A Multilingual Algorithm of Texts’ Semantic-Syntactic Analysis for Adaptive Planning Systems

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

vdrfom@aha.ru and vfomichov@hse.ru

Abstract

The natural language texts (NL-texts) from the newspapers, e-mail lists, various blogs, etc. are the important sources of information being able to stimulate the elaboration of a new plan of actions. The paper describes a new formal approach to developing multilingual algorithms of semantic-syntactic analysis of NL-texts. It is a part of the theory of K-representations - a new theory of designing semantic-syntactic analyzers of NL-texts with the broad use of formal means for representing input, intermediary, and output data. The current version of the theory is set forth in a monograph published by Springer in 2010. One of the principal constituents of this theory is a complex, strongly structured algorithm SemSynt1 carrying out semantic-syntactic analysis of texts from some practically interesting sublanguages of the English, German, and Russian languages. An important feature of this algorithm is that it doesn’t construct any syntactic representation of the inputted NL-text but directly finds semantic relations between text units. The other distinguished feature is that the algorithm is completely described with the help of formal means, that is why it is problem independent and doesn’t depend on a programming system. The peculiarities and some central procedures of the algorithm SemSynt1 are analyzed.

Keywords

Semantics-oriented natural language processing; semantic representation; theory of K-representations; formal model of a linguistic database; SK-languages; multilingual algorithm of semantic-syntactic analysis

Introduction

An important source of information being able to stimulate the elaboration of a new plan of actions are the natural-language texts (NL-texts) from newspapers, e-mail lists, various blogs, etc. There are numerous situations when the information being able to change a plan of actions can be obtained from the sources in several natural languages. For instance, it is the case of planning the delivery of the loads across different countries with several languages. It would be very expensive to develop for each concrete language of possible interest a separate conceptual information...
retrieval system with the ability of understanding just this particular language. That is why during last years many researchers have indicated the necessity of designing multilingual algorithms of semantic-syntactic analysis of NL-texts (see, e.g., (Wilks and Brewster 2006)).

In the monograph (Fomichov 2010) a new theory of designing multilingual semantic-syntactic analyzers of NL-texts with the use of formal means for representing input, intermediary, and output data is proposed. This theory is called the theory of K-representations (knowledge representations). Let’s consider its structure.

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

The expressions of SK-languages will be called the K-strings. If $Expr$ is an expression in natural language (NL) and a K-string $Semrepr$ can be interpreted as a semantic representation of $Expr$, then $Semrepr$ will be called a K-representation (KR) of the expression $Expr$.

The second basic constituent of the theory of K-representations is a broadly applicable mathematical model of a linguistic database. 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 third basic constituent of the theory of K-representations is a complex, strongly structured algorithm carrying out semantic-syntactic analysis of texts from some practically interesting sublanguages of English, Russian, and German languages. The algorithm $SemSynt1$ transforms a NL-text in its semantic representation being a K-representation (Fomichov 2010). The input texts can be from the English, German, and Russian languages. That is why the algorithm $SemSynt1$ is multilingual.

An important feature of this algorithm is that it doesn’t construct any syntactic representation of the inputted NL-text but directly finds semantic relations between text units. The other distinguished feature is that a complicated algorithm is completely described with the help of formal means, that is why it is problem independent and doesn’t depend on a programming system. The algorithm is implemented in the programming languages PYTHON and C++.

The principal goals of this paper are as follows: (a) to attract the attention of the researchers to a new method of developing multilingual algorithms of semantic-syntactic analysis of texts (an implementation of this method is described in Chapters 7 – 10 of the monograph (Fomichov
2010)); (b) to illustrate the peculiarities of the central procedure of the algorithm \textit{SemSynt1}, allowing for the discovery of possible semantic relations in the combinations “Verbal form + Preposition (possibly, empty) + Noun Group”; (c) to explicitly add the parameter \textit{language} to the input data of the algorithm \textit{SemSynt1} and to add the attributes with the index \textit{language} to the attributes of several semantic-syntactic dictionaries (relations) being the parts of the considered relational linguistic database.

\section*{Morphological and Classifying Representations of an Input Text}

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

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

\textbf{Classifying representation.} From informal point of view, we will say that a classifying representation (CR) of the text \textit{T} coordinated with the morphological representation \textit{Rm} of the text \textit{T} is a two-dimensional array \textit{Rc} with the number of the rows \textit{nt} and the column with the indices \textit{unit}, \textit{tclass}, \textit{subclass}, \textit{mcoord}, in which its elements are interpreted in the following way. Let \textit{k} be the number of any row in the array \textit{Rc} i.e. \(1 \leq k \leq nt\). Then \textit{Rc}[k, unit] is one of elementary meaningful units of the text \textit{T}, i.e. if \textit{T} = \textit{t1} ... \textit{tn}, then such position \(p\), where \(1 \leq p \leq nt\), can be found that \textit{Rc}[k, unit] = \textit{tp}. If \textit{Rc}[k, unit] is a word, then \textit{Rc}[k, tclass], \textit{Rc}[k, subclass], \textit{Rc}[k, mcoord] are correspondingly a part of speech, a subclass of the part of speech, the sequences of the values of morphological properties. If \textit{Rc}[k, unit] is a construct (i.e. a value of a numeric parameter), then \textit{Rc}[k, tclass] is the string \textit{constr}, \textit{Rc}[k, subclass] is the designation of the subclass of informational units corresponding to this construct, \textit{Rc}[k, mcoord] = 0.

\textbf{Example}. Let \textit{T1} = "Has the management board of the firm “Rainbow” changed in May?". Then a classifying representation \textit{Rc1} of the text \textit{T1} coordinated with the morphological representation \textit{Rm1} of \textit{T1} may be the following array:
If T1germ = “Hat der Verwaltungsrat der Firma “Rainbow” in Mai veraendernt sich?”, then a classifying representation $Rc2$ of the text $T1germ$ coordinated with the MR $Rm2$ of $T1$ may have the following form:

<table>
<thead>
<tr>
<th>unit</th>
<th>tclass</th>
<th>subclass</th>
<th>mcoord</th>
</tr>
</thead>
<tbody>
<tr>
<td>has-changed</td>
<td>verb</td>
<td>verb-in-indic-mood</td>
<td>1</td>
</tr>
<tr>
<td>the management-board</td>
<td>noun</td>
<td>common-noun</td>
<td>2</td>
</tr>
<tr>
<td>of</td>
<td>prep</td>
<td>nil</td>
<td>3</td>
</tr>
<tr>
<td>the-firm</td>
<td>noun</td>
<td>common-noun</td>
<td>4</td>
</tr>
<tr>
<td>“Rainbow”</td>
<td>artif-name</td>
<td>nil</td>
<td>0</td>
</tr>
<tr>
<td>in</td>
<td>prep</td>
<td>nil</td>
<td>5</td>
</tr>
<tr>
<td>May</td>
<td>noun</td>
<td>proper-noun</td>
<td>6</td>
</tr>
<tr>
<td>?</td>
<td>marker</td>
<td>nil</td>
<td>0</td>
</tr>
</tbody>
</table>

The Projections of the Components of a Linguistic Basis on the Input Text

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

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

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

<table>
<thead>
<tr>
<th>ord</th>
<th>sem</th>
<th>st1</th>
<th>st2</th>
<th>st3</th>
<th>comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>change1</td>
<td>event</td>
<td>nil</td>
<td>nil</td>
<td>Yves has</td>
</tr>
</tbody>
</table>
The city has changed very much in the 1990s - 2000s

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

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

**Matrix Semantic-Syntactic Representations of NL-texts**

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

\[ locunit, nval, prep, posdir, reldir, mark, qt, nattr, \]

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

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

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

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

Let's consider the purpose of introducing the column group

\[ posdir (posdir_1, posdir_2, ..., posdir_n), \]

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

There are no governing lexical units for the verbs in the principal clauses of the sentences, that is why for the row with the ordered number \( m \) associated with a verb, \( \text{Matr}[m, \text{posdir}, d] = 0 \) for any \( d \) from 1 to \( n \). Let's agree that the nouns govern the adjectives as well as govern the designations of the numbers (e.g. "5 scientific articles"), cardinal numerals, and ordinal numerals.

The group of the columns \( \text{reldir} \) consists of semantic relations whose existence is reflected in the columns of the group \( \text{posdir} \). For filling in these columns, the templates (or frames) from the arrays \( \text{Ar}l, \text{Ar}vfr, \text{Ar}frp \) are to be used; the method can be grasped from the analysis of the algorithm \( \text{BuildMatr1} \) constructing a matrix semantic-syntactic representation of an input NL-text stated in (Fomichov 2010).

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

According to the method introduced in Chapter 8 of (Fomichov 2010), a MSSR of a NL-text \( T \) is used as an intermediary data structure for constructing a semantic representation of \( T \) being an expression of a certain SK-language (that is, being a K-representation of \( T \)). This transformation is performed by the algorithm of semantic assembly \( \text{BuildSem1} \) described in Chapter 10 of (Fomichov 2010).

**Example.** Let \( T1 \) be the question "Has the management board of the firm “Rainbow” changed in May?", and \( T1\text{germ} \) be the same question in German “Hat der Verwaltungsrat der Firma “Rainbow” in Mai veraendernt sich?”. Then it is possible to associate both with \( T1 \) and with \( T1\text{germ} \) the same K-representation \( \text{Semrepr1} \) of the form

\[
\text{Question}(x1, (x1 \equiv \text{Truth-value}(\text{Situation}(e1, \text{change}2 \ast (\text{Focus-object}, \text{certn manag-board} \ast (\text{Assoc-company}, \text{certn company1} \ast (\text{Name1, “Rainbow”) : x3}) : x2) \\
(\text{Time, Last-month}(\text{May, current-year})))))) .
\]

**Key Ideas of a Multilingual Algorithm Discovering Semantic Connections of the Verbs**

Let us consider the conditions required for the existence of a semantic relationship between a meaning of a verbal form and a meaning of a word or word combination depending in a sentence on this verbal form. Let's agree to use the term "noun group" for designating the nouns or the nouns together with the dependent words representing the concepts, objects and sets of objects. For example, let \( S1 = "\text{When and where two aluminum containers with ceramic tiles have been delivered from?}" \), \( S2 = "\text{When the article by professor P. Somov was delivered?}" \) and \( S3 = "\text{Put the} \)
blue box on the green case". Then the phrases "two aluminum containers", "the article by professor P. Somov", "blue box" are the noun groups.

Let's call "a verbal form" either a verb in personal or infinitive form or a participle or a gerund. A discovery of possible semantic relationships between a verbal form and a phrase including a noun or an interrogative pronoun is playing an important role in the process of semantic-syntactic analysis of NL-texts.

Let's suppose that posvb is the position of a verbal form in the representation Rc, posdepword is the position of a noun or an interrogative pronoun in the representation Rc. The input data of the algorithm "Find-set-relations-verb-noun" are the integers posvb, posdepword, and two-dimensional arrays Arls, Arvfr, where Arls is the projection of the lexico-semantic dictionary Lsdic on the input text, Arvfr is the projection of the dictionary of verbal-prepositional frames Vfr on the input text.

The purpose of the algorithm "Find-set-relations-verb-noun" is in the first place to find the integer number nrelvbdep - the quantity of possible semantic relationships between the values of the text units with the numbers p1 and p2 in the classifying representation Rc. Secondly, this algorithm should build an auxiliary two-dimensional array Arrelvbdep keeping the information about possible semantic connections between the units of Rc with the numbers p1 and p2. The rows of this array represent the information about the combinations of a meaning of the verbal form and a meaning of the dependent group of words (or one word).

The structure of each row of the two-dimensional array Arrelvbdep with the indices of columns linenoun, linevb, role, example is as follows. For the filled in row with the number k of the array Arrelvbdep (k ≥ 1), linenoun is the ordered number of the row of the array Arls corresponding to the word in the position p1; linevb is the ordered number of the row of the array Arls corresponding to the verbal form in the position p2; role is the designation of the semantic relationship (thematic role) connecting the verbal form in the position p2 with the dependent word in the position p1; example is an example of an expression in NL realizing the same thematic role.

The search of the possible semantic relationships between a meaning of the verbal form (VF) and a meaning of the dependent group of words (DGW) is done with the help of the projection of the dictionary of verbal-prepositional frames (d.v.p.f.) Arvfr on the input text. In this dictionary such a frame (or a template) is searched that it would be compatible with the certain semantic-syntactic characteristics of the VF in the position posvb and the DGW with the number posdepword in Rc. Such characteristics include, first of all, the set of codes of grammatic cases Grcases associated with the text-forming unit having the ordered number - value posdepwd ("the position of dependent word") in Rc. Let's suppose that Rc[posvb, tclass]=verb. Then Grcases is the set of grammatic cases corresponding to the noun in the position posdepword.

Description of an Algorithm Discovering Semantic Relations Between a Verb and a Noun Group
**Purpose of the Algorithm "Find-set-relations-verb-noun"**

The algorithm is to establish a thematic role connecting a verbal form in the position \( posvb \) with a word (noun or connective word) in the position \( posdepword \) taking into account a possible preposition before this word. As a consequence, to select one of the several possible values of a verbal form and one of the several possible values of a word in the position \( posdepword \). In order to do this, three enclosed loops are required: (1) with the parameter corresponding to a possible meaning of the word in the position \( posdepword \); (2) with the parameter corresponding to a possible meaning of the verbal form; (3) with the parameter corresponding to a verbal-prepositional frame associated with this verbal form.

**External specification of the algorithm "Find-set-relations-verb-noun"**

**Input:**
- `input-lang` – string with the values `eng`, `germ`, `rus` denoting the English, German, and Russian languages;
- `Rc` - classifying representation,
- `nt` - integer - quantity of the text units in the classifying representation \( Rc \), i.e. the quantity of rows in \( Rc \),
- `Rm` - morphological representation of the lexical units from \( Rc \),
- `posvb` - integer - position of a verbal form (a verb in a personal or infinitive form, or a participle or a gerund),
- `posdepword` - integer - position of a noun,
- `Matr` - initial value of MSSR of the text;
- `Arls` - array - projection of the lexico-semantic dictionary \( Lsdic \) on the input text \( T \);
- `Arvfr` - array - projection of the dictionary of verbal-prepositional frames \( Vfr \) on the input text \( T \).

**Output:**
- `arrelvbdep` - one-dimensional array designed to represent the information about (a) a meaning of a dependent word, (b) a meaning of a verbal form, and (c) about a semantic relationship between the verbal form in the position \( posvb \) and the dependent word in the position \( posdepword \);
- `nrelvbdep` - integer – the quantity of meaningful rows in the array `arrelvbdep`.

**External specification of the auxiliary algorithm "Characteristics-of-verbal-form"**

**Input:**
- `p1` - the number of a row from the classifying representation \( Rc \) corresponding to a verb or a participle.

**Output:**
- `form1`, `refl1`, `voice1` - strings; their values are defined in the following way. If \( p1 \) is the position of a verb, then `form1` may have one of the following values: `indic` (the sign of the indicative mood), `infinit` (the sign of the infinitive form of a verb), `imperat` (the sign of the imperative mood). If \( p1 \) is the position of a participle, then `form1 := indic`. The string `refl1` takes the values `rf` (reflexive verb) or `nrf` (non-reflexive verb). The string `voice1` takes the value `actv` (the sign of the active voice) or `passv` (the sign of the passive voice). The values of the parameters `form1`, `refl1`, `voice1` are calculated based on the set of the numeric codes of the values of the morphological characteristics of the text unit with the ordered number \( p1 \).

**External specification of the auxiliary algorithm "Range-of-sort"**

**Input:**
- `z` - sort, i.e. an element of the set \( St (B (Cb (Lingb))) \), where \( Lingb \) is a linguistic basis (see Chapter 7 of (Fomichov 2010)), \( Cb (Lingb) \) is a marked-up conceptual basis, \( B (Cb (Lingb)) \)
is the conceptual basis being the first component of $Cb (Lingb)$, $St (B (Cb (Lingb)))$ is the set of sorts determined by the conceptual basis $B (Cb (Lingb))$.

**Output:** spectrum - set of all sorts being the generalizations of the sort $z$, including the sort $z$ itself.

**Algorithm "Find-set-relations-verb-noun"**

Begin  Characteristics-of-verbal-form (posvb, form1, refl1, voice1)
    nrelvbdep := 0
Comment
Now the preposition is being defined
End-of-comment
    prep := leftprep
Comment
Calculation of $posn1$ - position of the noun that defines the set of sorts of the text unit in the position $posdepword$
End-of-comment
    posn1 := posdepword
Comment
Then the set of grammatic cases $Grcases$ is being formed. This set will be connected with the word in the position $posdepword$ in order to find a set of semantic relationships between the words in the positions $posvb$ and $posdepword$.
End-of-comment
    t1 := Rc [posvb, tclass]
    t2 := Rc [posvb, subclass]
    p1 := Rc [posdepword, mcoord]
    Grcases := Cases (Rm [p1, morph])
    line1 := Matr [posn1, locunit]
    numb1 := Matr [posn1, nval]
Comment
The quantity of the rows with the noun meanings in $Arls$
End-of-comment
    loop for $i1$ from $line1$ to $line1 + numb1 - 1$
Comment
A loop with the parameter being the ordered number of the row of the array $Arls$ corresponding to the noun in the position $posn1$
End-of-comment
    Set1 := empty set
    loop for $j$ from 1 to $m$
Comment
$m$ - semantic dimension of the sort system $S(B(Cb (Lingb)))$, i.e. the maximal quantity of incomparable sorts that may characterize one entity
End-of-comment
    current-sort := Arls [$i1$, $stj$]
    if current-sort ≠ nil
then Range-of-sort (current-sort, spectrum)
    Set1 := the union of the set Set1 and the set spectrum
end-if
Comment
For an arbitrary sort z the value spectrum is the set of all sorts being the generalizations of the
sort z including the sort z itself
End-of-comment
    end-of-loop
Comment
End of the loop with the parameter j
End-of-comment
Comment
Then the loop with the parameter corresponding to a meaning of the verbal form follows
End-of-comment
    line2 := Matr [posvb, locunit]
    numb2 := Matr [posvb, nval]
Comment
The quantity of the rows with the meanings of the verbal form in Arls
End-of-comment
    loop for i2 from line2 to line2 + numb2 - 1
Comment
A loop with the parameter being the ordered number of a row of the array Arls corresponding to
the verb in the position posvb
End-of-comment
    current-pred := Arls [i2, sem]
    loop for k1 from 1 to narvfr
        if Arvfr [k1, semsit] = current-pred
            then begin
                s1 := Arvfr [k1, str]
                if ((input-lang = Arvfr [k1, lang] and (prep = Arvfr [k1, sprep])
                and (s1 belongs to Set1) and (form1 = Arvfr [k1, fm])
                and (refl1 = Arvfr [k1, refl]) and (voice1 = Arvfr [k1, vc]))
                    then grc := arvfr [k1, grcase]
                    if (grc belongs to Grcases)
                        then
    Comment
The relationship exists
End-of-comment
    nrelvbdep := nrelvbdep + 1
    arrelvbdep [nrelvbdep, linevb] := i2
    arrelvbdep [nrelvbdep, linenoun] := i1
    arrelvbdep [nrelvbdep, gr] := grc
    arrelvbdep [nrelvbdep + 1, role] := arvfr [k1, trole]
end-if
end-if
Commentary on the Algorithm "Find-set-relations-verb-noun"

The quantity \( nrelvbdep \) of the semantic relationships between the verbal form and a noun depending on it in the considered sentence is found. Let's consider such sublanguages of English, German, and Russian languages that in all input texts a verb is always followed (at certain distance) by at least one noun.

The information about such combinations of the meanings of the verb \( V \) and the noun \( N1 \) that give at least one semantic relationship between \( V \) and \( N1 \) is represented in the auxiliary array \( arrelvbdep \) with the indices of the columns \( linenoun \), \( linevb \), \( role \), \( example \). For arbitrary row of the array \( arrelvbdep \), the column \( linenoun \) contains \( c1 \) - the number of such row of the array \( Arls \) that \( Arls [c1, ord] = posn1 \) (position of the noun \( N1 \)). For example, for \( Q1 = "When and where 3 aluminum containers have been delivered from?" \) \( Arls [c1, sem] = container1 \).

The column \( linevb \) contains \( c2 \) - the number of a row of the array \( Arls \) for which \( Arls [c2, ord] = posvb \), i.e. the row \( c2 \) indicates a certain meaning of the verb \( V \) in the position \( posvb \). For example, for \( Q1 = "When and where 3 aluminum containers have been delivered from?" \) the column \( Arls [c2, sem] = delivery2 \).

The column \( role \) is designed to represent the possible semantic relationships between the verb \( V \) and the noun \( N1 \). If \( nrelvbdep = 0 \) then the semantic relationships have not been found. Let's assume that this is not possible for the considered input language. If \( nrelvbdep = 1 \) then the following meanings have been clearly defined: the meaning of the noun \( N1 \) (by the row \( c1 \) ), the meaning of the verb \( V \) (by the row \( c2 \) ), and the semantic relationship \( arrelvbdep [nrelvbdep, role] \). For example, for the question \( Q1 \) the following relationships are true: \( V = "delivered" \), \( N1 = "containers" \), \( nrelvbdep = 1 \), \( arrelvbdep [nrelvbdep, role] = Object1 \). If \( nrelvbdep > 1 \) then it is required to apply the procedure that addresses clarifying questions to the user and to form these questions based on the examples from the column \( example \).

Example. Let T2 be the sentence in German "Dr. Kurt Stein hat in Mai den Verwaltungsrat der Firma "Rainbow" eingetreten" (an English version of T2 is the sentence T2eng = “Dr. Kurt Stein joined in Mai the management board of the firm “Rainbow”). The German verb “eintreten” has, in particular, the following meanings: (1) to stand up for, (2) to make comfortable shoes, (3) to join an organization. The conceptual analysis of T2 will enable a hypothetical applied intelligent system to positively answer the question \( T1 = "Has the management board of the firm “Rainbow” changed in May?" \).
Let’s show some details of analyzing T2. Suppose that the projection of the dictionary of verbal-prepositional semantic-syntactic frames $Vfr$ on the input text T1 may include the following subarray $Arvfr2fragm$ of the array $Arvfr2$:

<table>
<thead>
<tr>
<th>nb</th>
<th>semsit</th>
<th>lang</th>
<th>fm</th>
<th>refl</th>
<th>vc</th>
<th>trole</th>
<th>sprep</th>
<th>grc</th>
<th>str</th>
<th>expl</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>standing-up-for</td>
<td>germ</td>
<td>indic</td>
<td>nrf</td>
<td>actv</td>
<td>Supported-person</td>
<td>fuer</td>
<td>4</td>
<td>ints</td>
<td>expl1</td>
</tr>
<tr>
<td>4</td>
<td>making-comfortable</td>
<td>germ</td>
<td>indic</td>
<td>nrf</td>
<td>actv</td>
<td>Object-to-wear</td>
<td>nil</td>
<td>4</td>
<td>shoes</td>
<td>expl2</td>
</tr>
<tr>
<td>4</td>
<td>jsoining2</td>
<td>germ</td>
<td>indic</td>
<td>nrf</td>
<td>actv</td>
<td>New-org</td>
<td>nil</td>
<td>4</td>
<td>org</td>
<td>expl3</td>
</tr>
</tbody>
</table>

The number 4 in the column nb indicates the 4th position of the text unit “hat eingetreten” in the classifying representation of T2. The elements germ, indic, nrf, actv are interpreted as the values German, indicative-mood, non-reflexive, active-voice of the properties language, form-of-verb, reflexivity, voice; the elements Supported-person, Object-to-wear, New-org are the designations of thematic roles (or conceptual cases). The number 4 in the column grc designates the article Akkusativ in the German language. The elements ints, shoes, org are interpreted as the sorts intelligent-system, shoes, organization. The examples expl1, expl2, expl3 are defined by the relationships

expl1 = “Paul hat fuer seinen Freund Jens eingetreten” (“Paul has stood up for his friend Jens”);
expl2 = “Jean hat seine Schuhe eingetreten” (“Jean has made comfortable his shoes”);
expl3 = “Helene hat die Firma IBM in Maerz eingetreten” (“Helene joined in March the company IBM”).

Suppose that the algorithm “Find-set-relations-verb-noun” looks for possible semantic relations (thematic roles) in the combination

(hat eingetreten, den Verwaltungsrat)                            (1)

from the text T2. Then input-lang = germ, posvb = 4 (the position of the combination “hat eingetreten” in the classifying representation (CR) of T2), posn1 = 7 (the position of the combination “den Verwaltungsrat” in CR of T2).

The values of the parameter $i1$ of the first loop of the algorithm correspond to different meanings of the noun group in the position 7. Different values of the parameter $i2$ of another loop of the algorithm correspond to three meanings of the verb “eintreten” reflected in the considered subarray $Arvfr2fragm$ of the array $Arvfr2$.

The frame represented by the first row of the subarray $Arvfr2fragm$ doesn’t matches the combination (1) due to the lack in T2 of the German preposition “fuer” (“for” in English). The second frame determined by the subarray $Arvfr2fragm$ doesn’t matches the combination (1) too, since any reasonably designed lexico-semantic dictionary $Lsdic$ doesn’t allow for associating the sort “shoes” with the semantic unit manag-board (corresponding to the word combinations “a management board” and “ein Verwaltungsrat”. But the third frame from the subarray $Arvfr2fragm$ matches the combination (1), hence the semantic relation New-org will be discovered.
Step by step, the modified algorithm $\textit{SemSynt1b}$ (it includes the modified procedure “Find-set-relations-verb-noun”) will build the following K-representation $\textit{Semrepr2}$ of T2:

$\text{Situation (e2, joining2 }\ast\text{ (Agent1, certn person }\ast\text{ (First-name, “Kurt”) (Surname, “Stein”) : x1)) (Time, Last-month(May, current-year))}$

$\text{(New-org, certn org }\ast\text{ (Isa, mang-board)(Company-part, certn company1 }\ast\text{ (Name1, “Rainbow”) : x3) : x2)).}$

Obviously, it is not difficult to include into the knowledge base of a hypothetical information retrieval system such fragments that this system would give a positive reply to the initial question $T1 = \text{“Has the management board of the firm “Rainbow” changed in May?“}$, proceeding from the information conveyed by the K-representation $\textit{Semrepr2}$.

Conclusions

The new method of developing the algorithms of semantic-syntactic analysis of NL-texts introduced in (Fomichov 2010) was modified and illustrated above. The method has a number of significant advantages in comparison with other known methods of developing the algorithms of the kind. Firstly, the explicitness and fullness of the description of the algorithm $\textit{SemSynt1}$ in (Fomichov 2010) is many times higher than it is typical for the scientific publications on this problem (see, e.g., the paper (Popescu et al. 2003)). Secondly, the method doesn’t foresee the construction of a pure syntactic representation of the analyzed NL-text: it is oriented at discovering the semantic relations between the elementary meaningful units of a text.

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

It seems that the algorithm $\textit{SemSynt1}$ in its modified form described above can be used as a basis for designing multilingual conceptual information retrieval systems of the computer intelligent systems with adaptive planning capabilities.

References

2. Popescu, A.-M., Etzioni, O., Kautz, H. (2003); Towards a Theory of Natural Language Interfaces to Databases. In: Proceedings of the 8th International Conference on Intelligent User Interfaces; Miami, FL (pp. 149-157)