Formalization of scientific and technical sentences in Russian language

Y Vishniakov and R Vishniakov
Kuban State University, Stavropolskaya ulitsa, 149, Krasnodar, Krasnodar Krai, Russland, 350040
E-mail: jury.vishnyakov@gmail.com

Abstract. The article presents the research results on the computational theory of semantic interpretation in Russian language. These interpretations represent a generalizing formalization of a scientific and technical style sentences. Furthermore, the concept of a partition system of a word scale, its classes and a method of forming partition classes is defined. We showed that the word scale corresponds to many partition systems that are specified by a special rule for selecting instances from classes. Additionally, we proposed generalized algorithms for constructing a partition system.

1. Introduction
Great difficulty in processing of natural language texts, in areas such as information retrieval, classification and clustering of documents, is a lack of the universal mathematical semantics formalization. This leads to a lack of solutions that are acceptable in terms of efficiency and accuracy. In particular, difficulties are intensified by the increase of stored and constantly updated information volume, as well as by increasing demands of users to the quality and accuracy of the information processing. Thus, according to some estimates, the amount of information in the Internet increases exponentially each year. The implicitly accepted frequency paradigm for modeling semantic relevance, based on the properties of words frequency and the solutions proposed within it do not significantly increase the efficiency and accuracy of processing text information. The numerous scientific publications and forums devoted to the semantic relevance problem confirm that the problem of processing natural languages is still far from its final solution [1-6].

Apparently, the success of the existing solutions depends on the vast variety of the associations in the natural language. Those associations are a form of information presentation and its interpretation by the subject (semantic content introduced by the subject). However, linguistic scientific vocabulary and meta-annotations are based on the natural language and require mathematical formalization [7].

This paper describes a part of the research carried out in the framework of the computational approach developed by the authors in [8-10]. The work aimed at developing formalizations of natural language texts with an emphasis on the scientific and technical style.

2. Methods
The approach to the interpretation of semantics in relation to the subject of study of this work can be explained as follows. Let some string q of words ai of the form be given:

\[ q = a_1, a_2, ..., a_n \]  

(1)
representing a meaningful text fragment (sentence or part thereof). Integrity assumes that all the words of the q are included in phrases, forming an unambiguous relationship of direct subordination on the set of words.

We denote by $S(ai)$ the set of semantic meaning of the word ai, then for a text fragment q the set of semantic meaning $S(q)$ can be represented by a functional $\Phi$ of the general form:

$$S(q) = \Phi(S(a_1), S(a_1), \ldots, S(a_n))$$  \hfill (2)

Let's discuss how the functionality is formed. If in a phrase the word a is the main word, and the word b is the dependent one, then the dependent word always clarifies the meaning of the main word and plays the role of a meaningful context. We represent this dependence by the form:

$$\overrightarrow{a: b}$$  \hfill (3)

where the arrow indicates the direction of the word dependency. If the word a is included in the main word in several phrases with dependent words $b_1, b_2, \ldots, b_p$, then this circumstance is represented by a the form:

$$\overrightarrow{a: \{b_1, b_2, \ldots, b_p\}}$$  \hfill (4)

and call it the contextual connective of the word a.

From the above it follows that the semantic meaning of the phrase is a subset of the semantic meanings of its main word. The semantic meaning is determined by the meaning of the dependent word that forms the context in the phrase. We write this in the following expression.

$$S(\overrightarrow{a: b}) = S(a) \mid_{S(b)}$$  \hfill (5)

For this expression, the following relationships are true:

$$S(\overrightarrow{a: b}) = S(a) ; S(\overrightarrow{a: b}) \neq S(\overrightarrow{b: a}) ; S(\overrightarrow{a: a}) = S(a)$$  \hfill (6)

Now we introduce and define the operation of contextual clarification of the meaning of the main word a by the dependent word b as follows:

$$S(\overrightarrow{a: b}) = S(a) \cap S(b).$$  \hfill (7)

Here $\cap$ is the operation of contextual clarification of meaning, where the arrow defines the direction of words dependence in the phrase. Extending (8) to the contextual connective (3), we obtain:

$$S(\overrightarrow{a: (b_1, \ldots, b_p)}) = S(\overrightarrow{a: b_1}) \cap S(\overrightarrow{a: b_2}) \cap \ldots \cap S(\overrightarrow{a: b_p})$$  \hfill (8)

or otherwise:

$$S(\overrightarrow{a: (b_1, \ldots, b_p)}) = \cap_1^p S(\overrightarrow{a: b_i}).$$  \hfill (9)

Here $\cap$ – is the operation of intersecting sets.

Now, knowing the dependencies of words and using the operations of contextual refinement of the meaning and intersection of sets, we can construct a formulaic representation of the functional meaning for the whole string q. It should be noticed that if the word a is the main word in the q string, then the relation is always true:

$$S(q) \subset S(a)$$  \hfill (10)

Thus, the meaning of the text fragment q is a subset of the set of meanings of its main word a.

The authors' works, in particular, in [3], analyze the functional of meaning, methods for constructing it, computing, forms and algorithms for representing computational procedures, comparing text fragments for semantic proximity by type of functional. Though, it should be noted that the analytical
construction of the functional of meaning a priori assumes that the relation of subordination of words in a text fragment is already known.

At the same time, identifying relationships by a type of a test fragment and determining the structure of a textual fragment is not a trivial task. The authors are not aware of formalizations of this kind of relationships.

Let us refer to verbal formalizations used in linguistics, for example, in [2]. Regarding the structure of the proposal there is the following (our detente) “In natural language, the structure of a sentence can be considered from the point of view of the connection of words. However, it does not come down only to word relationships in phrases. Words can also form larger groups connected by a single meaning. Thus, the linear structure of a sentence, based on some intuitive criteria, can be divided into integral fragments, called segments, which can consist of more than one word. When forming the segments, the condition must be met: if two segments have a common part, then one of them must be completely embedded in the other. The sentence itself should be considered as the largest segment. The resulting segments are termed as components, and their combination – a system of components.”

As a matter of fact, such a verbal formalization is understandable for a specialist in linguistics, but it is clearly not enough for computer processing and it needs mathematical representation, estimation, rules and construction algorithms, which is the purpose and content of this work.

We introduce the concept of a word scale and consider its main properties [2]. Let some sentence α be given, represented by a nonempty string of words \( x_1, x_2, ..., x_m \). Let’s represent a sentence on a straight line and each word, in the order of its sequence in the sentence, is mapped to the numbered and indivisible segments of unit length. The first segment is matched with the beginning of the line, and the last with its end. We will name such a line as a word scale, and by its length we mean the number of elementary segments forming it.

A partition \( R \) of a word scale is its own representation as a set of disjoint adjacent segments, in the general case having different lengths. Obviously, the set \( \mathcal{R} \) of all partitions of the word scale is finite and its power is determined by the length of the word scale. We include in this set the partition \( R_1 \), corresponding to the largest segment - the entire word scale (sentence). Also, the partition \( R_m \) is included, which represents only elementary segments (sentence words) of the word scale. The partitions \( R_1 \) and \( R_m \) are unique and represent extreme cases between which there are many other partitions included in the set \( \mathcal{R} \).

Let’s construct partition classes on the set \( \mathcal{R} \), where the class \( \mathcal{R}_i = \{ R_{i1}, R_{i2}, ..., R_{ij}, ..., R_{ip} \} \) includes all partitions consisting of \( i \) segments. Moreover, it’s plain to see that the number of such classes is \( m \). Note that the classes \( \mathcal{R}_1 = \{ R_1 \} \) and \( \mathcal{R}_m = \{ R_m \} \) are represented by only one instance - the partitions \( R_1 \) and \( R_m \), respectively.

A combinatorial estimate \( N(m) \) of the set \( \mathcal{R} \) of all partitions of the word scale is carried out by constructing the numbering of its elements. In the numbering \( N(m) \), the classes are arranged in the order of numbers \( i = 1, 2, 3, ..., m \). Thus, the required numbering has the form:

\[
N(m) = \binom{1}{m}, \binom{2}{m}, ..., \binom{i}{m}, ..., \binom{m}{m},
\]

and the numbering term \( \binom{i}{m} \) is a subnumbering of partitions of the class \( \mathcal{R}_i \).

From the previous explanations it follows that if \( i = 1 \), then the class \( \mathcal{R}_1 \) includes only one partition representing the entire word scale. We assign the first number in the numbering and the name \( R_1 \) to this partition. If \( i = m \), then the class \( \mathcal{R}_m \) is also represented by just one partition, to which we assign the last number and name \( R_m \). The partition \( R_m \) consists of \( m \) elementary segments of the word scale and is the last member of the numbering. In [2], the mechanism of the formation of subnumberations is considered and a combinatorial estimate of the cardinality of the set \( \mathcal{R} \) is derived in the form of the relation \( N = 2^{m-1} \).

Now let’s move on to other concepts. Let some partitions \( R_i \) and \( R_j \) \( (i > j) \) be given in the set \( \mathcal{R} \) of all partitions of a word scale with a length \( m \), such that all segments of the partition \( R_i \) either do not coincide with the segments of the partition \( R_j \) or are embedded in them. For partitions \( R_i \) and \( R_j \), we
introduce the embedding operation $. The result of this operation is represented as a certain combined form of partitions (combined partition) $R_c(i, j)$, which is represented by all segments of the partition $R_j$ and the segments of the partition $R_i$ embedded in them. We define the mathematically combined form of partitions $R_i$ and $R_j$ as the result of the embedding operation and present it in the following form:

$$R_c(i, j) = R_i \vdash R_j.$$  \hfill (12)

Figure 1 shows a schematic illustration of the operation of embedding partitions $R_i$ and $R_j$. Its result - the combined form of the partition:

![Figure 1. Operation of embedding partitions.](image)

We expand the concept of a combined partition into the whole set of partitions $\mathcal{R}$ of the word scale. So, the word scale of the combined partition from partitions of all numbering classes $N(m)$ of the form will be named as the system of the partition $\Psi(m)$ of:

$$\Psi(m) = R_1 \vdash R_{2,k_2} \vdash \cdots \vdash R_{i,k_i} \vdash \cdots \vdash R_m,$$

where: $R_{i,k_i} \in \mathcal{R}_i.$ \hfill (13)

From the definition of the operation of embedding and numbering subclasses, it follows that only one instance is included in the partition system $\Psi(m)$ from each partition class (subnumbering). The way to select an instance from the class is determined by a rule, which we call an $S$ - rule.

The partition system $\Psi(m)$ can be constructed in two ways. The first method is characterized by the fact that the construction begins with the partition $R_1$ and then proceeds to the sequential construction of $R_2, R_3, \ldots$. We will call this method a top-down strategy. An enlarged top-down strategy can be represented by the following algorithm:

**Top-down strategy:**
1. For a specific sentence $\alpha$, construct a word scale and a partition $R_1$ represented by one segment.
2. Sequentially for $i = 2, \ldots, m$, applying S-rules, select an instance of the partition from the class $\mathcal{R}_i$ and include it in the partition system $\Psi_S(m)$.

The second method starts with the partition $R_m$ and moves on to the construction of the partitions $R_{m-1}, R_{m-2}, \ldots, R_1$. This method is a so-called a bottom-up strategy. The strategy can also be aggregated by the following algorithm:

**Bottom-up strategy:**
1. For a specific sentence $\alpha$, construct a word scale and a partition $R_m$ represented by elementary segments.
2. Sequentially for $i = m - 1, m - 2, \ldots, 1$, by applying the S-rule, select an instance of the partition from the class $\mathcal{R}_i$ and include it in the partition system $\Psi_S(m)$.

Here, the index $S$ in $\Psi_S(m)$ means that the strategy uses the corresponding S-rule.

Let’s illustrate the strategies for constructing the partition system $\Psi(m)$ with examples using the simplest S-rules. The first S-rule implies the split-off of the left unit segment from the composite segment. This rule is used in the strategy from top to bottom. The second S-rule - involves the attachment of the first left unit segment to the composite segment on the right and is used in the bottom-up strategy.
Step-by-step execution of the algorithm for constructing a system of partitions from top to bottom is presented in Figure 2.

You may notice that the partition $R_1$ by definition is included in the partition system at the first step (s. 1). At the second step (s. 2), for $i = 2$, from the class $\mathcal{R}_2$ it is selected as follows: 1) in the previous partition $R_1$, the leftmost unit segment is split off, 2) the unit segment and the remainder from $R_1$ form a new partition $R_2 \in \mathcal{R}_2$. The process is repeated until the composite segment turns into a single segment. After that, the partition system $\Psi_S(m)$ is represented by all the resulting partitions $R_1, R_2, ..., R_m$. Note that all segments of the partition system are embedded in each other.

In a bottom-up strategy, a composite segment is formed on the left side of the word scale by attaching to it the first left unit segment, which prescribes the S-rule of this strategy. The step-by-step algorithm of the bottom-up strategy is presented in Figure 3.
Figure 3. Building a partition system. Bottom-up strategy.

Note that the partition systems of Figures 2 and 3 are different, since different S-rules are used to construct them. It should also be noticed that the S-rules used to construct partition systems can be different. This allows, based on specific goals, to create unique configurations of the partition system $Ψ(m)$ that possess predetermined properties.

The construction of S-rules is not discussed in this paper and is the subject of a separate discussion, but, nevertheless, the following is an example that is directly related to the formalization of natural language sentence structures.

3. Results

So, we will consider the sentence $α$ presented in Figure 4. This sentence is projective and has word dependences marked by arcs.

Figure 4. An example of a projective sentence with word dependencies.

Let for the word scale of this sentence be constructed: the set of partitions the classes of partitions $\{R_1, R_2, ..., R_i, ..., R_m\}$. We construct the S-rule in order to include instances of partitions from the class $R_i$ in the partition system $Ψ(α)$ assuming that the construction of the partition system $Ψ(α)$ is based on the top-down strategy considered above. The selected instance of the partition from the class $R_i$ must have the following properties:

1. Must contain a word - the main word of the segment, in which the arc of dependence directed into it comes from a word lying outside of this segment.
2. For all other words of the segment, except the main one, all arcs of dependencies included in them come from words that are within the given segment.
3. The main word of the segment corresponding to the partition \( R_1 \) represents the predicate (axiom) \([2]\).

Now we will analyze the formation of segments. Let some partition \( R_{i,q} \in \mathcal{R}_i \) be given, including some segment \( \beta \). From the definition of the concept of a partition system, some, and in a particular case - all, segments of the partition \( R_{i+1,t} \in \mathcal{R}_{i+1} \) should be embedded in the segment \( \beta \). We construct these segments of the partition \( R_{i+1,t} \in \mathcal{R}_{i+1} \) as follows. We hang the arcs of dependencies on the words of the segment \( \beta \), define in it the main word \( x_Y \) and the dependent word \( x_i \) associated with it.

The operation of splitting the segment \( \beta \) by the main word \( x_k \) is its division into two adjacent subsegments according to the rule:
1. In the segment \( \beta \), find the dependency arc starting from the main word \( x_k \);
2. Draw a perpendicular line of splitting that cuts the segment next to the main word \( x_k \) so that it intersects the found arc of dependence according to s. 1. There are no other words between the main word and the line of splitting.
3. After the splitting operation is completed, the segment is divided along the splitting line into two adjacent sub-segments \( \beta_1 \) and \( \beta_2 \). One of the sub-segments contains the main word of the segment \( \beta \).
4. The segments \( \beta_1 \) and \( \beta_2 \) resulting from the splitting of the segment \( \beta \) are included in the partition \( R_{i+1,t} \in \mathcal{R}_{i+1} \).

The principle of such a splitting operation is illustrated in Figure 5.

**Figure 5.** Schematic representation of the operation of splitting a segment. Here the «линия расщепления» – splitting line and «подчинительная связь» - subordinate connection.

Let's consider the algorithm for constructing the partition system \( \Psi(\alpha) \) of the sentence \( \alpha \) based on the splitting operation:
1. Put the variable \( i = 1 \) (index counter).
2. Select the partition \( R_i \) and find in it a segment that has an arc of dependence starting from the main word and ends with a word in this segment.
3. Apply to this segment the splitting operation according to the found dependence arc. Then using the results of the operation, construct the partition \( R_{i+1} \) and include it in the partition system \( \Psi(\alpha) \). Increase the counter \( i = i + 1 \).
4. Repeat the step 2 until the partition \( R_{i+1} \) contains segments of dimension more than one word.
5. The process is complete, the partition system \( \Psi(\alpha) \) is constructed.

The partition system of the word scale of the sentence \( \alpha \), constructed on the basis of the above described algorithm, is called the normal partition system and denoted by \( \bar{\Psi}(\alpha) \).

Below is given an example of step-by-step execution of the algorithm for sentence \( \alpha \), represented in Figure 4. Its normal partition system \( \bar{\Psi}(\alpha) \) is presented in Figure 6.
Figure 6. Construction of a normal partition system $\Psi(\alpha)$.

The main words of the composite segments of the partitions are highlighted in bold arcs. The unit segments of the word scale (words) are renumbered by numbers of the natural series. Then the indexed names $\beta_i$ are assigned to the segments of the normal partition system in the order they are identified by the algorithm.

The normal partition system $\Psi(\alpha)$ can be represented as a graph, applying the following algorithm:

1. To each segment map a tree node.
2. To map the root of the tree to the segment corresponding to the whole sentence. Then to map the leaves to the elementary segments corresponding to the words.
3. In a tree of two different nodes, place the node of the segment belonging to the partition $R_i$ below the node of the segment of the partition $R_j$ if $i > j$.
4. If for some $i = j + 1$ a segment of the partition $R_i$ includes many segments of the partition $R_i$, then from the nodes representing them to form a bush of the node of the partition segment $R_j$.

The tree $\Psi(\alpha)$ is a result of applying the algorithm to the above constructed normal partition system of a sentence $\alpha$, see Figure 7.
Figure 7. The tree of the normal partition system $\Psi(\alpha)$.

Table 1. Grammatical roles and their abbreviations.

| №  | Grammatical role         | Abbreviation | №  | Grammatical role                        | Abbreviation          |
|----|--------------------------|--------------|----|-----------------------------------------|-----------------------|
| 1  | Sentence                 | Прж          | 9  | Group of secondary part                 | Гр_Всч                 |
| 2  | Group of the subject     | Гр_Пдж       | 10 | Group of objective compliment           | Гр_Доп                 |
| 3  | Subject                  | Пдж          | 11 | Objective compliment                    | Доп                   |
| 4  | Predicate group          | Гр_Скз       | 12 | Group of modifiers                      | Гр_Опр                 |
| 5  | Predicate                | Скз          | 13 | Modifier                                | Опр                   |
| 6  | Simple verb predicate    | ПрГ_Скз      | 14 | Group of adverbial                      | Гр_Обс                 |
| 7  | Compound Verb            | СГ_Скз       | 15 | Adverbial                               | Обс                   |
| 8  | Compound nominal predicate| СИ_Скз       | 16 |                                        |                       |

Here the rectangles are not the end leaves of the tree. They only explain the specific word meanings of the end nodes.

An example of representing the normal partition system $\Psi(\alpha)$ of the sentence $\alpha$ in the notation of the system of components is shown in Figure 8.
Figure 8. The system of components of the normal partition system $\Psi(\alpha)$. The sentence «Профессиональная аккредитация является основой международного признания образовательных программ российских вузов» in English translation is «Professional accreditation is the basis of international recognition of educational programs of Russian universities».

We modify the tree of the normal partition system $\Psi(\alpha)$ by assigning grammatical roles to strings of words included in the segments. In this case, the grammatical role of the segment is determined by the role of the main word of the segment in the sentence. Note that such attribution of the role to the segments is possible [2], since each of them has semantic completeness relative to the main word. Rename each grammatical role by the abbreviated name (abbreviation), as shown in Table 1.

For the considered example of the sentence $\alpha$, we assign the roles to the segments of the normal partition system $\Psi(\alpha)$ as shown in Table 2.

| Substring                                                                 | Syntactic role |
|---------------------------------------------------------------------------|----------------|
| $\beta_1 = \alpha$                                                       | Прж           |
| $\beta_2 =$ профессиональная аккредитация                                   | Гр_Пдж         |
| $\beta_3 =$ является основой международного признания образовательных программ российских вузов | Гр_Скз         |
| $\beta_4 =$ профессиональная                                              | Опр           |
| $\beta_5 =$ аккредитация                                                 | Пдж           |
| $\beta_6 =$ является                                                     | Скз           |
| $\beta_7 =$ основой международного признания образовательных программ российских вузов | Гр_Доп         |
| $\beta_8 =$ основой                                                     | Доп           |
| $\beta_9 =$ международного признания образовательных программ российских вузов | Гр_Опр         |
| $\beta_{10} =$ международного признания                                  | Гр_Опр         |
Now we modify the tree of the normal partition system $\Psi(\alpha)$ by assigning roles to its nodes and call it labeled. In addition to each rectangle representing a separate sentence word, we assign a label according to the type of part of speech. After this, the marked tree will take the form shown in Figure 9.

**Figure 9.** The labeled tree of the normal partition system $\Psi(\alpha)$

Here the nodes correspond to the syntactic roles of the segments, and the rectangles correspond to the sentences. Rectangles are marked with the corresponding names of the parts of speech.

### 4. Conclusion

So, the part of the research on the computational theory of semantic interpretation presented in the work represents a generalizing formalization of natural language sentences. The sentences are a part of the
class of a components system. For this purpose, the concept of the partition system $\Psi(m)$, formed by the combined partitions of all classes of the word scale, is formally defined. A method for selecting instances from all classes of partitions to include them in the partition system $\Psi(m)$, defined by the concept of an S-rule, is defined. It is shown that the word scale can correspond to many partition systems, which are determined by S-rules for including instances from the corresponding partition classes in the system. Generalized algorithms for constructing a system of partitions according to top-down and bottom-up strategies are presented. Also, a generalizing formalization of formalizing the sentence structure widely used in linguistics as a system of components is introduced.

References
[1] Shelmanov A O Smirnov I V 2014 Methods for semantic role labeling of Russian texts In Computational Linguistics and Intellectual Technologies. Proceedings of International Conference Dialog Vol 13 No 20 pp 607-620
[2] Cruse A 2011 Meaning in language: An introduction to semantics and pragmatics Oxford University Press UK.
[3] Dobrov B V and Loukachevitch N V 2011 Multiple Evidence for Term Extraction in Broad Domains RANLP pp 710-715.
[4] Tseng Y-H Lin C-J and Lin Y-I 2007 Text mining techniques for patent analysis Information Processing & Management 43(5) pp 1216–1247 Elsveir.
[5] Feldman R and Sanger J 2007 The text mining handbook: advanced approaches in analyzing unstructured data Cambridge university press
[6] Efimenko I V 2006 Obrabotka yestestvennoyazykovykh tekstov: ontologichnost’ v lingvistike i diskurivnost’ v izvlechenii znaniy URL: http://www. rai. org/resurs/papers/kii-2006/doklad/ Efimenko_2.doc.
[7] Testelets Y G 2001 Vvedeniye v obschhiy sintaksis Moscow Study Guide Publishing House of the Russian University for the Humanities p 830 (т Russian)
[8] Vishnyakov R Y Vishnyakov Y M 2017 Vychislitel’naya teoriya semantichestkoy interpretatsii tekstov nauchno-tekhnicheskogo stilya yestestvennogo yazyka In proceedings Aktual’nyye problemy prikladnoy matematiki, informatiki i mekhaniki pp 219-226 (т Russian)
[9] Vishnyakov Y M Vishnyakov R Y 2016 Vychislitel’naya semantichestkaya interpretatsiya tekstov nauchno-tekhnicheskogo stilya In Sovremennyye naukoyemiye tekhnologii Vol 12-2 pp 236-242
[10] Vishnyakov R Y Vishnyakov Y M 2013 Sistema razbijeniya slovnoy shkaly. postrojenyiie i svoystva. Informatizatsiya i svyaz’ Vol 2 pp 9-11 (т Russian)