УДК 004.89: 004.912

G.V. Dorokhina

Institute of Artificial Intelligence MES of Ukraine and NAS of Ukraine Ukraine, 83048, c. Donetsk, Artema st., 118 b

The Algorithm of Syntactic Analysis Based on Grammatical Rules

Г.В. Дорохина

Институт проблем искусственного интеллекта МОН Украины и НАН Украины, г. Донецк Украина, 83048, г. Донецк, ул. Артема, 118 б

Алгоритм синтаксического анализа на основе грамматических правил

Г.В. Дорохіна

Інститут проблем штучного інтелекту МОН України і НАН України Україна, 83048, м. Донецьк, вул. Артема 118 б

Алгоритм синтаксичного аналізу на базі граматичних правил

Изложен разработанный алгоритм синтаксического анализа, выполняющий построение дерева зависимостей для простого распространённого неосложнённого предложения русского языка. Алгоритм находит слов пары, между которыми возможна синтаксическая связь в соответствии с грамматическими правилами. Для выделения предикативного ядра предложения используются шаблоны минимальных структурных схем предложений.

Ключевые слова: синтаксический анализ, грамматические правила, предложения русского языка, предикативное ядро предложения.

The developed algorithm is called upon to build a dependency tree by the simple expanded Russian sentence. The algorithm finds the words pairs between which the syntactic connection is possible. The algorithm finds these pairs according to grammatical rules. The minimal structure schemes are used to represent a predicative base of sentence.

Key words: syntactic analysis; grammatical rules; Russian sentence; predicative base of sentence.

Викладено розроблений алгоритм синтаксичного аналізу, який будує дерево залежностей для простого поширеного неускладненого речення російської мови. Алгоритм знаходить слів пари, між якими можливий синтаксичний зв'язок у відповідності з граматичними правилами. Для виділення предикативного ядра пропозиції використовуються шаблони мінімальних структурних схем речень.

Ключові слова: синтаксичний аналіз, граматичні правила, речення російської мови, предикативне ядро речення.

The difficulty of automatic syntactic analysis of Russian texts is caused by features of the Russian language: free word order and the phenomenon of homonymy on morphological and higher levels of language. Therefore, this problem is usually solved using the statistical methods. Under this approach a large annotated textual corpora is used. A creation of such corpora is a time-consuming task. Moreover, the solving of particular tasks of text analysis only by statistical methods says few new for the fundamental linguistics. Nowadays tools for automatic text analysis that based on linguistic methods are not developed enough. This fact points to the importance of the research.

The object of research is a simple sentence of Russian language that is expanded but not semi-composite. It means that the sentence does not contain following constructions: parentheses, absolute participial clause, homogeneous parts of the sentence. Considered in the article sentences also should not contain conjunction, connective word, particle, interjection.

The subject of research is the method of building a dependency tree by the sentence.

The goal of research is to create the syntactic analysis algorithm based on grammatical rules.

The proposed algorithm of Sentence Processing consists of following stages.

- 1. The morphological analysis of wordfroms.
- 2. Searching for pairs of potentially connected wordforms in the sentence.
- 3. Reduction a quantity of pairs of potentially connected wordforms.
- 4. Building of dependency tree.

The morphological analysis of wordfroms.

We execute the morphological analysis by using the Module of morphological analysis of Russian words RDMA_IAI. This module is a dynamic link library for Windows. The database of the module contains the paradigms of Russian words. Each paradigm is a set of wordforms connected with their lemma (dictionary form). Lemma is also considered a wordform. All wordforms are represented by pairs: spelling and morphological information (MI). In this paper, the term «morphological information» is used to mean a set of values of grammatical categories (e.g. Person: 1st person, 2nd person, 3rd person; Number: singular, plural; Case: nominative case, genitive case, etc.)

The Module RDMA_IAI able to solve following problems: normalizing of word-forms to the dictionary form (lemma); synthesis of all wordforms (paradigm) of word.

The output of function that normalizes a wordform is an array of pairs: lemma's spelling, MI of wordform. After the stage of morphological analysis the sentence S that consists of N wordforms is represented by the vector:

$$S = (s_1, \dots s_r, \dots s_N). \tag{1}$$

Here *i* means the wordform's number in the sentence, s_i is an array of interpretations s_i^i of *i*-th wordform:

$$s_{i} = \{s_{1}^{i}, \dots s_{j}^{i}, \dots s_{n_{i}}^{i}\}. \tag{2}$$

Each interpretation s_j^i is represented by pair: lemma's spelling w_j^i and morphological information m_j^i of wordform:

$$s_j^i = (w_j^i, m_j^i) . (3)$$

Searching for pairs of potentially connected wordforms in the sentence.

At the second stage we search pairs of potentially connected interpretations of wordforms. Let us introduce a relation $\eta(x,y,t)$. It accepts value 1, if the connection of type t is possible between the interpretations of wordforms x and y, such $x \in s_i$, $y \in s_j$, $i \neq j$, $t \in T$. Herewith x is the main word of syntactic connection, y is the dependent word and T is the set of syntactic connection types.

The set of syntactic connection types T is union of two subsets: Tm and Ta:

$$T = Tm \cup Ta$$
, $Tm \cap Ta = \emptyset$.

Here Ta is a set of types of relation with the minor sentence parts (categorial agreement, government, joining). The Tm is a set of relation's types between the principal sentence parts. This set we build using the minimal structure schemes (MSS) [1, p. 742-727] that declare a predicative base of Russian sentences. The following reference designations are used in the

table I. The following reference designations are used in the table. The predicative aspect of sentence is presented as: finite verb (V_f); finite copulative verb (Cop_f) (e.g. "быть" – "to be", "казаться" – "to seem", "становиться" – "to become"); infinitive (Inf), transfering the specific modal meaning; impersonal forms of copulative verb – singular or plural of copula in 3-rd face (Cop_{s3} , Cop_{pl3}). The nominative aspect of the sentence is presented using name forms and adverbs: the noun forms of nominative and instrumental cases ($N_{1/5}$) also the non propositional and propositional forms of any oblique case which are capable to be combined with the copula ($N_{2...pr}$); the adjective and passive particles forms of nominative and instrumental cases, short form and comparative degree of the adjective ($Adi_{1/5/6}$); the adverbs, which are capable to be combined with the copula (Adv_{pr}).

Table 1 – The minimal structure schemes

№	Minimal structure scheme	Examples of sentence				
1	$N_1 V_f$	Грачи прилетели.				
2	N. Con. Adi	Ночь тихая (тиха). Ночь была тихая (тихой, тиха).				
	$N_1 Cop_f Adj_{1/5/f}$	Ночь была тише дня.				
3	$N_1 Cop_f N_{I/5}$	Он – студент. Он был студент. Он был студентом.				
4	N_1 Cop_f N_{2pr}	Дом был с лифтом. Чай – с сахаром. Глаза – навыкате.				
	N_1 Cop_f N_{2pr} $/Adv_{pr}$	Глаза были навыкате. Подарок – сыну. Подарок был				
	/Auv _{pr}	сыну.				
5	$Inf V_f$	Курить воспрещалось.				
6	$Inf Cop_f N_{1/5}$	Дозвониться – проблема (было проблемой).				
	$Inj Cop_f N_{1/5}$	Любить иных – тяжелый крест.				
7		Промолчать – разумно. Промолчать – самое разумное.				
	$Inf Cop_f Adj_{1/5/f}$	Промолчать было разумно. Промолчать было самым				
		разумным.				
8	Inf $Cop_f N_{2pr}$	Молчать было в его правилах. Молчать – в его				
	$ \begin{array}{ll} Inf & Cop_f N_{2pr} \\ /Adv_{pr} & \end{array} $	правилах. Молчать было некстати. Идти трудно.				
	•	В магазин идти (было) сыну.				
9	Inf Copf Inf	Отказаться было обидеть.				
10	V_{s3}	Смеркается. Ему нездоровится.				
11	V_{pl3}	Его обидели. В классе зашумели.				
12	$Cop_{pl} N_{2pr} /Adv_{pr}$	От него были в восторге. С ним были запросто.				
13	$Cop_f N_1$	Будет дождь. Была зима. Осень.				
14	$Cop_{s3} Adj_{fsn}$	Было темно.				
15	$Cop_{pl3} Adj_{fpl}$	Результатом были довольны.				
16	Inf	Быть по-вашему.				
17	$Cop_{s3} N_{2pr} / Adv_{pr}$	Будет без осадков. Было поздно.				

Let's set out each MSS by some templates (see table 2). The template is a sequence of notations of rules which define if the relation $t \in Tm$ between x and y is possible. These rules are described in tables 3, 4. We build templates only for sentences, which predicative base consists of two words and more. Therefore, the table 2 does not contain a template for the MSS N016.

The table 3 contains simple rules that define relations $\eta(x, y, t)$ between the principal sentence parts $(t \in Tm)$. There is defined if syntactic connection is potentially possible using only following information:

- part of speech of principal word x and part of speech of dependent word y;

- sequence order of principal word x and dependent word y in the sentence ('direct' x stands before y; 'indirect' y stands before x; 'any' sequence order of principal word x and dependent word y in the sentence is unimportant);
 - is a dash must be placed between the of principal word and dependent one.

Table 2 – Templates of the minimal structure schemes

1 K1 Грачи прилетели. 2 K2 Ночь тихая (тихая). 2 KNC_L + KCAdj Ночь была тихая (тихой, тиха). 3 KNC_L + KCN Он был студент. 4 KNC_L + K3 6 Он был студентом. 4 KNI Preph K Pr Nobj Чай – с сахаром. 4 K Nom Obj Подарок – сыну. 5 KN C L + KCP + K Pr Nobj Дом был с лифтом. 6 KCI Nom + K3 6 Дозвонитья было проблемой. 6 КСI Nom + KCN Дозвониться была проблема. 7 КСI Nom + KCAdj Промолчать было разумно. 8 КI Pred + K Pr Nobj Молчать в его правилах. 8 КI NomObj В магазин идти сыну. 8 КI Pred Молчать было в его правилах. 8 КI Pred Молчать было в его правилах. 8 КI NomObj В магазин идти сыну. 8 КI NomObj В магазин идти сыну. 9 КСI Nom + KCP Молчать было в его правилах. 12 КСI Nom + KCP Молчать было некстати. <t< th=""><th>MSS</th><th>MSS template</th><th colspan="3">Example</th></t<>	MSS	MSS template	Example		
2 KNC_L + KCAdj Ночь была тихая (тихой, тиха). 8 OH - студент. 8 KNC_L + KCN OH был студент. 6 KNC_L + K3 6 OH был студентом. 6 KNI Prep+ K Pr Nobj Чай - с сахаром. 6 KN Pred Глаза - навыкате. 6 K Nom Obj Подарок - сыну. 6 K Nom L + KCP+ K Pr Nobj Дом был с лифтом. 6 K CL Nom + K3 6 Дозвонитья было проблемой. 6 КСІ Nom + KCN Дозвонитья было проблемой. 6 КСІ Nom + KCN Дозвониться была проблема. 6 КСІ Nom + KCN Дозвониться было проблемой. 7 КСІ Nom + KCAdj Промолчать было проблема. 8 КІ Pred + K Pr Nobj Молчать - разумно. 8 КІ Pred + K Pr Nobj Молчать - в его правилах. 8 КІ Pred Молчать было в его правилах. 8 КСІ Nom + КСР Молчать было в его правилах. 9 КСІ Nom + КСІ Отказаться было обидеть. 9 КСР + К Pr Nobj От него были в восторге. <td>1</td> <td>_K1</td> <td colspan="3">Грачи прилетели.</td>	1	_K1	Грачи прилетели.		
Hoчь тише дня.		_K2	Ночь тихая (тиха).		
K3 Он – студент. KNC L + KCN Он был студент. KNC L + K3 6 Он был студентом. KNI Prep+ K Pr Nobj Чай – с сахаром. KN Pred Глаза – навыкате. KN Nom_Obj Подарок – сыну. KNC L+ KCP+ K Pr Nobj Дом был с лифтом. KNC L+ KCP Pr Глаза были навыкате. 5 K5 Курить воспрещалось. KCI Nom + K3 6 Дозвониться была проблема. K6 Дозвониться она проблема. K6 Дозвониться – проблема. K7 Промолчать было разумно. K1 Pred + K Pr Nobj Молчать – в его правилах. K1 Pred + K Pr Nobj Молчать некстати. K1 Pred Молчать было в его правилах. КСI Nom + КСР + Молчать было некстати. K CI Nom + KC Pr Молчать было некстати. КСI Nom + КС Рг Отказаться было обидеть. К9 Отказаться обыле в восторге. КСР - К Pr От него были в восторге. КС Pr С ним были запросто. КСР - К Р Nobj Без осадков. КСР + К Pr Nobj Без осадков. Было поздно.	2	_KNC_L + _KCAdj	Ночь была тихая (тихой, тиха).		
3 KNC_L + KCN Он был студент. KNC_L + K3_6 Он был студентом. KN1 Prep+ K Pr Nobj Чай – с сахаром. KN Pred Глаза – навыкате. K Nom_Obj Подарок – сыну. KNC_L + KCP+ K Pr Nobj Дом был с лифтом. KNC_L + KC Pr Глаза были навыкате. 5 K5 KCI Nom + K3_6 Дозвонитья было проблема. K6 Дозвониться – проблема. K6 Дозвониться – проблема. K7 Промолчать было разумно. K1 Pred + K Pr Nobj Молчать – в его правилах. K1 NomObj В магазин идти сыну. K1 Pred Молчать некстати. KCI Nom + KCP + Молчать было в его правилах. K Pr Nobj Молчать было некстати. 9 КСІ Nom + КС Рг Молчать было обидеть. 4 КО Р Отказаться было обидеть. КР КОР + К Pr Nobj От него были в восторге. КСР - С ним были запросто. КСР + К Pr Nobj Будет дождь. КСР + К Pr Nobj Без осадков. КСР + К Pr Nobj Будет без осадков.<			Ночь тише дня.		
KNC L + K3 6 Он был студентом. KN1 Prep+ K Pr Nobj Чай – с сахаром. KN Pred Глаза – навыкате. KN Pred Подарок – сыну. KN C L + KCP+ K Pr Nobj Дом был с лифтом. KNC L + KC Pr Глаза были навыкате. 5 K5 Курить воспрещалось. KCI Nom + K3 6 Дозвониться было проблема. K6 Дозвониться было разумно. K6 Дозвониться – проблема. K7 Промолчать было разумно. K1 Pred + K Pr Nobj Молчать – в его правилах. K1 NomObj В магазин идти сыну. K1 Pred Молчать некстати. K2I Nom + KCP Нолчать было в его правилах. K1 Pred Молчать было некстати. K1 Pred Молчать было некстати. К1 Nom + KC Pr Молчать было некстати. К2 Nom + KC Pr Молчать было обидеть. К2 Nom + KC Pr Отказаться было обидеть. К2 Pr От него были в восторге. КС Pr С ним были запросто. К2 Pr Nobj Будет дождь. К2 Pr Nobj Будет без осадков.		_K3	Он – студент.		
KN1 Prep+ K Pr Nobj Чай – с сахаром. KN Pred Глаза – навыкате. KN Pred Глаза – навыкате. KN Pred Глаза – навыкате. KNC L+ KCP+ K Pr Nobj Дом был с лифтом. KNC L+ KC Pr Глаза были навыкате. KNC L+ KC Pr Глаза были навыкате. KS Курить воспрещалось. KCI Nom + K3 6 Дозвониться было проблема. K6 Дозвониться – проблема. K6 Дозвониться – проблема. K7 Промолчать было разумно. K1 Pred + K Pr Nobj Молчать – в его правилах. K1 Pred + K Pr Nobj Молчать некстати. K1 Pred Молчать было в его правилах. K1 Pred Молчать было в его правилах. K1 Pre Nobj Молчать было некстати. K1 Nom + KC Pr Молчать было некстати. K2 Nom + KC Pr Молчать было обидеть. K9 Отказаться – обидеть. K2 Pr Nobj От него были в восторге. КСР - K Pr Nobj Было темно. К2 Pr Nobj Без осадков. КСР + K Pr Nobj Было поздно.	3	_KNC_L + _KCN	Он был студент.		
KN Pred Глаза – навыкате. K Nom Obj Подарок – сыну. KNC_L+_KCP+_K_Pr Nobj Дом был с лифтом. KNC_L+_KCP Глаза были навыкате. KS Курить воспрещалось. KCI Nom + K3 6 Дозвонитья было проблемой. K6 Дозвониться была проблема. K6 Дозвониться – проблема. K7 Промолчать было разумно. K1 Pred + K Pr Nobj Молчать – в его правилах. K1_NomObj В магазин идти сыну. K1_Pred Молчать некстати. KCI_Nom + KCP Молчать было в его правилах. K Pr Nobj КСІ Nom + БСР КСІ Nom + КСР Отказаться было обидеть. КР КОР К Рг Nobj От него были в восторге. КСР - К Рг Nobj От него были в восторге. КСР - К Рг Nobj Было темно. КР Р Nobj Без осадков. КСР - К Рг Nobj Будет без осадков. КСР - К Рг Nobj Было поздно.		_KNC_L + _K3_6	Он был студентом.		
4 K Nom_Obj Подарок – сыну. KNC_L+_KCP+_K_Pr_Nobj Дом был с лифтом. KNC_L+_KC_Pr Глаза были навыкате. 5 K5 Курить воспрещалось. 6 КСІ_Nom+_KCN Дозвониться было проблема. 7 КСІ_Nom+_KCAdj Промолчать было разумно. 8 КІ_Pred+_K_Pr_Nobj Молчать – в его правилах. КІ_NomObj В магазин идти сыну. КІ_Pred Молчать некстати. КСІ_Nom+_KCP+ Молчать было в его правилах. КСІ_Nom+_KCP Молчать было некстати. 9 КСІ_Nom+_KC 12 КСР+_K_Pr_Nobj Отказаться было обидеть. 12 КСР+_K_Pr_Nobj От него были в восторге. 13 КСN Будет дождь. 14 К14 Было темно. 15 К15 Результатом были довольны. 16 КСР+_K_Pr_Nobj Будет без осадков. КСР+_K_Pr_Nobj Было поздно.		_KN1_Prep+_K_Pr_Nobj	Чай – с сахаром.		
KNC L+ KCP+ K Pr Nobj Дом был с лифтом. KNC L+ KC Pr Глаза были навыкате. KS Курить воспрещалось. KCI Nom + K3 6 Дозвониться была проблема. K6 Дозвониться – проблема. K6 Дозвониться – проблема. 7 КСІ Nom + KCAdj Промолчать было разумно. КІ Pred + K Pr Nobj Молчать – в его правилах. КІ NomObj В магазин идти сыну. КІ Pred Молчать было в его правилах. КГ Nom + КСР + Молчать было в его правилах. КСІ Nom + КСР КСІ Nom + КСІ Отказаться было обидеть. 9 КСІ Nom + КСІ КСР - КР Nobj От него были в восторге. КСР - КР Nobj От него были в восторге. КСР - КР Nobj Было темно. 15 К15 Результатом были довольны. КР - Nobj Будет без осадков. КСР + К Pr Nobj Будет без осадков. КСР - КР Поbj Было поздно.		_KN_Pred	<u> </u>		
KNC L + KC Pr Глаза были навыкате. 5 K5 Курить воспрещалось. 6 КСІ Nom + K3 6 Дозвониться было проблемой. 7 КСІ Nom + КСАфі Промолчать было разумно. 7 КСІ Nom + КСАфі Промолчать – вего правилах. 8 КІ Pred + К Pr Nobj Молчать – вего правилах. 8 КІ NomObj В магазин идти сыну. 8 КІ Pred Молчать было вего правилах. 6 КСІ Nom + КСР Молчать было некстати. 8 КСІ Nom + КСР Молчать было некстати. 9 КСІ Nom + КСІ Отказаться было обидеть. 9 КСР + К Pr Nobj От него были в восторге. 12 КСР + К Pr Nobj От него были в восторге. 13 КСN Будет дождь. 14 К14 Было темно. 15 К15 Результатом были довольны. 17 КСР + К Pr Nobj Будет без осадков. КСР + К Pr Nobj Было поздно.	4		Подарок – сыну.		
5 K5 Курить воспрещалось. 6 КСІ Nom + КЗ 6 Дозвонитья была проблема. 7 КСІ Nom + КСАdj Промолчать было разумно. 8 КІ Рred + К Рг Nobj Молчать – разумно. 8 КІ Pred + К Рг Nobj Молчать – в его правилах. 8 КІ Pred Молчать некстати. 9 КСІ Nom + КСР + Молчать было в его правилах. 10 КСІ Nom + КСР Молчать было обидеть. 12 КСР - К Рг Nobj От него были в восторге. 12 КСР - К Рг Nobj От него были в восторге. 13 КСN Будет дождь. 14 К14 Было темно. 15 К15 Результатом были довольны. 17 КСР + К Рг Nobj Будет без осадков. КСР + К Рг Nobj Будет без осадков. КСР + К Рг Nobj Было поздно.		_KNC_L+_KCP+_K_Pr_Nobj	Дом был с лифтом.		
КСІ Nom + КЗ 6 Дозвонитья было проблемой. КСІ Nom + КСN Дозвониться была проблема. КСІ Nom + КСАфі Промолчать было разумно. КІ Ргеф + К Рг Nobi Молчать – в его правилах. КІ Рred Молчать некстати. КІ Рred Молчать было в его правилах. КІ Рred Молчать было в его правилах. КСІ Nom + КСР + Молчать было в его правилах. КСІ Nom + КСР Рг Уксі Nom + КСІ Отказаться было обидеть. Отказаться – обидеть. От него были в восторге. КСР - К Рг Nobi От него были запросто. КСР Будет дождь. КР Р Nobi Без осадков. КСР + К Рг Nobi Будет без осадков. КСР - К Рг Nobi Будет без осадков. КСР - К Рг Nobi Было поздно.		_KNC_L + _KC_Pr	Глаза были навыкате.		
6 КСІ Nom + КСN Дозвониться была проблема. 7 КСІ Nom + КСАфі Промолчать было разумно. 8 КІ Pred + К Pr Nobi Молчать – в его правилах. 8 КІ NomObi В магазин идти сыну. 8 КІ Pred Молчать некстати. 10 КСІ Nom + КСР + КРг Nobi Молчать было в его правилах. 10 КСІ Nom + КС Рг Молчать было некстати. 10 КСІ Nom + КС Рг Отказаться было обидеть. 12 КСР + К Рг Nobi От него были в восторге. 13 КСР С ним были запросто. 13 КСN Будет дождь. 14 К14 Было темно. 15 К15 Результатом были довольны. КР Р Nobi Без осадков. КСР + К Рг Nobi Будет без осадков. КСР + К Рг Nobi Было поздно.	5	_K5	Курить воспрещалось.		
K6 Дозвониться – проблема. КСІ Nom + КСАфі Промолчать было разумно. КІ Pred + К Pr Nobi Молчать – в его правилах. КІ Pred Норі КІ Pred Молчать некстати. КІ Pred Молчать было в его правилах. КІ Рг Nobi Отказаться было обидеть. Отказаться обидеть. Отказаться – обидеть. КСР - К Рг Nobi От него были в восторге. КСР - К Рг Отказаться обили запросто. КК Рг Поbi Было темно. КСР - К Рг Nobi Было поздно.		_KCI_Nom + _K3_6	Дозвонитья было проблемой.		
7 KCI_Nom + KCAdj Промолчать было разумно. KI_Pred + K_Pr_Nobj Молчать – в его правилах. KI_NomObj В магазин идти сыну. KI_Pred Молчать некстати. KCI_Nom + KCP + Молчать было в его правилах. K_Pr_Nobj Молчать было некстати. SKI_Nom + KC_Pr Молчать было некстати. Отказаться было обидеть. Отказаться – обидеть. КРР - К_Pr_Nobj От него были в восторге. КС_Pr С ним были запросто. ККР - К-Рг Было темно. К_Pr_Nobj Без осадков. КС_Pr Было поздно.	6	_KCI_Nom + _KCN	Дозвониться была проблема.		
К7 Промолчать – разумно. 8 KI Pred + K Pr Nobj Молчать – в его правилах. KI NomObj В магазин идти сыну. KI Pred Молчать некстати. KCI Nom + KCP + Молчать было в его правилах. K Pr Nobj Отказаться было обидеть. KOI Nom + KCI Отказаться было обидеть. КОР + К Pr Nobj От него были в восторге. КСР + К Pr Nobj От него были запросто. Вудет дождь. Отказаться – обидеть. КСР - К Рг Поврј Было темно. КСР - К Рг Nobj Без осадков. КСР + К Рг Nobj Будет без осадков. КСР - К Рг Поврј Было поздно.		_K6	Дозвониться – проблема.		
К7 Промолчать – разумно. КІ_Pred + K_Pr Nobj Молчать – в его правилах. КІ_Pred Молчать некстати.	7	_KCI_Nom + _KCAdj	Промолчать было разумно.		
8 KI_NomObj В магазин идти сыну. KI_Pred Молчать некстати. KCI_Nom + _KCP + _Mолчать было в его правилах. K_Pr_Nobj КСI_Nom + _KC_Pr Молчать было некстати. 9 KCI_Nom + _KCI Отказаться было обидеть. K9 Отказаться – обидеть. 12 КСР + _K_Pr_Nobj От него были в восторге. КС_Pr С ним были запросто. 13 КСN Будет дождь. 14 К14 Было темно. 15 К15 Результатом были довольны. K_Pr_Nobj Без осадков. KCP + _K_Pr_Nobj Будет без осадков. KC_Pr Было поздно.	,	_K7	Промолчать – разумно.		
8 KI Pred KCI Nom + KCP + MONЧАТЬ БЫЛО В ЕГО ПРАВИЛАХ. K Pr Nobj MOЛЧАТЬ БЫЛО В ЕГО ПРАВИЛАХ. KCI Nom + KC Pr MOЛЧАТЬ БЫЛО НЕКСТАТИ. 9 KCI Nom + KCI ОТКАЗАТЬСЯ БЫЛО ОБИДЕТЬ. K9 ОТКАЗАТЬСЯ — ОБИЛИ В ВОСТОРГЕ. КС Pr С ним были запросто. 13 KCN БУДЕТ ДОЖДЬ. 14 K14 15 K15 К Pr Nobj Без осадков. K Pr Nobj Будет без осадков. КСР + К Pr Nobj Будет без осадков. КСР - К Рг Было поздно.			Молчать – в его правилах.		
KCI_Nom + KCP + KCP + КСР + КСР Nobj Нолчать было в его правилах. KCI_Nom + KC Pr Молчать было некстати. 9 KCI_Nom + KCI Отказаться было обидеть. K9 Отказаться – обидеть. 12 КСР + К Pr Nobj От него были в восторге. КС Pr С ним были запросто. Будет дождь. 14 К14 Было темно. 15 К15 Результатом были довольны. К Pr_Nobj Без осадков. Будет без осадков. КСР + К Pr_Nobj Будет без осадков. Было поздно.		_KI_NomObj	В магазин идти сыну.		
_KCI_Nom + _KCP + _KPr_Nobj Молчать было в его правилах. 9 _KCI_Nom + _KC_Pr Молчать было некстати. 9 _KCI_Nom + _KCI	Q		Молчать некстати.		
KCI_Nom + KC_Pr Молчать было некстати. 9 KCI_Nom + KCI Отказаться было обидеть. 12 KCP + K_Pr_Nobj От него были в восторге. 12 KCP + K_Pr_Nobj С ним были запросто. 13 KCN Будет дождь. 14 K14 Было темно. 15 K15 Результатом были довольны. K_Pr_Nobj Без осадков. KCP + K_Pr_Nobj Будет без осадков. KC_Pr Было поздно.	0		Молчать было в его правилах.		
9 KCI_Nom + KCI Отказаться было обидеть. 12 KCP + K_Pr_Nobj От него были в восторге. 12 KCPr С ним были запросто. 13 KCN Будет дождь. 14 K14 Было темно. 15 K15 Результатом были довольны. K Pr_Nobj Без осадков. KCP + K_Pr_Nobj Будет без осадков. KC_Pr Было поздно.					
K9 Отказаться – обидеть. 12 КСР + К Pr Nobj От него были в восторге. КС Pr С ним были запросто. 13 КСN Будет дождь. 14 К14 Было темно. 15 К15 Результатом были довольны. К Pr Nobj Без осадков. КСР + К Pr Nobj Будет без осадков. КС Pr Было поздно.		_KCI_Nom + _KC_Pr	Молчать было некстати.		
K9 Отказаться – обидеть. 12 КСР + _K_Pr_Nobj От него были в восторге. КС Pr С ним были запросто. 13 КСN Будет дождь. 14 К14 Было темно. 15 К15 Результатом были довольны. КР P Nobj Без осадков. КСР + _K_Pr_Nobj Будет без осадков. КС Pr Было поздно.	9	_KCI_Nom + _KCI	Отказаться было обидеть.		
12 _KC_Pr С ним были запросто. 13 _KCN Будет дождь. 14 _K14 Было темно. 15 _K15 Результатом были довольны. _K_Pr_Nobj Без осадков. _KCP+_K_Pr_Nobj Будет без осадков. _KC_Pr Было поздно.		l <u>—</u>			
_KC_Pr С ним были запросто. 13 _KCN Будет дождь. 14 _K14 Было темно. 15 _K15 Результатом были довольны. _K_Pr_Nobj Без осадков. _KCP + _K_Pr_Nobj Будет без осадков. _KC_Pr Было поздно.	12				
14 K14 Было темно. 15 K15 Результатом были довольны. K Pr Nobj Без осадков. KCP + K Pr Nobj Будет без осадков. KC Pr Было поздно.			-		
15 _K15 Результатом были довольны. _K_Pr_Nobj Без осадков. _KCP + _K_Pr_Nobj Будет без осадков. _KC_Pr Было поздно.					
K_Pr_Nobj Без осадков. KCP + _K_Pr_Nobj Будет без осадков. KC_Pr Было поздно.		_	Было темно.		
17 _KCP + _K_Pr_Nobj Будет без осадков. _KC_Pr Было поздно.	15	-	Результатом были довольны.		
			Без осадков.		
_KC_Pr Было поздно.	17	_KCP+_K_Pr_Nobj	Будет без осадков.		
К17 Цветов было!	' '	_KC_Pr	* *		
		_K17	Цветов было!		

Symbol * (see tab. 3, 5) is used in notation of rules in order to specify that case of depend word y is hang on the preposition x.

Notation of rule	Part of speech of principal (x) and dependent (y) words		Dash	Sequence order of words
K5	x	Verb (infinitive)		onv
_ K 3	У	Verb (non infinitive)		any
K9	х	Verb (infinitive)	+	any
_K3	у	Verb (infinitive)	ı	
KI Pred	х	Verb (infinitive)		direct
_KI_FIEU	у	Adverb		direct
KI Prep	х	Verb (infinitive)	+	direct
_Ki_ricp	у	Preposition		
_KC_Pr	х	Copula		onv
_KC_F1	у	Adverb		any
KN Pred		Noun	+	any
_KIN_I ICU		Adverb	ı	
KCI	x	Copula		direct
_KCI	У	Verb (infinitive)		direct
KCI Nom	x	Copula		indirect
_KCI_NOIII	У	Verb (infinitive)	manect	
KCP	х	Copula		direct
_1XC1	у	Preposition		direct
K Pr Nobj *	x	Preposition		direct
_IX_I I_INUUJ	У	Noun		uncci

Table 3 – Notation of simple rules that define syntactical connections $t \in Tm$

The table 4 contains more complex rules that define relations $\eta(x, y, t)$ between the principal sentence parts $(t \in Tm)$. These rules also use morphological information of principal and dependent words, sequence order of principal and dependent word in the sentence, acceptable parts of speech of words standing between principal word and dependent one (separator), is a dash must be placed between the of principal and dependent words. We apply following reference designations for grammatical categories of morphological information and their values:

- Number (N_x, N_y) takes the values: 'singular' ('sing.') and plural ('pl.');
- Case (C_x, C_y) takes the values: 'nominative' ('nom.'), 'genitive' ('gen.'), 'dative' (dat.), 'accusative' ('acc.'), 'instrumental' ('in.'), 'locative' ('loc.');
 - Tense (T_x, T_y) takes the values: 'past', 'present' ('pres.'), 'future' ('fut');
 - Gender (Gx, Gy);
 - Person (F_x, F_y) takes the values: 'first'(1), 'second' (2), 'third'(3);
- Form of adjective $(AdjF_x, AdjF_y)$ takes the values: 'the positive degree', 'the comparative degree' ('comp.'), 'the superlative degree', 'short form' ('short').

These rules also include: logical operators 'AND' (&), 'OR' (V), 'NOT'(!); wOrder – the sequence order of words ('direct', 'indirect').

Let us consider the rules that belong to the set Ta. They identify relations with the minor sentence parts: categorial agreement, government, joining (see tab. 5, 6).

Table 4 – Notation of complex rules that define syntactical connections $t \in Tm$

Notation of rule	Part of speech of principal (x) and dependent (y) words		Sepa- rators	Rule	
	x V	Terb (non infinitive)	Any	$N_x = N_y \& C_y = \text{`nom.'} \& ((T_x = \text{`past'} \lor F_x = 3) \& (N_x = \text{`sing.'} \& G_x = G_y \lor N_x = \text{`pl.'}) \lor T_x = \text{`fut.'} \lor T_x = \text{`pres.'})$	
_K1		Yerb (non infinitive) ersonal pronoun	Any	$N_x=N_y\&C_y=$ 'nom.' & $(T_x=$ 'past' $\vee F_x=3)$ & $(N_x=$ 'sing.' & $G_x=G_y\vee N_x=$ 'pl') $\vee F_x=F_y$ & $(T_x=$ 'fut.' $\vee T_x=$ 'pres.'))	
		rerb (non infinitive) ardinal numeral	Any	C_y = 'nom.'	
_K7		Terb (infinitive) djective	Adverb	$(C_y = \text{`nom.'} \lor AdjF_y = \text{`short'} \lor AdjF_y = \text{`comp.'}) & Dash$	
_KI_NomObj		Terb (infinitive) Toun VAdjective	Any	C_y = 'nom.'	
_K_NomObj		oun oun	!Verb	C_y = 'nom.' C_y = 'nom.'	
_K14		opula djective	!Verb	$N_x = N_y$ & $N_x = \text{sing.'}$ & $F_x = 3$ & $AdjF_y = \text{short'}$	
_K15		opula djective	!Verb	$N_x = N_y$ & $N_x = \text{`pl.'}$ & $F_x = 3$ & $AdjF_y = \text{`short'}$	
_KCAdj		opula djective	!Verb	$((C_y = \text{`nom.'} \lor C_y = \text{`in.'}) \& (T_x = \text{`past'} \& G_x = G_y \lor T_x = \text{`fut.'} \lor T_x = \text{`pres.'})) \lor AdjF_y = \text{`short'} \lor AdjF_y = \text{`comp.'}$	
_K3_6	-	opula Ioun	Any	$N_x = N_y$ & $C_y = $ in.	
KCN	x C	opula	!Verb	wOrder='direct'& N_x = N_y & C_y ='nom.'& ((T_x ='past' $\vee F_x$ =3)&(N_x ='sing.'& G_x = G_y	
_		opula		$\vee N_x = \text{`pl'}) \vee T_x = \text{`fut.'} \vee T_x = \text{`pres.'})$ wOrder='indirect' & $N_x = N_v \& C_v = \text{`nom.'}$	
_KNC_L		oun	!Verb	&($(T_x = \text{`past'} \lor F_x = 3)$ &($N_x = \text{`sing.'}$ & $G_x = G_y \lor N_x = \text{`pl'}) \lor T_x = \text{`fut.'} \lor T_x = \text{`pres.'})$	
_KC_Pr		opula Ioun	!Verb	C_y ='gen.' $\vee C_y$ ='dat.' $\vee C_y$ ='acc.'	
_K2	x N	oun	!Verb	C_x ='nom.'&(N_x = N_y & G_x = G_y & (C_y ='nom.' \vee	
K3	x N	djective Ioun	!Verb	$AdjF_v$ ='short') $\vee AdjF_v$ ='comp.') C_x ='nom.' & C_v ='nom.' & $Dash$	
_KN1_Prep	x N	oun oun	any	C_x ='nom.' & Dash	
K6	x V	reposition Terb (infinitive) Toun	!Verb	C_y ='nom.' & Dash	
_K17	x C	opula Toun	!Verb	$C_y \neq$ 'nom.'	

Table 5 – Notation of simple rules that define syntactical connections $t \in Ta$

Notation of rule	Part of speech of principal (x) and dependent (y) words		Separators	Sequence order of words	
,	x	Verb			
t_{a1}	y	Adverb	Any	Any	
+	x	Verb	- !Verb	Any	
t_{a2}	У	Adverbial participle	! VEID		
<i>t</i> .	x	Verb	- !Verb	Anv	
t_{a3}	y	Verb (infinive)	: V C10	Any	
t_{a4}	x	Verb	- !Verb	Any	
1 _a 4	y	Preposition	: V CI U	Any	
t_{a5}	x	Noun	Adjective ∨ Participle	Anv	
i _{a5}	y	Preposition	Adjective v i articipie	Any	
t_{a6}	x	Noun	Adjective ∨ Participle	Direct	
<i>i</i> a6	y	Verb (infinive)	Adjective v i articipie		
t_{a7}	x	Adjective	None	Any	
la/	y	Adverb	None		
t_{a8}	x	Adjective	- None	Direct	
1 a8	y	Preposition	TVOIC	Direct	
t_{a9}	x	Adjective	- Adverb	Direct	
<i>i</i> a9	y	Verb (infinive)	Auvelo	Direct	
<i>t</i> _{a10} *	x	Preposition	Adjective ∨ Participle	Direct	
ι_{a10}	y	Noun	Aujective V ranticipie		
<i>t</i>	x	Adverb	- None	Direct	
t_{a11}	у	Adverb	TVOILC		
t 10	x	Adverb	- None	Direct	
t_{a12}	у	Preposition	TYONG		

Table 6 – Notation of complex rules that define syntactical connections $t \in Tm$

Notation of rule	Part of speech of principal (x) and dependent (y) words		Separators	Sequence order of words	Rule
t_{a13}	v	Verb Noun	Any	Any	$C_y \neq$ 'nom.'
t_{a14}	$\frac{x}{y}$	Noun Adjective	Adverb	Any	$(C_x=C_y\lor AdjF_y=\text{`comp.'})\&$ $N_1=N_2\&(N_1=\text{`pl.'}\lor G_1=G_2)$
t_{a15}	<i>x y</i>	Noun Noun	Adjective ∨ Participle	Direct	$C_y \neq$ 'nom.'
t_{a16}	<i>x y</i>	Adjective Noun	Adjective V Participle	Direct	C_y ='gen.' \vee C_y ='dat.' \vee C_y ='in.'

The rules presented in the tables 3-6 define the set of threes (x, y, t) for which $\eta(x, y, t) = 1$. It is possible to express all types of syntactic connections using threes (x, y, t). A syntactic connection between x and y that is achieved by connective word z (a prepositional government) we express using two threes: (x, z, t_1) , (z, y, t_2) .

We search for pairs of potentially connected wordforms in the sentence using these rules. Let us save founded pairs of potentially connected wordforms in the sentence S (2)-(3) as a set R of threes (x, y, t):

$$R = \{(x, y, t)\} : \eta(x, y, t) = 1, x \in s_i, y \in s_i, i \neq j$$
 (4)

Reduction a quantity of pairs of potentially connected wordforms.

The set of first components of this threes set will be marked as A (the set of principal words), the set of second components the threes set will be marked as B (the set of dependent words).

$$A = \{x\} : \exists (x, y, t) \in R$$

$$B = \{y\} : \exists (x, y, t) \in R$$
(5)

We can build a dependence tree for the sentence if all it's wordforms are connected with one or more wordforms by syntactic connection.

Let us introduce the criterion of sentence's connectedness: "At least one interpretation of each wordform must belong to the set of principal words or to the set of dependent words."

$$\forall i = \overline{1, N} \ \exists z \in s_i : z \in (A \cup B)$$
 (6)

The sentence, which doesn't satisfy criterion (6), is not syntactically connected. It is possible that the sentence is written with error. The analysis of such sentences stops.

Let's form vector S' that describe sentence S. Each element s_i' of vector S' is a subset of s_i . Members of s_i' should belong to the set of principal words A or to the set of dependent words B.

$$S' = (s'_1, \dots s'_i, \dots s'_N),$$

$$s'_i \subseteq s_i : \forall z \in s'_i \ z \in (A \cup B).$$
(7)

Thus we reduce a quantity of wordforms interpretations due to using as sentence's representation the vector S^\prime instead of vector S .

Using a sentence's representation S' we build a set of morphological sentence's markings. The D set of morphological sentence's markings can be received as the Cartesian product of sets which are the elements of a vector S'.

$$D = s'_{1} \times ... \times s'_{i} \times ... \times s'_{N},$$

$$D = \{d_{k} : d_{k} = (d_{1}^{k}, ..., d_{k}^{k}, ...d_{N}^{k})\}$$
(8)

Most of morphological sentence's markings d_k are invalid. We reject such morphological sentence's markings using outlined below criterions.

Let's create following sets in order to apply the criterion of sentence's connectedness to the morphological sentence's marking d_k : F_k – a set of components of morphological sentence's marking d_k ; R_k – a subset of set R (4) which contain an information about pairs of potentially connected by syntactic relationship interpretations of wordforms d_i^k and d_j^k ; A_k – a set of the set of principal words of these pairs; B_k – a set of dependent words.

$$F_{k} = \{d_{i}^{k}\}, i = \overline{1, N}$$

$$R_{k} = \{(x, y, t) : (x, y, t) \in R, x \in F_{k}, y \in F_{k}\}$$

$$A_{k} = \{x : (x, y, t) \in R_{k}\}$$

$$B_{k} = \{y : (x, y, t) \in R_{k}\}$$
(9)

The following condition allows checking if the morphological sentence's marking d_k satisfies the criterion of sentence's connectedness.

$$d_i^k \in (A_k \cup B_k), i = \overline{1, N} \tag{10}$$

The morphological sentence's marking d_k which does not satisfy criterion (10) is unacceptable.

For the sentences considered in the paper it is possible to build a dependency tree if the morphological sentence's marking d_k satisfy following criterion: "A count of wordforms which belong to the set of principal words but not belong to the set of dependent words should not be more then 1."

$$\mid A_k \setminus B_k \mid \le 1 \tag{11}$$

The next criterion deals with a prepositional government. Let *P* is the set of prepositions of Russian language. The criterion is following: "Prepositions belong to both the set of principal words of the sentence and the set of dependent words of the sentence."

$$\forall z \in P \cap F_k \ z \in B_k \cap A_k \tag{12}$$

We will continue further analysis of morphological sentence's marking d_k if it satisfies the criteria (10)- (12).

$$Rm_i = \{(x, y, t) : (x, y, t) \in R_k, t = t_i, t_i \in h\}$$

Building of dependency tree.

The pair (F_k, R_k) describes the directed edge-labeled graph G_k . The set F_k is a nodes set of this digraph. The set Rk is a set of labeled edges (x,y,t). Here pair (x,y) is an arc from x to y and t is a label of the edge. Required dependence tree is a subgraph of digraph G_k .

But not all of them are the trees of syntax subordination (TSS). The decision on the reasonableness of morphological sentence marking and admissibility of separate connections from the R_k set will be made in terms of the next criteria.

Digraphs simple connectedness which is designated by F_k and R_k connections subsets not contradicting the minimal structure scheme templates.

Equality of 1 demidegree of these digraphs peaks stopping. The Rk correspondence to the h minimal structure scheme template is analyzed. For this the Rm={Rmi} set will be put, where Rmi ∈ Rk is one type and this type is included to the h template

$$Rm_{i} = \{(x, y, t) : (x, y, t) \in R_{k}, t = t_{i}, t_{i} \in h\}$$
(12)

If |Rm| < |h| the sentence doesn't correspond to h.

Let we put $RM = \{rm_V\}$, where $\subseteq Rm_1 \times ... \times Rm_i \times ... \times Rm_l$ and $rm_V = ((x_1, y_1, t_1), ..., (x_l, y_l, t_l))$: if l > 1 $x_1 = x_2, \forall i > 1$ $x_i + 1 = y_l$.

The rm_V element is a base for the TSS creation by the h template. Let $g = \{(x, y, t)\}$, where (x, y, t) is the elements of rm_V vector. It is necessary to add the minor connections of the c set.

$$c = \{(x', y', t') : (x', y', t') \in (R_k \cap Ta) \neg \exists (x, y, t) \in g : (y' = y \lor x' = x)\}$$
(13)

Let's mark $g' = g \cup c$. If the digraph (g', R_k) is not single-connected, so it's impossible to create the correct TSS. Otherwise it's necessary to solve a peaks problem with an in-degree more than 1. For each such peak only one connection is left, according to the requirement, that the way length from the root vertex to it is maximum. If there is one peak in which the n of competitive connections is brought on the ways of identical length, it is considered that there is a syntax homonymy and all the n connections are correct and n different TSS corresponds to the pair (Fk, g').

The list of syntactically connected words pairs is a recognized correct g' connections combining, which are built on all Rm_i for every F_k and h template.

Experimental system of sentence parsing.

As a result of developed algorithm programs implementation is created the experimental system of sentence parsing. Input system's data is the text, which consists of Russian words, the sentences are ended with punctuation marks (".", "!", "?", "..."), all the sentence's words are in the lower register (except the first word). The text, which is input from file, should be in the Windows-1251 coding. See examples of system responses in tab. 7.

Table 7 – Examples of responses

Sentence		x, y, t	ДСП		
Schichee	x	y	t	деп	
Meranganyan	сидели	Мы	_K1	<u>⊟</u> -сидели	
Мы сидели на восьмом	сидели	на	Upr	— Мы (_K1) ⊟- на (Upr)	
этаже	на	этаже	Upr	⊟ этаже (Upr)	
31 43 110	этаже	восьмом	Sogl	восьмом (Sogl)	
Июльская	ночь	Июльская	Sogl	⊟-была	
ночь была	была	ночь	_KNC_L	⊟- ночь (_KNC_L) — Июльская (Sogl)	
тихая	была	тихая	_KCAdj	тихая (_KCAdj)	
Парень был	был	Парень	_KNC_L	⊟- был Парень (_KNC_L)	
спортсменом	был	спортсменом	_K3_6	спортсменом (_К3_6)	
Полеболого бол	был	Дом	_KNC_L	⊟-был	
Дом будет без лифта	был	без	_KCP	— Дом (_KNC_L) ⊟- без (_KCP)	
лифта	без	лифта	_K_Pr_Nobj	лифта (_K_Pr_Nobj)	
Курить воспрещалось	воспрещалось	Курить	_K5	⊡- воспрещалось Курить (_K5)	
Дозвониться	Дозвониться	проблемой	_KCI_Nom	⊡-было	
было проблемой	было	Дозвониться	_KC_Pr	— Дозвониться (_KCI_Nom) — проблемой (_K3_6)	
Промолчать	было	Промолчать	_KCI_Nom	⊡- было	
было разумнее	было	разумнее	_KCAdj	— Промолчать (_KCI_Nom) — разумнее (_KCAdj)	
Уступить	было	Уступить	_KCI_Nom	⊟- было	
было в	было	В	_KCP	— Уступить (_KCI_Nom) ⊟-в (_KCP)	
правилах	В	правилах	_K_Pr_Nobj	□ В (_KCP) □ правилах (_K_Pr_Nobj)	

Список литературы

1. Белошапкова В.А. Современный русский язык. М.: Азбуковник, 1997. 928с.

References

1. Beloshapkova VA Modern Russian language. M.: Azbukovnyk, 1997. 928 p.

RESUME

G.V. Dorokhina

The Algorithm of Syntactic Analysis Based on Grammatical Rules

The automatic syntactic analysis of Russian texts is usually solved using the statistical methods. But the solving of particular tasks of text analysis only by statistical methods says few new for the fundamental linguistics. Nowadays tools for automatic text analysis that based on linguistic methods are not developed enough. This fact points to the importance of the research. The object of research is a simple sentence of Russian language that is expanded but not semi-composite. The subject of research is the method of building a dependency tree by the sentence. The goal of research is to create the syntactic analysis algorithm based on grammatical rules.

The article propose the The proposed algorithm of Sentence Processing consists of following stages.

- 1. The morphological analysis of wordfroms.
- 2. Searching for pairs of potentially connected wordforms in the sentence.
- 3. Reduction a quantity of pairs of potentially connected wordforms.
- 4. Building of dependency tree.

We search for pairs of potentially connected wordforms in the sentence using grammatical rules. All types of syntactic connections are expressed as using connections between pairs of words. The paper contains the description of rules that allow defining if two words are potentially connected. The connections between words forming predictive base of sentence are considered in detail. Such connections are called the main connections. The rest of connections we call the minor connections.

Each wordform in Russian may have some interpretation on morphological level due to the phenomenon of homonymy. The way to reduce a quantity of pairs of potentially connected wordforms is proposed. It allows reducing the computational complexity of the algorithm.

We build the dependency tree as following. For the first we choose the sets of main connections needed to form the predictive base of sentence. Then for each of this set we build dependency trees by adding the minor connections.

Статья поступила в редакцию 05.06.2014.