A New Version of a Broadly Applicable, Cross-lingual Meaning Representation Formalism and Its Significance for Biomedical Sciences

Vladimir A. Fomichov
Department of Systems Modeling and Design Automation, Moscow Aviation Institute (National Research University)
Moscow, Russian Federation
E-mail: vfomichov@gmail.com

Keywords: automatic information extraction, natural language processing, semantic parsing of scientific definitions, scholarly knowledge, biomedical sciences, abstract meaning representation, uniform meaning representation, theory of K-representations, SK-language

Received: March 10, 2024

The first purpose of the paper is to attract the attention of the scholars in natural language (NL) processing to a new version of a broadly applicable, cross-lingual meaning representation formalism: a new version of the theory of SK-languages (standard knowledge languages) introduced in early 2000s by the theory of K-representations (knowledge representations), or TKR. A collection of original expressive mechanisms for constructing semantic representations (SRs) of scientific notions' definitions is suggested, its joint work is demonstrated. A special attention is paid to indicating the advantages of TKR-based approach to building SRs of scientific definitions in comparison with Universal Conceptual Cognitive Annotation, Abstract Meaning Representation, and Uniform Meaning Representation. New precious and broad prospects of describing semantic structure of NL-texts pertaining to biomedical sciences are indicated. The second purpose is to improve an algorithm (introduced in a previous paper of the author) constructing SRs of scientific notions' definitions and to illustrate its principal ideas. The output SRs are the expressions of SK-languages. The methodological basis is TKR. In particular, the suggested algorithm of scientific definitions' semantic parsing includes a complex procedure based on the algorithm of semantic parsing SemSynt1 given by TKR. The original features of the suggested algorithm constructing SRs of scientific definitions are shown.

Povzetek: Članek uvaja izboljšano različico SK-jezikov za medjezikovno reprezentacijo pomena v obdelavi naravnega jezika, s poudarkom na biomedicinskih znanostih, ter izboljšuje algoritme za semantično razčlenjevanje znanstvenih definicij.

1 Introduction

During two last decades, a considerable progress has been achieved in many branches of the field of studies aimed at processing written texts and oral speech in natural language (NL). In particular, it applies to NL-interfaces to mobile devices, OWL-based ontologies, autonomous intelligent systems (robots), the computer systems extracting knowledge from NL-texts with the goal of forming and updating ontologies, question-answering systems dealing with full-text databases, NL-interfaces to the giant semantic information system Linked Open Data (LOD).

In the context of this progress, many specialists in NL processing (NLP) from various countries have realized the existence of a fundamental problem to be solved in order to continue this progress. This fundamental problem is the construction of broadly applicable (desirably, universal), cross-lingual meaning representation (MR) formalisms and of the algorithms transforming NL-texts into their meaning representations (or semantic representations, SRs) with respect to the used knowledge base and, in the case of discourses, in the context of the previous fragment of the analyzed discourse.

The recent impressive achievements in constructing NLP systems based on large languages models and black-box neural networks could stimulate many computer science specialists to draw the conclusion that the problem of developing broadly applicable, cross-lingual MR formalisms is not worth to concentrate much attention on solving it. However, the authors of [1, 2] underline that the black box nature of neural networks makes it difficult to know when and where to correct the used models for eliminating errors or at least anticipating errors. But it is absolutely necessary to have the possibilities of the kind, because interpretability and controllability in NLP systems are critical in high-stake application scenarios, in particular, for realizing reasonable human-intelligent robot interaction and in medicine for making reasonable expectations and conclusions.

The above said explains the reasons for organizing the First International Workshop on Designing Meaning Representations (DMR 2019, Florence, Italy, August 2019) [3]. The principal objectives of the workshop DMR 2019 were as follows: (a) to gain a deeper understanding of the key elements of meaning representations (MRs) that are the most valuable to the NLP community; (b) to critically examine existing MRs with the goal of using the findings to inform the design of next-generation MRs; (c)

to explore the opportunities and identify the challenges in the design and use of MRs in multilingual settings [3]. The Fifth International Workshop on Designing Meaning Representations (DMR 2024) took place in May 2024 in Torino, Italy [4].

The proceedings of the workshop DMR 2019 include the paper [5], its authors (three specialists from two IBM research centres in San Jose, California) underline that "creating a universal semantic representation that works across a large number of languages is an important objective for the NLP community".

Wonderfully, the first version of mathematical framework being appropriate for creating a universal, cross-lingual semantic representation was suggested 23 years before the start of the workshop DMR 2019 in the paper [6]. This paper introduced a mathematical model (having an original form) describing a system of ten partial operations on conceptual structures and introducing a new class of formal languages – the class of restricted standard knowledge languages (RSK-languages).

The paper [7] shortly explains the essence of these ten partial operations on conceptual structures. This information is used in the mentioned paper as the ground for demonstrating very broad expressive possibilities of RSK-languages. In particular, the possibilities of building SRs of compound scientific notions and of complex discourses. A considerable part of the paper [7] is devoted to explicating the advantages of the RSK-languages in comparison with Universal Networking Language (UNL) elaborated under the guidance of the UNO Institute for Advanced Studies, Tokyo University. This language is based on the idea of representing the meanings of separate sentences by means of binary relations.

The second version of a mathematical framework being appropriate for creating a universal, cross-lingual MR was published nine years before the workshop DMR 2019. It was done in the monograph [8] introducing the theory of K-representations (knowledge representations) – an original theory of designing semantic-syntactic parsers of NL with the broad use of formal means for representing input, intermediary, and output data. One of the constituents of the theory of K-representations (TKR) is the theory of SK-languages (standard knowledge languages).

The difference between the theory of RSK-languages and the theory of SK-languages can be explained as follows. The first theory assumes the existence of only one possible angle of look at any considered entity from an application domain. For instance, people are considered as intelligent objects, the cars – as dynamic physical objects, and the firms – as the organizations. However, each concrete person is simultaneously both an intelligent object and a dynamic physical object, the IT-companies develop, in particular, programming environments and have the locations. That is why the theory of SK-languages gives the possibility to consider the compound semantic characteristics (compound types) of the entities from application domains. For instance, each person may be associated with the type *intel. system * dyn. phys. object* and each firm with the type *org * intel. system * space. object*.

The essence of the principal constituents of TKR is shortly explained below in the sections 3 and 4.

Taking into account the objectives of the present paper, it is important to say that the monograph [8] includes the following conjecture: "The designers of NL processing systems have received a system of the rules for constructing well-formed formulas (besides, a compact system, it consists of only ten main rules) allowing for (according to the hypothesis of the author) building semantic representations (SRs) of arbitrary texts pertaining to numerous fields of humans' professional activity, i.e., SRs of the NL-texts on economy, medicine, law, technology, politics, etc." [8, p. ix-x].

Since the release of the monograph [8], no scholars have put forward any objections against this hypothesis.

There are serios reasons to believe that the construction of the theory of SK-languages in [8] means that the barrier of complexity in developing a universal meaning representation formalism was overcame nine years before the workshop DMR 2019 and, as it is clear now, outstripped the time of its creation.

The first objective of the present paper is to attract the attention of the NLP community to a new version of the SK-languages' theory and, as a consequence, of TKR being more compact and convenient for practical usage in comparison with the version of TKR stated in the monograph [8]. A new version is constructed due to introducing a compact mathematical model of a system of primary units of conceptual level used by an applied intelligent system (AIS). Formally, this goal is achieved due to introducing and considering the notion of *optimized conceptual basis* instead of the notion of conceptual basis.

The present paper explicates the broadest prospects opened for biomedical sciences by SK-languages. Two principal directions of reasoning are combined: (a) the demonstration of the possibilities to use SK-languages for building SRs of complex scholarly discourses and complex definitions of the notions; (b) the explication of the advantages of SK-languages in comparison with Universal Conceptual Cognitive Annotation [9], Abstract Meaning Representation [10], and Uniform Meaning Representation [1].

The second objective of the paper is to improve the algorithm AlgSemDef1 introduced in [11] constructing SRs of scientific notions' definitions and to illustrate the principal ideas of the suggested algorithm AlgSemDef2. The output SRs of the definitions are the expressions of SK-languages.

In Conclusions, the following hypothesis is formulated: the considered optimized version of the SK-languages' theory (a part of TKR) may be interpreted as the *starting universal meaning representation formalism.*

2 Computational semantics and biomedical sciences

During last decade, the development of the huge semantic information system Linked Open Data (LOD) [12, 13] caused the emergence of a big family of projects in the field of NL processing aimed at extracting factual

information in the form of triples. Mainly due to this reason, the attention of the major part of researchers in NLP was not focused on the problem of representing semantic structure of scientific texts in NL (the NL-texts) in a formal way.

A strong impulse for "switching-on" the interest of the researchers in NLP to this problem was given in the middle of the 2010s by the change of the paradigm in the field of constructing large knowledge bases (KBs) in biomedical sciences. The first limitation of the previous period of constructing biomedical KBs is that, most often, they were manually built and curated. The second limitation was that the prior work on automatic information extraction (AIE) from NL-texts was that it was focused on several distinguished classes of scientific publications describing, in particular, protein–protein interactions or gene-drug relationships or drug effects, etc.

The main aspects of changing (in the middle of the 2010s) the paradigm of AIE for constructing biomedical KBs is the transition to considering, as knowledge sources, not only arbitrary scientific publications in the field but also the sources like the health portals and popular online discussion forums. This transition is considered as the ground for developing new, comprehensive approaches linking diverse entity types, spanning genes, diseases, symptoms, drugs, drug effects, anatomic parts, etc. [14].

As a consequence, the designers of computer systems for constructing large knowledge graphs for biomedical sciences faced the need of strong and flexible formal means for representing semantic structure of complex NL-expressions.

During last fifteen years, the researchers working in NL processing have received four main formalisms for designing semantics-oriented NLP systems extracting knowledge from biomedical texts. The first one was suggested by V.A. Fomichov in the invited keynote address delivered at the opening session of the 18th international TALN conference (Traitement Automatique des Langues Naturelles, June 27 – July 1, 2011, France, University Montpellier 2) [15]. The principal subject of the keynote address was the broad prospects opened for bioinformatics and Semantic Web by the theory of K-representations (TKR) stated in [8].

Taking into account the objectives of the present paper, it should be noted that the keynote address [15] attracted, in particular, the attention of the researchers in bioinformatics to the expressive mechanisms being precious for representing scholarly knowledge, first of all, the definitions of scientific notions.

The approach UCCA [9] is qualified by its authors as a semantic representation. However, it seems that it may be qualified more exactly as a kind of semantic-syntactic representation (SSR). The reason for this conclusion is that the basic elements of the UCCA structures are not semantic units but the words (including the articles "a", "the', the pronouns "he", "his", etc.) and word combinations. The UCCA approach suggests to use directed acyclic graphs as SSRs of the sentences in NL. The authors of [9] suggested 12 original labels for marking the vertices of the graph. The basic expressions are called *scenes*, they may be of two types: *processual scenes (PS)* and *state scenes (StS)*.

One of the edges starting from the root of PS has the label P (Process), and one of the edges starting from the root of StS has the label S (State). The label A (Participant) denotes a participant in a scene in a broad sense. E.g., a linear representation of a scene associated with the sentence "Cukor encouraged the studio to accept her demands" is the expression

$$Cukor_A\ encouraged_P\ [the_E\ studio_C]_A\ [to_R\ [accept\ her\ demands]_C\ ]_A.$$

Here the label "C" (Center) is used for the conceptualization of the parent unit, "E" (Elaborator) marks a non-scene relation which applies to a single Center, "R" (Relation) corresponds to all other types of non-scene relations.

The main drawbacks of the UCCA approach seem to be as follows: (a) the basic elements are the words but not semantic units; (b) there are no expressive mechanisms for constructing SRs of the expressions in NL listed in the Table 1.

The semantic formalism AMR was introduced in the year 2013 in the ACL publication [10] by a group consisting of ten researchers from UK and USA. The central idea of AMR approach is to use rooted, directed acyclic graphs for representing the meaning of a sentence. The nodes in the graphs (called concepts) stem from the words and word combinations in a sentence, and the edges correspond to semantic-syntactic relationships between the words in the sentence. During last decade, a number of research groups in different countries realized the experimental projects aimed at parsing a sentence into its AMR [16-20].

The subject of the paper [21] is biomedical events extraction using AMR, the events are protein-protein interactions. An example of such sentence is as follows: "This LRA-induced rapid phosphorylatin of radixin was significantly suppressed in the presence of C3 toxi, a potent inhibitor of Rho".

The most recent specification of AMR is dated by May 1, 2019 [22].

During mainly last seven years, a number of the projects have emerged aimed at investigating the possibilities of using AMR for representing the meanings of sentences in Chinese, Turkish, Korean, Portuguese, Spanish, Vietnamese. The common feature of the publications describing such projects is that the researchers have considered only rather simple sentences of concrete languages being structurally very far from complex sentences encountered in scholarly documents.

AMR considers only sentences but not the discourses (the sequences of sentences interrelated by their meanings). For overcoming this restriction, the Uniform Meaning Representation was suggested in [1]. UMR consists of an AMR-based sentence level representation of a text that focuses on predicate-argument structures, the senses and named entities and a document level representation (in other words, a discourse level representation) that captures semantic relations going beyond the sentence boundaries. It applies to the time and causal relations and also to the coreference relations

establishing the identity of several designations of the same entity.

At the sentence level, UMR adds to named entities and words senses that are already in AMR the means to express the information of syntactic character. E.g., a sentence level UMR for the sentence "He denied any wrong-doing" associates the pronoun "he" with the concept *"a person"* that has a *ref-person* attribute with the value *3rd* and a *ref-numb* attribute with the value *Singular*. That is why, as in case of UCCA, UMR may be better qualified as semantic-syntactic representation.

The paper [23] describes an UMR-writer – a Web-based software helping to created te UMR annotations of NL-texts. The paper [24] reports the first release of the UMR dataset consisting of six languages – Chinese, English, Arapaho, Kukavra, Navajo, and Sanapana where the last four are low resources languages that have quite distinct linguistic properties.

The focus of this section is on the significance of main modern branches of formal NL semantics for designing computer systems being able to extract knowledge from the sentences and discourses pertaining to biomedical sciences. In this connection, let's formulate several properties of semantic formalisms seeming to be crucial for the design of NLP systems extracting knowledge from biomedical NL-texts.

Property P1. The possibility to form compound designations of the notions ("an enzyme for converting fibrinogen to fibrin during coagulation", etc.). It should be mentioned that the creators of the system KnowLife indicated in [14] the necessity of representing compound notions.

Property P2. The possibility to construct a definition of a scientific notion connecting an introduced notion with its explanation. For instance, it is the case of the definition D1 = "The genotype of an organism is the collection of all its chromosomes".

Property P3. The availability of effective formal means for building SRs of arbitrarily complex descriptions of the sets. As an example, let's consider the definition D2 = "Type A blood group are the persons who possess type A isoantigen on red blood cells and anti-B agglutinin in plasma".

Property P4. The possibility to build SRs of the infinitives with dependent words or gerundial constructions. Such expressions may express the goals, commitments, the destinations of the processes, wishes, etc. As a example, let's consider the expression "an enzyme which helps to convert fibrinogen to fibrin during coagulation" as a fragment of the definition D3 = "Thrombin is an enzyme which helps to convert fibrinogen to fibrin during coagulation".

Property P5. The possibility to construct SRs of the compound constructions formed from several infinitives with dependent words with the help of the logical connectives NOT, AND, OR.

Property P6. The possibility to build SRs of complex sentences and discourses with the references to the meanings of phrases and larger parts of discourse. For example, this phenomenon is realized in the definition D4 = "All granulocytes are polymorphonuclear; that is, they have multilobed nuclei".

Property P7. The availability of formal means allowing for constructing object-oriented SRs of definitions, that is, formal structures including such slots as the list of the authors of a definition, the year of its creation, a relevant thematic domain, and SR of the definition.

This property is significant for the process of inscribing the constructed SR of a definition into an appropriate ontology.

Table 1 shows whether the approaches UCCA, AMR, UMR, TKR possess the listed properties 1 – 7.

Table 1. The possession of the properties P1 – P7 by the approaches UCCA, AMR, UMR, TKR.

Pro-perty	UCCA	AMR	UMR	TKR
P1	-	-	-	+
P2	-	-	-	+
P3	-	-	-	+
P4	-	-	-	+
P5	-	-	-	+
P6	-	-	-	+
P7	-	-	-	+

The present paper continues the line of the paper [25], where it was shown that much broader prospects in comparison with AMR for creating semantic languages – intermediaries are opened by TKR. The content of Table 1 shows that this conclusion applies also to the approaches UCCA and UMR.

One of the objectives of the present paper is to remind the designers of the systems extracting information from scientific texts for constructing large biomedical knowledge graphs of the powerful and flexible expressive mechanisms of TKR.

3 A new version of a broadly applicable, cross-lingual meaning representation formalism

3.1 A new version of a mathematical model describing primary units of conceptual level and their interrelations

The first constituent of TKR is a mathematical model (Model 1) describing primary units of conceptual level used by an applied intelligent system and the interrelations of these units.

Let's use an analogy for explaining the essence of Model 1. It is known that arbitrary language from the class of first order logic languages (or the languages of the first order predicates logic, or FOL languages) is determined by the choice of non-empty set *Const* of symbols called constants, non-empty set *Var* of symbols called variables,

non-empty set *F* of functional symbols (the names of functions), non-empty set *Pred* of predicate symbols (the names of *n*-ary predicates, where $n \geq 1$), and a mapping *numb-arg* from the union of *F* and *Pred* associating each symbol from this set with a positive integer *k* interpreted as the number of arguments. That is why it is possible to say that arbitrary language from the class of FOL languages is determined by a complex formal object (a five-tuple) *Logbs* = (*Const, Var, F, Pred, numb-arg*), this formal object may be called *a logical basi*s. Generalizing, we can say that arbitrary language from the class of FOL languages is determined by a five-tuple *Logbs* of the form (*c1, c2, c3, c4, c5*), where *c1, c2, c3, c4, c5* are some formal objects.

Formally, the first version of the Model 1 has the form of the definition of a new class of compound formal objects called *conceptual bases* [8]. According to [8], the construction of an arbitrary conceptual basis (c.b.) *B* is equivalent to selecting a finite sequence of the form *c1, c2, …, c15*, where *c1, c2, …, c15* are some formal objects. Each c.b. *B* determines a formal language *Ls(B)* called SK-language (standard knowledge language) in the basis *B*. The expressive mechanisms of SK-languages open broad new prospects for building semantic representations (SRs) of arbitrarily complex sentences and discourses in NL.

The recent version of the Model 1 determines (see the appendix to the present paper) the class of new formal objects called *optimized conceptual bases* (*o.c.b.*). This definition was introduced in [26]. The construction of an arbitrary o.c.b. *Bs* is equivalent to selecting a finite sequence of the form *c1, c2, …, c9*, where *c1, c2, …, c9* are some formal objects. Here *c1 = St* is a finite set of symbols called the sorts and denoting the most general notions from the considered application domains. For instance, *St* may include the elements *phys. ob, org, sit, event* interpreted as the denotations of the notions "physical object", "organization", "situation", "event" (a dynamic situation).

Comparing the number of components 15 of arbitrary c.b. with the number of components 9 of arbitrary o.c.b., it is possible to conclude that, simplifying the Model 1 of TKR, we obtain an optimized version of the theory of SK-languages being the core of TKR. The first basic mathematical model of TKR looks now much simpler, and it will be a great advantage for introducing the university students to TKR.

3.2 A model describing a system of ten partial operations on conceptual structures

The second constituent of TKR is a mathematical model (Model 2) describing a system of such 10 partial operations on structured meanings (SMs) of NL-texts that, using primitive conceptual units as "blocks", we are able to build SMs of arbitrary NL-texts (including articles, textbooks, etc.) and arbitrary pieces of knowledge about the world [8]. These partial operations will be denoted below as Op[1],…, Op[10]. A system consisting of ten partial operations Op[1] – Op[10] is completely mathematically defined in the Chapter 4 of the monograph [8].

Discussing the Model 2, let's mean the updated version of this model. In order to obtain the updated version of the Model 2, it is sufficient to interpret the term "conceptual basis" as the term "optimized conceptual basis" in all definitions and comments of the Chapter 4 of the monograph [8] and to assume that any considered conceptual basis B has the form (*S, Cexp*), where *S* is a sort system, and *Cexp* is an expanded concept-object system coordinated with *S* (see the appendix to the present paper).

Let's consider the algebraic essence of the Model 2. As it is known, a partial algebra on a non-empty set *Y* is any pair *A* of the form (*Y, Z*), where *Z* is a finite set consisting of the partially defined functions from Y^n into *Y*, where $n \geq 1$. The functions from *Z* are called *partial operations on the set Y*. The set *Y* is called *the carrier of the partial algebra A*.

Let *Bs* be an arbitrary optimized conceptual basis, and $L = Ls\ (Bs)$, $L^1 = L$, $L^2 = L \times L$ (the Cartesian square of the language), for $k > 2$, $L^k = L \times L \times \ldots \times L$ (the Cartesian *k*-degree of *L*), and *Sem-carrier (Ls)* be the union of L^k for all $k \geq 1$. Then each partial operation from the list *Op [1],., Op[10]* is an unary (one-argument) partial operation on the set *Sem-carrier(Ls)*. E.g., let w1 be the logical connective ∧ (AND), *w2 = Bratislava, w3 = Ljubljana, w4 = Zagreb*. Then the expression (*Bratislava ∧ Ljubljana ∧ Zagreb*) is the value of the partial operation *Op [7]* on the four-tuple (*w1, w, w3, w4*). This expression may be a component of an SR of the phrase "George would like to visit this summer Bratislava, Ljubljana, and Zagreb".

Let's see (without numerous mathematical details) how the partial operations Op [1] - Op [10] do work.

The operation Op [1] allows us to join intensional quantifiers and designations (simple or compound) of notions, in particular, for constructing the formulas *certain medicine1, certain medicine1 * (Country-manufacturer, France), certain medicine1 * (Country-manufacturer, (France ∨ Italy)), all medicine1 * (Target-disease, malaria).*

The operation Op[2] is used for constructing the expressions of the form $f(t_1, \ldots, t_n)$, and Op[3] enables us to build the expressions of the form ($c \equiv$ d). Examples: *Height (certain person1), Quantity (all medicine1 * (Target-disease, malaria)) and (Height (certain person1) ≡ 183/cm).*

One uses the operation Op[4] for building the expressions of the form $rel(t_1, \ldots, t_n)$, where *rel* is the name of a relation with *n* attributes (example: *Belongs (Thomas-Hunt-Morgan, Creators(genetics))).*

The operation Op[5] provides the possibility to mark a formula or its part by means of a variable. Example: *all medicine1 * (Country-manufacturer, India) : S12.*

The operation Op[6] allows us to join the negation connective ¬ to a formula (example: ¬*antibiotic*).

The operation Op[7] governs the usage of the logical connectives ∧ (and) and ∨ (or). Example: *car1 * (Manufacturer, (BMW ∨ Opel)).*

Using the operation Op[8] at the last step of an inference, it is possible to construct compound

designations of notions. Example: *medicine1 * (Target-disease, malaria) (Country-manufacturer, India).*

The operation Op[9] allows us to use the universal quantifier and existential quantifier ($\forall$ и $\exists$) in formulas. The operation Op[10] enables us to construct the SRs of finite sequences as the strings of the form $< c_1 , ..., c_n >$, where $c_1 , ..., c_n$ are the elements of a sequence.

The first and second constituents of TKR form the theory of SK-languages (standard knowledge languages), stated, in particular, in [8]. As a consequence of defining the notion of optimized conceptual basis, the paper [26] introduced a new, more compact version of the theory of SK-languages.

4 From the theory of SK-languages to the theory of K-representations

4.1 Two formal models of a linguistic database

The third constituent of TKR is formed by two broadly applicable mathematical models of a linguistic database (Model 3a and Model 3b). The Model 3a is oriented at Russian language and is introduced in the monograph [27]. The Model 3b is oriented at Russian, English, and German languages and is described in Chapter 7 of the monograph [8]. Both models describe the frames expressing the necessary conditions of the existence of semantic-syntactic 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”.

4.2 A family of the semantic parsing algorithms

The fourth basic constituent of TKR is formed by a family of complex, strongly structured algorithms carrying out semantic-syntactic analysis (or semantic parsing) of texts from some practically interesting sublanguages of NL. These algorithms transform NL-texts into their semantic representations being K-representations – the SRs being the expressions of SK-languages. The first algorithm called SemSyn is published in Russian in the monograph [27], the input texts of this algorithm form a sublanguage of the Russian language. The next algorithm SemSynt1 is presented in the second part of the monograph [8]. The input texts can be from English, German, and Russian languages. That is why the algorithm SemSynt1 is multilingual.

An important feature of the algorithms SemSyn and SemSynt1 is that they don’t construct any syntactic representation of the inputted NL-text but directly find semantic-syntactic relations between text units. The other distinguished feature is that complex algorithms are completely described with the help of formal means, that is why they are problem independent and don’t depend on a programming system.

The paper [28] introduces a highly compact way of describing formal structure of linguistic databases (semantic-syntactic component) and of presenting the algorithms of semantic parsing. The paper contains the algorithm of semantic parsing SemSyntRA, developed under the framework of the proposed approach.

4.3 Contributions to several branches of computer science

The fifth constituent of TKR is a collection of scientific results expanding theoretical foundations of advanced ontologies, cross-lingual conceptual information access, Multilingual Semantic Web, the design of agent communication languages in multi-agent systems and recording the content of e-negotiations, semantic parsing of irregular NL-texts: the texts with metonymic phrases and metaphoric expressions [7, 8, 11, 15, 25, 26, 29-31].

5 New prospects for describing semantic content of scientific texts

Let’s illustrate only several properties being the principal advantages of TKR by means of constructing semantic representations (SRs) of the definitions D1 – D4 from Section 2.

Property Q1. The possibility to construct SRs of the sets (in essence, it coincides with the property P3 stated in Section 3).

Property Q2. The possibility to construct compound designations of the concepts qualifying the sets (it is a partial case of the property P1).

Example 1. The definition D1 = “The genotype of an organism is the collection of all its chromosomes” from Section 2 (Restriction 2) may have the following SR Semrepr1:

(*Genotype (arbitrary organism: x1) ≡ all gene * (Location, arbitrary chromosome * (Part1, x1))).*

Here the semantic unit *Genotype* is interpreted as the name of a function with one argument.

Example 2. The definition D2 = “Type A blood group are the persons who possess type A isoantigen on red blood cells and anti-B agglutinin in plasma” from Section 2 (Restriction 3) may be associated with the following SR Semrepr2 using a different, general (for the theory of K-representations) structured model:

*Definition (type-A-blood-group, certain set * (Quality-composition, person) : S1, Description(arbitrary person * (Element, S1) : y1, Situation (e1, possessing1 * (Agent1, y1)(Object1, (certain isoantigen * (Type1, ‘A’)(Location, arbitrary cell1 * (Part1, certain blood1 * (Color, red)(Part1, y1))) ^ certain agglutinin * (Type1, ‘anti-B’)(Location, certain plasma1 * (Part1, y1))))).*

Property Q3 (in essence, it coincides with the property P4 stated in Section 3). Contrarily to expressive possibilities of first order logic, UCCA, AMR, and UMR, TKR allows us to build formal semantic analogues of the goals, commitments, etc. expressed by infinitive and gerundial constructions.

Property Q4. The possibility to construct SRs of the notions in the form *conc * (rel$_1$, d$_1$) … (rel$_k$, d$_k$)*, where *conc* is a semantic unit designating a notion, $k \geq 1$, for $m \geq 1$, *rel$_m$* designates either a binary relation or a function with one argument, d_m designates either the second attribute of a binary elation or the value of the function with the name *rel$_m$*.

Property Q5. The possibility to build SRs of the notions' definitions in the form

*(concept1 ≡ concept2 * (rel$_1$, d$_1$) … (rel$_k$, d$_k$))*, where the unit *concept1* designates the notion to be explained, and *concept2* designates the basic notion used in an explanation of *concept1*.

Example 3. Due to the properties 3 - 5, the definition D3 = "Thrombin is an enzyme which helps to convert fibrinogen to fibrin during coagulation" from Section 2 (Restriction 4) may have the following SR Semrepr3:

(*thrombin ≡ enzyme * (Main-function, helping * (Objective-role, converting1 * (Start-matter, certain fibrinogen) (Final-matter, certain fibrin) (Covering-process, certain coagulation))))*.

Property P6. The possibility to build SRs of complex sentences and discourses with the references to the meanings of phrases and larger parts of discourse.

Example 4. The definition D4 = "All granulocytes are polymorphonuclear; that is, they have multilobed nuclei" from Section 1 (Restriction 5) may have the following SR Semrepr4:

(Property (arbitrary granulocyte: x1, polymorphonuclear): P1 ∧ Explanation (P1, Implies (Part1(x1, arbitrary nucleous: x2), Property (x2, multilobed)))).

Property P7. The availability of formal means allowing for constructing object-oriented SRs of definitions, that is, formal structures including such slots as the list of the authors of a definition, the year of its creation, a relevant thematic domain, and SR of the definition.

Example 5. Let D5 = "Control gene is a gene which can turn other genes on or off". Then let

Semrepr5 = (*control-gene ≡ gene * (Is-able, (turning-on * (Object-bio, some gene: Set1) ∧ turning-off * (Object-bio, Set1))))*.

It is possible to construct a different SR of the definition Def5, it will reflect the metadata of information piece, indicating the edition, the authors, and year of publication. In this case

*Semrepr-with-metadata = certain inform-object * (Content1, Semrepr5) (Authorship, (D. Turnpenny ∧ S. Ellard))(Publishing-house, Elsevier)(Year, 2005) (Title, "Emery's Elements of Medical Genetics")(Edition-number, 12)*.

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 of NL-texts (i.e., for representing structured meanings of NL-texts in a formal way).

SK-languages allow also 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 TKR explains the regularities of structured meanings of, likely, arbitrary sentences and discourses pertaining to biomedicine and other fields of professional activity of people.*

6 The principal ideas of processing the notions' definitions by the algorithm AlgSemDef1

The input language of the semantic parsing algorithm SemSynt1 introduced in the monograph [8] and mentioned in the subsection 4.2 doesn't include the definitions of the notions. In order to overcome this restriction, the algorithm AlgSemDef1 is presented in [11]. This algorithm uses a little modified form of the algorithm SemSynt1 as a big procedure; the input language of AlgSemDef1 consists of the scholarly notions' definitions having often encountered syntactic structures.

6.1 Input language of an algorithm of definitions' semantic parsing

Let's characterize the input language of the algorithm AlgSemDef1 by means of several examples.

Example 1. Let Def1 = "The Eustachian tube is a canal that leads from the middle ear to the pharynx".

Example 2. Let Def2 = "The Eustachian tube is a canal leading from the middle ear to the pharynx".

Example 3. Let Def3 = "A silent mutation is a mutation not altering the polypeptide product of the gene".

Example 4. Let Def4 = "Pyramid system is the principal efferent pathway of the cortex transmitting the movement impulses, originating in the forward central gyrus of the cortex and reaching the motor neurons of the spinal cord".

Example 5. Let Def5 = "A jack is a technical device for moving physical objects in the vertical plane".

Let's say that the definition Def1 has the type 1, Def2 – Def4 have the type 2, and the definition Def5 has the type 3.

While processing a definition of the type 2, our algorithm initially transforms it into the synonymic definition of the type 1. The meaning of the definition Def5 is the same as of the definition Def 6 = "A jack is a technical device moving physical objects in the vertical plane". That is why, while processing a definition of the type 3, our algorithm will transform it into the synonymic definition of the type 1.

Consider the basic assumptions about the input definitions.

Assumption 1. While elaborating an algorithm of definitions' semantic parsing, we consider not real tokens (nouns, articles, verbs, etc.) but the so called elementary meaningful text units: "a gene", "a technical device", "the piston", "has received", etc.

Example. Consider the definition Def1 = "The Eustachian tube is a canal that leads from the middle ear to the pharynx". This definition will be regarded as the sequence $t_1\ t_2\ \ldots\ t_n$, where t_1 = "The Eustachian tube", t_2 = "is", t_3 = "a canal", t_4 = "that", t_5 = leads", t_6 = "from", t_7 = "the middle ear", t_8 = "to", t_9 = "the pharynx", t_{10} = dot.

Assumption 2. Let's suppose that all input definitions have the form *t1 t2 t3 descr*, where t*1* is a lexical representation (LR) of the notion to be explained, *t2* is the word "is", t3 is a LR of the basic notion, *descr* is a fragment describing the properties possessed by the objects qualified by the notion to be explained.

6.2 The central ideas of definitions' semantic parsing

The monograph [8] introduces an original, strongly structured algorithm of semantic parsing (in other terms, an algorithm of semantic-syntactic analysis) called the algorithm SemSynt1. This algorithm is multilingual: the input texts may belong to restricted sublanguages of English, German, and Russian (in the last case the texts in Latin transcription are considered). The input texts may be the questions of many kinds, the statements, and the commands. The algorithm SemSynt1 transforms an input text into its semantic representation being its K-representation (knowledge representation), i.e., an expression of a certain SK-language.

The algorithm of semantic parsing SemSynt1 [8] provided by TKR opens broad prospects of developing useful for practice algorithms for semantic processing of knowledge pieces expressed in NL, in particular, of the definitions of the notions.

Example. Let's consider the central ideas of processing the definition Def4 = "Pyramid system is the principal efferent pathway of the cortex transmitting the movement impulses, originating in the forward central gyrus of the cortex and reaching the motor neurons of the spinal cord" by the algorithm AlgSemDef1 introduced in [11].

Let's assume that the knowledge base includes the units *pyramid-system* and *pathway1*. The definition Def4 contains the verb produced forms "transmitting", "originating", and "reaching". That is why the variable *nvb* (the number of verb produced forms) receives the value 3. The execution of the algorithm AlgSemDef1 includes the construction and semantic parsing of the auxiliary sentences Sent1 = "The principal efferent pathway of the cortex transmits the movement impulses", Sent2 = "The principal efferent pathway of the cortex originates in the forward central gyrus of the cortex", and Sent3 = "The principal efferent pathway of the cortex reaches the motor neurons of the spinal cord".

In the cycle for *i* from 1 to *nvb*, the variable *phrase* receives the value Sent1 in case i = 1, the value Sent2 in case i = 2, and the value Sent3 in case i = 3.

The one-dimensional array Sem1 with the elements being strings is used for storing the primary semantic representations (SRs) of the sentences Sent1, Sent2, and Sent3. These primary SRs are constructed by the algorithm SemSynt1 (in a little modified form). The following assignments will be fulfilled (in the context of a certain linguistic database) in the cycle for *i* from 1 to *nvb* (the element *certn* is an intensional quantifier corresponding to the meaning of the word combination "a certain", the element *Qual-compos* denotes the binary relation "Qualitative composition of a set"):

Sem1[1] := *Situation(e1, trasmission1 * (Mediator1, certn pathway1 * (Importance1, principal)(Quality1, efferent)(Part, certn cortex1 : x2) : x1)(Object1, certn set * (Qual-compos, impulse1 * (Purpose1, movement)) : S3)),*

Sem1[2] := *Situation(e1, originating * (Focus-entity, certn pathway1 * (Importance1, principal)(Quality1, efferent)(Part, certn cortex1 : x2) : x1)(Departure-entity, certn gyrus1 * (Space-property1, forward)(Space-property2, central)(Part, certn cortex1 : x4) : x3)),*

Sem1[3] := *Situation(e1, reaching1 * (Focus-entity, certn pathway1 * (Importance1, principal)(Quality1, efferent)(Part, certn cortex1 : x2) : x1)(Destination-entity, certn set * (Qual-compos, neuron1 * (Purpose1, setting-in-motion)(Part, certn spinal-cord : x4) : S3)).*

The purpose of the next stage of executing the algorithm AlgSemDef1 is to improve the referents of various entities mentioned in the sentences Sent1, Sent2, and Sent3. Firstly, it is easy to see that the variables x2 in the expression Sem1[1] and the variable x4 in the expression Sem1[2] are the marks of the same cortex. That is why the variable x4 is to be replaced by the variable x2 in the expression Sem1[2]. Secondly, the variable S3 denotes a set of movement impulses in Sem1[1] and a set of motor neurons in Sem1[3]. That is why the variable S3 is to be replaced by the variable S4 in the expression Sem1[3].

Taking into account these considerations, let's fulfill a number of actions. Let *Concrepr2* be an expression obtained from the expression *Sem1[2]* by means of replacing the variable *e1* by *e2* and the variable *x4* by the variable *x2*. Let *Concrepr3* be an expression obtained from the expression *Sem1[3]* by means of replacing the variable *e1* by *e3* and the variable *S3* by the variable *S4*. Then let

Sem2[1] := Sem1[1], Sem2[2] := Concrepr2, Sem2[3] := Concrepr3.

Thus, the array Sem2 will have the configuration reflected on Table 2.

Table 2: The configuration of the array Sem2 constructed from the input definition Def4 by the algorithm AlgSemDef1.

*Situation (e1, trasmission1 * (Mediator1, certn pathway1 * (Importance1, principal) (Quality1, efferent) (Part, certn cortex1 : x2) : x1)(Object1, certn set * (Qual-compos, impulse1 * (Purpose1, movement)) : S3))*
*Situation (e2, originating * (Focus-entity, certn pathway1 * (Importance1, principal) (Quality1, efferent) (Part, certn cortex1 : x2) : x1)(Departure-entity, certn gyrus1 * (Space-property1, forward)(*

Space-property2, central)(Part, certn cortex1 : x2) : x3))
*Situation (e3, reaching1 * (Focus-entity, certn pathway1 * (Importance1, principal) (Quality1, efferent) (Part, certn cortex1 : x2) : x1)(Destination-entity, certn set * (Qual-compos, neuron1 * (Purpose1, setting-in-motion)(Part, certn spinal-cord : x4) : S4))*

Then the algorithm AlgSemDef1 builds a K-representation of Def4 as the formula *Sem-final* of the form *Definition* (*pyramid-system, pathway1, x1,* (*Sem2[1] ∧ Sem2[2] ∧ Sem2[3]*)).

This formula is interpreted as follows: an arbitrary *pyramid system* is denoted by the variable *x1*, the physical object x1 is a pathway1 (a pathway in a biological object), and the formula (*Sem2[1] ∧ Sem2[2] ∧ Sem2[3]*) describes the properties of the physical object *x1*.

7 Improvement of the algorithm AlgSemDef1

Analysing the structure of the one-dimensional array Sem2 constructed by the algorithm AlgSemDef1 from the input definition Def4 (see Table 2), we see that each of the elements Sem2[1], Sem2[2], Sem2[3] includes the same fragment being a semantic image of the word combination "the principal efferent pathway of the cortex". It is the expression

*certn pathway1 * (Importance1, principal) (Quality1, efferent) (Part, certn cortex1 : x2) : x1.* (1)

Let's develop a simple algorithm Improve-actants replacing by x1 all occurences of a value associated with the first realized semantic role in the expressions Sem2[m], where m>1. In case of processing the definition Def4 (it explicates the notion "a pyramid system"), this algorithm will replace the second and third occurences of the expression (1) by the variable x1.

Let's agree that the elements of the considered informational universe X(B) and the variables from the set V(B), where B is a conceptual basis, will be interpreted below as symbols.

Algorithm Improve-actants
External specification

Input: one-dimensional array Sem2 containing the strings.
Output: transformed one-dimensional array Sem2.

Algorithm

Begin
Cycle for k from 2 to nvb
Begin
Let p be the position of the first occurrence (from the left) of the variable x1 in Sem2[k]. Then delete the symbols in the positions 10, 11, p-1 of the expression Sem2[k]
End

Example. Let Sem2 be the array constructed by the algorithm AlgSemDef1 for the input Def4 (see Fig. 1). Then the correspondence between the positions 1, 2, …, 12 and the symbols in these positions forming the string Sem2[2] is shown on Table 3.

Table 3: A correspondence between the positions of a left fragment of Sem2[2] and the elements (the symbols) in these positions.

position	element
1	Situation
2	(
3	e2
4	,
5	originating
6	*
7	(
8	Focus-entity
9	,
10	certn
11	pathway1
12	*

Then the first three symbols deleted by the algorithm *Improve-actants* will be the elements *certn*, *pathway1*, *. Let p be the position of the first occurrence of the variable x1 in the expression Sem2[2]. Then the last deleted symbol will be the colon in the position p-1 (the left neighbour of the element x1 in the expression Sem2[2]).

The algorithm AlgSemDef1 is the sequence of the auxiliary algorithms FirstStage, Rename-variables, and FinalStep. *Then let's define the algorithm AlgSemDef2 as the sequence of the auxiliary algorithms FirstStage, Rename-variables, FinalStep, and Improve-actants.*

Example. Processing the input definition Def4 (an explication of the notion "a pyramid system"), the algorithm AlgSemDef2 will construct the array Sem2 whose configuration is reflected on Table 4.

Table 4: The configuration of the array Sem2 constructed from the input definition Def4 by the algorithm AlgSemDef2.

*Situation (e1, trasmission1 * (Mediator1, certn pathway1 * (Importance1, principal) (Quality1, efferent) (Part, certn cortex1 : x2) : x1)(Object1, certn set * (Qual-compos, impulse1 * (Purpose1, movement)) : S3))*
*Situation (e2, originating * (Focus-entity, x1) (Departure-entity, certn gyrus1 * (Space-property1, forward) (Space-property2, central) (Part, certn cortex1: x2) : x3))*
*Situation (e3, reaching1 * (Focus-entity, x1) (Destination-entity, certn set * (Qual-compos, neuron1 * (Purpose1, setting-in-motion) (Part, certn spinal-cord: x4) : S4))*

Then the algorithm AlgSemDef2 builds a K-representation of Def4 as the formula *Sem-final* of the form *Definition* (*pyramid-system, pathway1, x1,* (*Sem2[1] ∧ Sem2[2] ∧ Sem2[3*)), that is, of the form

Definition (*pyramid-system, pathway1, x1,* (*Situation(e1, trasmission1 * (Mediator1, certn pathway1 * (Importance1, principal)(Quality1, efferent)(Part, certn cortex1 : x2) : x1)(Object1, certn set * (Qual-compos, impulse1 * (Purpose1, movement)) : S3)) ∧ Situation(e2, originating * (Focus-entity, x1)(Departure-entity, certn gyrus1 * (Space-property1, forward)(Space-property2, central)(Part, certn cortex1 : x2) : x3)) ∧ Situation(e3, reaching1 * (Focus-entity, x1)(Destination-entity, certn set * (Qual-compos, neuron1 * (Purpose1, setting-in-motion)(Part, certn spinal-cord : x4) : S4)))).*

8 Experimental validation of the algorithm AlgSemDef2

The suggested algorithm AlgSemDef2 is implemented in the computer program developed under the framework of the master thesis [32]. The program is realized by means of the language Java and uses the system PostgreSQL as a database management system. The program executes a two-step transformation of a knowledge piece description in Russian. The first step is fulfilled by the algorithm 1. Its content is the transformation "NL description of a knowledge piece => K-representation". The second step (fulfilled by the algorithm 2) is the transformation "K-representation of a knowledge piece => A collection of the records in the language OWL (Ontology Web Language)", the goal is to inscribe the created collection of the OWL-records into an OWL-based ontology.

The algorithm 1 includes the algorithm AlgSemDef2 as a complex procedure, but the input language is broader than in case of the algorithm AlgSemDef2. Any input of the algorithm 1 has the form *Def Addition 1…. Addition N*, where $N \geq 1$, the fragment *Def* is a definition of a notion *conc*, $N \geq 1$, the fragment *Addition 1…. Addition N* states additional information about the objects qualified by the notion *conc.* In particular, the input language of the algorithm 1 includes the knowledge piece "Crocodile is an animal from the order of water reptiles. The crocodiles live in the rivers, lakes, and marches. Geographical zone is tropics. The length may reach 7 m".

9 Conclusion

It is shown above that the theory of K-representations (TKR) provides a number of precious expressive mechanisms for describing in a formal way semantic structure of scholarly sentences and discourses in NL pertaining to biology and medicine. With this aim, TKR introduced (in several previous publications) a new class of formal languages – the class of SK-languages (standard knowledge languages). TKR suggests to use this class of languages for building semantic representations (SRs) both of sentences and discourses. The main attention is paid to the advantages of TKR in comparison with Universal Conceptual Cognitive Annotation, Abstract Meaning Representation, and Uniform Meaning Representation.

The are serious reasons to conjecture that the considered optimized version of the SK-languages' theory (a part of TKR) may be interpreted as the *starting universal meaning representation* (*MR*) *formalism*. In this connection, the new, very topical task for the computer linguists of many countries is to select the collections of semantic units being convenient for forming MRs of sentences and discourses in their native languages, using ten rules of the SK-languages' theory for combining these semantic units into (a) MRs of sentences and discourses, (b) conceptual structures emerged in the memory of applied intelligent systems (AIS) during a dialogue with a human being or another AIS.

The topicality of the problem of forming and updating ontologies by means of automatic extraction of knowledge from NL-texts motivated the development of an algorithm of building SRs of the scholarly notions' definitions (it was published in a previous work of the present paper's author and is called AlgSemDef1). TKR underpinned the creation of the algorithm AlgSemDef1. The present paper improves the algorithm AlgSemDef1 and defines the algorithm AlgSemDef2. The purpose is to make more compact the output of the algorithm. The main steps of processing complex scholarly definitions pertaining to biology and medicine by the algorithms AlgSemDef1 and AlgSemDef2 are illustrated.

Thus, it is shown above that TKR opens new precious prospects for the designers of NL processing systems in biomedical sciences.

Acknowledgement

I am grateful to the anonymous referees for precious comments helped to improve this paper.

Appendix: The definition of optimized conceptual basis

The definitions considered below were initially introduced in [26].

A. Sort systems

Let's imagine that we should start the formalization of a certain application domain or a group of domains. Our aim will be the ability to formally represent semantic structure of arbitrary NL-texts used in the considered domain. At the first step, let's select a finite set *St* of sorts and distinguish in *St* a certain sort *P* to be called the sort "a meaning of proposition" (in the sense "a meaning of statement"). This sort *P* will be interpreted as the mark (the type) of semantic representations (SRs) of sentences being statements and of arbitrary narrative discourses.

Let's suppose that the content of next two steps is to define two binary relations *Gen* and *Tol* on the set *St*. It means that we define two subsets of the Cartesian product $St \times St$. The relation *Gen* formalizes the hierarchy on the set of most general notions *St*. Our requirement will be that it is a reflexive, antisymmetric, and transitive relation (i.e., a partial order on *St*). Due to reflexivity, for arbitrary sort w, the pair (*w, w*) belongs to *Gen*. The pair (*real, integer*) belongs to *Gen* means that all integers form a subclass of all real numbers.

The binary relation *Tol* is to formalize the phenomenon of the existence of two or more different angles of look at some objects. For instance, a scholar with usual physical capabilities may be considered, on the one hand, as an intelligent system and, on the other hand, as a dynamic physical object (because she/he may run, spring, etc.). Taking this into account, the pair (*ints, dyn. phys. ob*) belongs to *Tol* in case the sorts *ints* and *dyn. phys. ob* are interpreted respectively as "an intelligent system" and "a dynamic physical object".

Definition 1 (introduced in [6]). A sort system is an arbitrary four-tuple *S* of the form (*St, P, Gen, Tol*), where *S* is an arbitrary finite set of symbols, *P* belongs to *St, Gen* is a non-empty binary relation on *St* being a partial order on *St* (a reflexive, antisymmetric, and transitive relation), *Tol* is a binary relation on *St* being antireflexive and symmetric, and the following conditions are satisfied:

(1) *St* doesn't include the symbols ↑, *, {,}, (,), [*entity*], [*concept*], [*object*], [↑*entity*], [↑*concept]*, [↑*object*].

(2) If *Concretizations(P)* is the set of all such *z* from the set *St* that (*P, z*) belongs to *Gen* then the set-theoretical difference *St* \ *Concretizations(P)* is not empty, and for every *u* from *St* \ *Concretizations(P)* and for every *w* from *Concretizations(P)* the sorts *u* and *w* are incomparable both for the relation *Gen* and for the relation *Tol*.

(3) For each *t, u* from *St*, it follows from (*t, u*) belongs to *Gen* or (*u, t*) belongs to *Gen* that *t* and *u* are incomparable for the relation *Tol*.

(4) For each *t1, u1* from *St* and each *t2, u2* from *St*, it follows from (*t1, u1*) belongs to *Tol*, (*t2, t1*) belongs to *Gen*, (*u2, u1*) belongs to *Gen* that (*t2, u2*) belong to *Tol*.

The elements of the set *St* are called sorts, *P* is called the sort "a meaning of proposition", the binary relations *Gen* and *Tol* on the set *St* are called respectively the *generality relation* and *tolerance relation*. If *t, u* belong to *St* and the pair (*t, u*) belongs to *Gen*, then we often use an equivalent notation *t* →*u* and say that *t* is a generalization of *u*, and *u* is a concretization of *t.*

The symbols listed in the requirement (1) play special roles in TKR. The requirement (2) is to be interpreted as follows. The sort *P* will play the role of the type (the mark) of semantic representations (SRs) of statements and narrative texts. Suppose that we have a pragmatic reason to introduce the particular cases of the sort *P:* the sort *P-act* as the type of the statements about the physical actions, *P-inf* as the type of the statements about information transmissions, and *P-comm* as the type of the statements being implicit commands ("It is necessary to close the dor", etc.). Then *Concretizations (P)* = {*P, P-act, P-inf, P-comm*}.

But we should have different sorts (let they form a subset *Usual-things-sorts* of the set *St*) for classifying things to be mentioned in the statements of the mentioned kinds. Obviously, for every *u* from the subset *Usual-things-sorts* and for every *w* from *Concretizations(P)* the sorts *u* and w should be incomparable both for the relation *Gen* and for the relation *Tol*.

B. The set of types corresponding to a sort system

The monograph [8] contains a definition associating arbitrary sort system *S* with the countable set *Tp(S)* of strings called types and interpreted as simple and compound semantic characteristics of the entities from the considered application domains.

Example. There is such sort system S_1 that the set $Tp(S_1)$ includes, in particular, the elements *ints * phys.ob*, {*ints * phys.ob*}, ↑*ints * phys.ob*, *integer*, {*integer*}, (integer, integer), {(*integer, integer*)}. The interpretation of these types will be considered below in an example immediately after the definition 2.

The arrow ↑ distinguishes the types of the notions; the types of the form {*z*} are the types of the sets; the types of the forms (*x, y*), (*x, y, z*) are respectively the types of a pair and of a triple, etc. The type {(*integer, integer*)}is to be interpreted as the type of various binary relations on the set of all integers; in particular, as the type of the relations < and >. The form of the type *ints * phys.ob* says that the pair (*ints, phys.ob*) belongs to *Tol.*

C. Expanded concept-object systems

Definition 2. Let *S* be any sort system of the form (*St, P, Gen, Tol*). Then a five-tuple *Cexp* of the form (*X, V, tp, F, ref*) is called an *expanded concept-object system* (*e.c.o.s.*) *coordinated with the sort system S* then and only then the following conditions are satisfied:

X and *V* are countable non-intersecting sets of symbols; *tp* is a mapping from the union of *X* and *V* into the set of types *Tp(S)*;

F is a subset of *X*; for each *h* from *F*, the string *tp(h)* has the beginning {(and the ending)};

The set of sorts *St* is a subset of *X*, and for each *w* from *St*, *tp(w)* = ↑*w; ref* is a distinguished element of the set-theoretical difference *X* \ *Y*, where *Y i*s the union of *St* and *F*;

The set of all such elements *var* from *V* that *tp(var)* = [*entity*] is countable.

The set *X* is called the *primary informational universe*; the elements of *V* are called *variables*. The elements of the subset *F* are called *functional symbols* (or the names of functions). The distinguished element *ref* is called the *referential quantifier*.

The referential quantifier *ref* is interpreted as the semantic unit designating the meaning of the expression "a certain". In the examples in English usually *ref* is the string *certain* or *certn*.

The mapping *tp* gives us a much more fine-grained structuring of application domains than first order logic.

Example. The sets *St, X* and the mapping *tp* may satisfy the following conditions: (a) *St* includes the elements (sorts) *dyn.phys.ob* (dynamic physical object), *ints* (intelligent system), *org* (organization), *inf.ob* (informational object); (a) *X i*ncludes the elements

L.Carroll, Alice-in-Wonderland, person, student-group, Suppliers, Authorship, and

*tp(person) =↑ ints * dyn.phys.ob, tp(L.Carroll) = ints * dyn.phys.ob; tp(Alice-in-Wonderland) = inf.ob, tp(Authorship) = {(ints, inf.ob)};*

*tp(student-group) = =↑ {ints * dyn.phys.ob}, tp(Suppliers) = {(org, {org})}.*

Here the symbol ↑indicates the type of a notion; *Suppliers* is the name of the function associating an enterprise with the set of all its suppliers.

D. Optimized conceptual bases

Definition 3. Let *S* be any sort system of the form (*St, P, Gen, Tol*), and a five-tuple *Cexp* of the form (*X, V, tp, F, ref*) be an expanded concept-object system coordinated with the sort system *S*. Then the ordered pair *B* = (*S, Cexp*) will be called an optimized conceptual basis if and only if the following conditions are satisfied:

The sets *X* and *V* don't include the symbols ',' (comma), '*', ':', '(', ')', '<', '>', '&';

The set of sorts *St* includes the symbols *eqv, neg, binlog, int1, int2, ext*; the primary informational universe *X* includes the universal and existential quantifiers and, besides, the symbols ≡, ¬ , ∨ , ∧;

The union of the set *Int1={y from X| tp(y) = int1}* and the set *Int2={w from X| tp(w) = int21}* belongs to the set-theoretical difference *X \ Y*, where *Y* is the union of *St* and *F*; *tp(ref) = int1*;

tp(≡) = *eqv*; tp(¬) = *neg*; tp(∨) = tp(∧) = *binlog*; the value of the mapping *tp* for the universal quantifier and existential quantifier is equal to ext.

The elements of the sets *Int1* and *Int2* will be called *intensional quantifiers.*

The symbols ≡, ¬ , ∨ , ∧ should be read as "identical to", "not", "or", "and". The elements of the set *Int1* are interpreted as semantic units denoting the meanings of the words and expressions "a certain", "any", "arbitrary", etc. The elements of the set *Int2* are interpreted as semantic units denoting the meanings of the words and expressions "all", "several", "many", "a few", etc.

Using these denotations and interpreting in Chapter 4 of [8] the symbol *B* as an optimized conceptual basis (but not as a conceptual basis), we are able to get the definition of the SK-language *Ls(B)* in the basis *B*. Thus, due to the simplification of the basic Model 1 of TKR, we obtain a simplified (or optimized) version of the theory of SK-languages.

References

[1] Van Gysel, J. E. L., Vigus, M. et al. (2021). Designing a Uniform Meaning Representation for Natural Language Processing. *Kuenstliche Intelligenz*, vol. 35, no. 3-4, pp. 1-18. https://doi.org/10.1007/s13218-021-00722-w

[2] Bonn, J., Flanigan, J., Hajic, J., Jindal, I., Li, Y., and Xue, N. (2024). Meaning Representations for Natural Languages: Design, Models and Application. In *Proceedings of the 2024 Joint Intern. Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 20-25 May 2024*, Language Resources Association (ELRA), 2024, pp. 13-18; https://aclanthology.org/2024.lrec-tutorials.3.pdf

[3] DMR-1 Workshop (2019). *Proceedings of DMR 2019: The First International Workshop on Designing Meaning Representation*. Italy, Florence, August 1, 2019 in conjunction with ACL 2019; https://aclanthology.org/W19-33.pdf

[4] DMR-5 Workshop (2024). *Proceedings of DMR 2024: The Fifth International Workshop on Designing Meaning Representation*, Italy, Torino, May 21, 2024 in conjunction with 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), Language Resources Association (ELRA), 2024; https://aclanthology.org/2024.dmr-1.0

[5] Zhu, H., Li, Y. and Chiticariu, L. (2019). Towards universal semantic representation. In *Proceedings of DMR 2019: The First International Workshop on Designing Meaning Representation, Italy, Florence, August 2019 in conjunction with ACL 2019*, pp. 177-181. https://doi.org/10.18653/v1/w19-3320

[6] Fomichov, V.A. (1996). A Mathematical Model for Describing Structured Items of Conceptual Level. *Informatica. An International Journal. of Computing and Informatics (Slovenia)*, vol. 20 no. 1, pp.15-32.

[7] Fomichov, V.A. (2008). A Comprehensive Mathematical Framework for Bridging a Gap between Two Approaches to Creating a Meaning-Understanding Web. *Intern. Journal of Intelligent Computing and Cybernetics*, vol. 1 no. 1, pp. 143-163. https://doi.org/10.1108/17563780810857176

[8] Fomichov, V.A. (2010a). *Semantics-Oriented Natural Language Processing: Mathematical Models and Algorithms.* Springer, New York, Dordrecht, Heidelberg, London, 352 p. https://doi.org/10.1007/978-0-387-72926-8

[9] Abend, O. and Rappoport, A, (2013). Universal Conceptual Cognitive Annotation (UCCA). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, August 4-9 2013. Association for Computational Linguistics, 2013, pp. 228-238; https://aclanthology.org/P13-1023

[10] Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M. and Schneider, N. (2013). Abstract Meaning Representation for Sembanking. *Proceedings of the 7th ACL Linguistic Annotation*

Workshop and Interoperability with Discourse, Sofia, Bulgaria, August 8-9, 2013 (www.aclweb.org/anthology/W13-2322; retrieved 2024-07-30)

[11] Fomichov, V.A. (2022). Semantic Mapping of Definitions for Constructing Ontologies of Business Processes. In *2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO), May 23 - 27, 2022, Opatija, Croatia. Proceedings*. Croatian Society for Information, Communication and Electronic Technology - MIPRO, Rijeka, 2022, pp. 1258-1263. https://doi.org/10.23919/mipro55190.2022.9803553

[12] Auer, S., Bryl, V. and Tramp, S. (Editors) (2014). *Linked Open Data -- Creating Knowledge Out of Interlinked Data. Results of the LOD2 Project*, Springer, 2014; open access at https://link.springer.com/book/10.1007/978-3-319-09846-3. https://doi.org/10.1007/978-3-319-09846-3

[13] Blaney, J. (2017). Introduction to the Principles of Linked Open Data. *Programming Historian*, 2017; open access at https://programminghistorian.org/en/lessons/intro-to-linked-data. https://doi.org/10.46430/phen0068

[14] Ernst, P., Siu, A., and Weikum, G. (2015). KnowLife: a versatile approach for constructing a large knowledge graph for biomedical sciences. *BMC Bioinformatics*, 16, 157 (2015); open access at https://doi.org/10.1186/s12859-015-0549-5 (retrieved 2024-07-30).

[15] Fomichov, V. A. (2011). The prospects revealed by the theory of K-representations for bioinformatics and Semantic Web. *Actes de la 18e conference sur le Traitement Automatique des Langues Naturels. Actes de la 15e Rencontre des Etudiants Cercheurs en Informatique pour le Traitement Automatique des Langues. France, Montpellie, 27th June - 1st July 2011 Vol. 1: Actes: articles longs*, AVL Diffusion, Montpellier, pp 5-20; https://aclanthology.org/2011.jeptalnrecital-invite.1.pdf

[16] Wang, C., Xue, N., and Pradnan, S. (2015). A transition-based algorithm for AMR parsing. In *Proc. of the 2015 Conf. of the North American Chapter of ACL: Human Language Technologies. Denver, Colorado*, ACL, 2015, pp. 366-375. https://doi.org/10.3115/v1/n15-1040

[17] Flanigan, J., Thomson, S., Carbonell, J., Dyer, C., and Smith, N.A. (2014). A discriminative graph-based parser for the abstract meaning representation". In *Proc. of the 52nd Annual Meeting of ACL, Vol. 1: Long Papers*, Baltimore, Maryland, ACL, 2017, pp. 1426-1436. https://doi.org/10.3115/v1/p14-1134

[18] Pust, M., Hermjakob, U., Knight, K. , Marcu, D. and May, J. (2015). Parsing English into abstract meaning representation using syntax-based machine translation. In *Proc. of the 2015 Conf. on Empirical Methods in Natural Language Processing. Lisbon, Portugal*, ACL, 2015, pp. 1143-1154. https://doi.org/10.18653/v1/d15-1136

[19] May, J. (2016). SemEval-2016 task 8: Meaning representation parsing, In *Proc. of the 10th Intern. Workshop on Semantic Evaluation (SemEval-2016)*. ACL, San Diego, California, 2016, pp. 1063-1073. https://doi.org/10.18653/v1/s16-1166

[20] Damonte, M., Cohen, S.B., and Satta, G. (2017). An incremental parser for abstract meaning representation. *Proc. of the 15th Conference of the European Chapter of ACL, April 2017, vol. 1, Long Papers*, Valencia, Spain, ACL, 2017, pp. 536-546. https://doi.org/10.18653/v1/e17-1051

[21] Rao, S., Marcu. D. and Hal Daume III. (2017). Biomedical event extraction using Abstract Meaning Representation. In *Proc. of the Intern. Conference BioNLP 2017, Vancouver, Canada, Aug. 2017.* ACL, 2017, pp. 126-135. https://doi.org/10.18653/v1/w17-2315

[22] Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M. and Schneider, N. (2019). *Abstract Meaning Representation (AMR) 1.2.6 Specification. May 1, 2019*; github.com/amrisi/amr-guidelines/blob/master/amr.md (retrieved 2024-07-30).

[23] Zhao, J., Xue, N., Van Gysel, J., and Choi, J. D. (2021). UMR-Writer: A Web Application for Annotating Uniform Meaning Representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Nov. 7-11, 2021.* Association for Computational Linguistics, 2021, pp. 160-167. https://doi.org/10.18653/v1/2021.emnlp-demo.19

[24] Bonn, J., Buchholz, M., et al. (2024). Building a Broad Infrastructure for Uniform Meaning Representation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 20-25 May 2024*. Language Resources Association (ELRA), 2024, pp. 2537-2547; https://aclanthology.org/2024.lrec-main.229.pdf

[25] Fomichov, V.A. (2017). SK-languages as a Powerful and Flexible Semantic Formalism for the Systems of Cross-Lingual Intelligent Information Access. *Informatica. An Intern. J. of Computing and Informatics (Slovenia)*, 2017, vol. 41, pp. 221-232.

[26] Fomichov, V.A. (2021). Intelligent Monitoring of News on Economics and Finance Based on Formal Semantics of the Movement Verbs. In *2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO), September 27 – October 1, 2021, Opatija, Croatia. Proceedings*, Croatian Society for Information, Communication and Electronic Technology - MIPRO, Rijeka, 2021, pp. 1253-1258. https://doi.org/10.23919/mipro52101.2021.9597096

[27] Fomichov, V.A. (2005). *The Formalization of Designing Natural Language Processing Systems*, Moscow, MAKS Press, 368 p. (in Russian).

[28] Razorenov, A. A. and Fomichov, V. A. (2016). A new formal approach to semantic parsing of instructions and to file manager design Database and Expert Systems Applications. In *Proceedings of the 27th International Conference, DEXA 2016, Porto, FomichovInformaticaCameraReadyPaper2024fin1.dotxPortugal, September 5-8, 2016*, Springer, Cham, 2016, vol. 9827. Part. I, pp. 416-430. https://doi.org/10.1007/978-3-319-44403-1_26

[29] Fomichov, V.A. (2010b). Theory of K-representations as a Comprehensive Formal Framework for Developing a Multilingual Semantic Web. *Informatica. An Intern. Journal of Computing and Informatics*, 2010, vol. 34, pp. 387-396.

[30] Fomichov, V.A. (2013). A Broadly Applicable and Flexible Conceptual Metagrammar as a Basic Tool for Developing a Multilingual Semantic Web. In *Metais, E., Meziane, F., Saraee, M., Sugumaran, V., Vadera, S. (eds.) NLDB 2013. LNCS, 2013*, vol. 7934. Springer, Heidelberg. 2013, pp. 249-259. https://doi.org/10.1007/978-3-642-38824-8_21

[31] Fomichov, V.A. (2023). A New Approach to Semantic Parsing of Metonymic Phrases by a Business Intelligence System. In *2023 46th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), May 22 - 26, 2023, Opatija, Croatia. Proceedings*, Croatian Society for Information, Communication and Electronic Technology - MIPRO, Rijeka, 2023, pp. 1305-1310. https://doi.org/10.23919/mipro57284.2023.10159720

[32] Urubkov, V. S. (2024). Development of a Semantic Analyzer of Natural Language Knowledge Descriptions. Master thesis (Academic Advisor Prof. V. A. Fomichov). Moscow Aviation Institute (National Research University), 2024.