge from NL-texts, to build the K-representations of knowledge pieces, and to inscribe these K-representations into the existing ontologies are being elaborated.
9. The generators of NL-texts (the recommendations for the users of expert systems or of recommender systems, the summaries of Web-documents, etc.) using the SK-languages for representing the meaning of a NL-text to be synthesized are being constructed. Besides, a reasonable direction of research seems to be the design of applied intelligent systems being able to present the semantic content of a message for the end user as an expression of a non-standard K-language being similar to a NL-expression but containing, may be, a number of brackets, variables, markers.

Fulfilling these steps, the international scientific community will create in a reasonable time a digital conceptual space unified by a general-purpose language platform. The realization of this strategy will depend on the results of its discussion by the international scientific community.

9 A new method of developing multilingual semantic-syntactic analyzers of NL-texts

9.1 The advantages of a new method

It seems that the complete potential of semantics-oriented approach to designing multilingual algorithms of processing NL-texts is far from being exhausted. A new implementation of this approach is described in the monograph [9]. In essence, [9] describes a new method of developing the algorithms of semantic-syntactic analysis of NL-texts. This method can be reconstructed from the study of the algorithm *SemSynt1* completely described in Chapters 8 - 10 of [9].

The input texts of the algorithm *SemSynt1* can be the sentences (statements, commands, and questions) from some practically interesting sublanguages of English, Russian (a Latin transcription is used), and German languages. The output of the algorithm is a semantic representation of the input text being its K-representation.

The principal advantages of the new method are as follows: (1) the algorithm *SemSynt1* uses an original formal model of a linguistic database (see Chapter 7 of [9]), this model is problem-independent; (2) an important feature of the algorithm is that it doesn't construct any syntactic representation of the inputted NL-text but directly finds semantic relations between text units; since numerous lexical units have several meanings, the algorithm uses the information from a linguistic database and linguistic *context* for choosing one meaning of a lexical unit among several possible meanings; (3) the other distinguished feature is that this complex algorithm is completely described with the help of formal tools, that is why its description doesn't uses any expressive mechanisms of any concrete programming system; (4) the main procedures of the algorithm (of the upper and middle levels) are the same for the English, Russian, and German languages; (5) the main procedures of the algorithm *SemSynt1* are described with the help of the terms being well known to the programmers (one- and two-dimensional arrays, a string, a set, a binary conceptual relation between two elements) and don't demand a command of complicated linguistic terminology, often being specific for a concrete natural language.

9.2 The input-output characteristics of the multilingual algorithm *SemSynt1*

Let's consider the examples illustrating the correspondence between the natural language 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*.

Example 1. Let T1eng = "The international scientific conference "DEXA-2009" took place in Linz, Austria, during August 31 – September 4, 2009", T1rus = "Mezhdunarodnaya nauchnaya konferentsiya "DEXA-2009" prokhodila v gorode Linz, Avstriya s 31 avgusta po 4 sentyabrya 2009 goda", T1germ = "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 2. Let T2 = "Find a description of the programming language PYTHON on the Web-site http://docs.python.org", T3rus = "Naydite opisanie yazyka programmirovaniya PYTHON na veb-sayte http://docs.python.org", T3germ = "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 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 Implementation of the algorithm *SemSynt1*

An expanded and modified version of the algorithm *SemSynt1* has been implemented with the help of the programming language PYTHON; as it is shown in [4], this language proved to be a convenient tool of developing NL processing systems. The input language of the elaborated NL-interface SEMANTIKA (E.K. Orlov, Faculty of Business Informatics, State University – Higher School of Economics, Moscow) is broader than the input language of the algorithm *SemSynt1*: it includes the statements, questions, and commands in Russian that can contain the participle constructions and attributive clauses. For instance, the input language of the program SEMANTIKA includes the question "What medicines offered by the pharmaceutical firm "GlaxoSmithKlein" are produced in Poland?".

The predecessor of the *SemSynt1* – the algorithm *SemSyn* described in [7] – was implemented in the Web programming language PHP. Chapter 11 of the monograph [9] contains the examples illustrating the principles of processing NL-texts by the experimental Russian-language interface NL-OWL1, implemented in the Web programming system PHP and developed on the basis of the algorithm *SemSyn*. An particular, the example associating the definition "Carburettor is a device for preparing a gas mixture of petrol and air" firstly with a K-representation and later with an OWL-expression is considered.

10 Conclusion

The main result of this paper is an original strategy of transforming, step by step, the existing Web into a Semantic Web of new generation (SW-2), where the principal distinguished feature of SW-2 would be the well-developed ability of NL processing. That is why SW-2 can be also qualified as a Meanings Understanding Web or as a Multilingual Semantic Web.

The aim of proposing this strategy is to increase the total successfulness, effectiveness of the mainly decentralized global ongoing process of endowing the existing Web with the ability of understanding texts in many natural languages.

The proposed strategy is based on a broad spectrum of new possibilities provided by the theory of K-representations (knowledge representations) developed by the author of this paper and presented in [9]. In particular, the paper illustrates a number of new precious opportunities of using SK-languages for building semantic annotations of informational sources, constructing complex definitions of the concepts in the advanced ontologies, and building semantic representations (or text meaning representations) of complex discourses pertaining to biology and medicine.

The final part of the paper describes the peculiarities and input-output characteristics of a new multilingual algorithm of semantic-syntactic analysis of NL-texts (from the sublanguages of English, Russian, and German languages). This algorithm, called *SemSynt1,* is a part of the theory of K-representations and is presented in Chapters 9 and 10 of [9]. An expanded and modified version of *SemSynt*1 has been implemented with the help of the programming language PYTHON.

References

[1] Angelova, G. (2005). Language Technology Meets Ontology Acquisition. In: Dau, F., Mugnier, M.-L., Stumme, G. (eds.) International Conference on Conceptual Structures 2005. LNCS, Vol. 3596, Springer, Heidelberg, pp. 367-380.

[2] Berners-Lee, T., Hendler, J., Lassila, O. (2001). The Semantic Web. Scientific American, pp. 34-43.

[3] Berners-Lee, T., Hall, W., Hendler, J.A., O'Hara, K., Shadboldt, N., and D.J. Weitzner (2006). A Framework for Web Science. Foundations and Trends in Web Science, Vol. 1, No. 1, now Publishers Inc.-134 p.

[4] Bird, S., Klein, E., and E. Loper (2009). Natural Language Processing with Python. O'Reily.

[5] 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.

[6] 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*.

[7] Fomichov, V.A. (2005). The Formalization of Designing Linguistic Processors. Moscow, MAX Press (in Russian).-368 p.

[8] 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).-176 p.

[9] Fomichov, V.A. (2010). Semantics-Oriented Natural Language Processing: Mathematical Models and Algorithms. Springer, New York, Dordrecht, Heidelberg, London.-354 p.

[10] Fomichov, V.A. (1996). A Mathematical Model for Describing Structured Items of Conceptual Level. Informatica. An Intern. J. of Computing and Informatics (Slovenia), 20 (1), pp. 5-32.

[11] Fomichov, V.A. (1998). Theory of Restricted K-calculuses as a Comprehensive Framework for Constructing Agent Communication Languages. In: Fomichov V.A., Zeleznikar A.P. (eds.). Special Issue on NLP and Multi-Agent Systems.

Informatica. An International J. of Computing and Informatics (Slovenia), 22 (4), pp. 451-463.

[12] Fomichov, V.A. (2000). An Ontological Mathematical Framework for Electronic Commerce and Semantically-structured Web. In: Zhang Y., Fomichov V.A., Zeleznikar A.P. (eds.), Special Issue on Database, Web, and Cooperative Systems. Informatica. An International J. of Computing and Informatics (Slovenia), Vol. 24, No. 1, pp. 39-49.

[13] Fomichov, V.A. (2008). A Comprehensive Mathematical Framework for Bridging a Gap Between Two Approaches to Creating a Meaning-Understanding Web. International J. of Intelligent Computing and Cybernetics (Emerald Group Publishing Limited, UK), Vol. 1, No. 1, pp. 143-163.

[14] 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.

[15] 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, pp. 39-50.

[16] Frank, A., Krieger, H.-U., Xu, F., Uszkoreit, H., Crysmann, B., Jrg, B., Schaefer, U. (2007). Question Answering from Structured Knowledge Sources. *Journal of Applied Logic,* Vol. 5, No. 1, pp. 20-48.

[17] Multilingual Semantic Web (2009) CFP: 1st Workshop on the Multilingual Semantic Web (collocated with WWW 2010); received on Monday, 21 December 2009; http://lists.w3.org/Archives/Public/semantic-web/2009Dec/0065.html; retrieved 12.03.2010.

[18] Navigli, R., Velardi, P. (2006). Through automatic semantic annotation of on-line glossaries. In Proc. of European Knowledge Acquisition Workshop (EKAW)-2006, LNAI 4248, pp. 126-140.

[19] 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 , pp. 149-157.

[20] Prince, V., Roche, M., eds (2009). *Information Retrieval in Biomedicine: Natural Language Processing for Knowledge Integration.* IGI Global. 460 pp.

[21] Reeve, L., Han, H. (2005). Survey of semantic annotation platforms. Proc. of the 20th Annual ACM Symposium on Applied Computing and Web Technologies.

[22] Shadboldt, N., Hall, W., Berners-Lee, T. (2006). Semantic Web Revisited. IEEE Intelligent Systems, Vol. 21, No. 3, pp. 96-101.

[23] 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 - 4, now Publishers Inc.-129 p.