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:

$$S = (s_1,... s_i,... s_N) . \quad (1)$$

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

$$s_i = \{s_1^i,...s_j^i,...s_{n_i}^i\} . \quad (2)$$

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

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

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:

$$T = Tm \cup Ta , \; Tm \cap Ta = \varnothing .$$

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 ($N_{2\ldots\text{pr}}$); the adjective and passive particles forms of nominative and instrumental cases, short form and comparative degree of the adjective ($Adj_{1/5/\text{f}}$); 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	$N_1\ Cop_f\ Adj_{1/5/f}$	Ночь тихая (тиха). Ночь была тихая (тихой, тиха). Ночь была тише дня.
3	$N_1\ Cop_f\ N_{1/5}$	Он – студент. Он был студент. Он был студентом.
4	$N_1\ Cop_f\ N_{2\ldots\text{pr}}/Adv_{\text{pr}}$	Дом был с лифтом. Чай – с сахаром. Глаза – навыкате. Глаза были навыкате. Подарок – сыну. Подарок был сыну.
5	$Inf\ V_f$	Курить воспрещалось.
6	$Inf\ Cop_f\ N_{1/5}$	Дозвониться – проблема (было проблемой). Любить иных – тяжелый крест.
7	$Inf\ Cop_f\ Adj_{1/5/f}$	Промолчать – разумно. Промолчать – самое разумное. Промолчать было разумно. Промолчать было самым разумным.
8	$Inf\ \ Cop_f\ N_{2\ldots\text{pr}}/Adv_{\text{pr}}$	Молчать было в его правилах. Молчать – в его 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	$Cop_{pl}\ N_{2\ldots\text{pr}}/Adv_{\text{pr}}$	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	$Cop_{s3}\ \ N_{2\ldots\text{pr}}/Adv_{\text{pr}}$	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 №16.

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

MSS	MSS template	Example
1	_K1	Грачи прилетели.
2	_K2	Ночь тихая (тиха).
	_KNC_L + _KCAdj	Ночь была тихая (тихой, тиха). Ночь тише дня.
3	_K3	Он – студент.
	_KNC_L + _KCN	Он был студент.
	_KNC_L + _K3_6	Он был студентом.
4	_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	Курить воспрещалось.
6	_KCI_Nom + _K3_6	Дозвонитья было проблемой.
	_KCI_Nom + _KCN	Дозвониться была проблема.
	_K6	Дозвониться – проблема.
7	_KCI_Nom + _KCAdj	Промолчать было разумно.
	_K7	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 + _K_Pr_Nobj	Молчать было в его правилах.
	_KCI_Nom + _KC_Pr	Молчать было некстати.
9	_KCI_Nom + _KCI	Отказаться было обидеть.
	_K9	Отказаться – обидеть.
12	_KCP + _K_Pr_Nobj	От него были в восторге.
	_KC_Pr	С ним были запросто.
13	_KCN	Будет дождь.
14	_K14	Было темно.
15	_K15	Результатом были довольны.
17	_K_Pr_Nobj	Без осадков.
	_KCP + _K_Pr_Nobj	Будет без осадков.
	_KC_Pr	Было поздно.
	_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.

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

Notation of rule	Part of speech of principal (x) and dependent (y) words		Dash	Sequence order of words
_K5	x	Verb (infinitive)		any
	y	Verb (non infinitive)		
_K9	x	Verb (infinitive)	+	any
	y	Verb (infinitive)		
_KI_Pred	x	Verb (infinitive)		direct
	y	Adverb		
_KI_Prep	x	Verb (infinitive)	+	direct
	y	Preposition		
_KC_Pr	x	Copula		any
	y	Adverb		
_KN_Pred		Noun	+	any
		Adverb		
_KCI	x	Copula		direct
	y	Verb (infinitive)		
_KCI_Nom	x	Copula		indirect
	y	Verb (infinitive)		
_KCP	x	Copula		direct
	y	Preposition		
_K_Pr_Nobj *	x	Preposition		direct
	y	Noun		

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);
- Fofm 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’ ($\vee$), ‘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
_K1	x	Verb (non infinitive)	Any	$N_x=N_y$ &C_y= ‘nom.’& ((T_x=‘past’∨ F_x=3) & (N_x=‘sing.’&$G_x=G_y \vee N_x$= ‘pl.’)∨ T_x=‘fut.’∨ T_x=‘pres.’))
	y	Noun		
	x	Verb (non infinitive)	Any	$N_x=N_y$&C_y=‘nom.’&(T$_x$=‘past’∨ F_x=3) & (N_x=‘sing.’&$G_x=G_y \vee N_x$=‘pl’) ∨ $F_x=F_y$ & (T_x=‘fut.’∨T_x=‘pres.’))
	y	Personal pronoun		
	x	Verb (non infinitive)	Any	C_y= ‘nom.’
	y	Cardinal numeral		
_K7	x	Verb (infinitive)	Adverb	(C_y= ‘nom.’∨ $AdjF_y$=‘short’ ∨ $AdjF_y$=‘comp.’) & *Dash*
	y	Adjective		
_KI_NomObj	x	Verb (infinitive)	Any	C_y= ‘nom.’
	y	Noun∨Adjective		
_K_NomObj	x	Noun	!Verb	C_y= ‘nom.’C_y= ‘nom.’
	y	Noun		
_K14	x	Copula	!Verb	$N_x=N_y$ & N_x=‘sing.’ & F_x =3& $AdjF_y$=‘short’
	y	Adjective		
_K15	x	Copula	!Verb	$N_x=N_y$ & N_x=‘pl.’ & F_x =3 & $AdjF_y$=‘short’
	y	Adjective		
_KCAdj	x	Copula	!Verb	((C_y=‘nom.’∨C_y=‘in.’) &(T_x=‘past’& $G_x=G_y \vee T_x$=‘fut.’∨T_x=‘pres.’)) ∨ $AdjF_y$=‘short’ ∨ $AdjF_y$=‘comp.’
	y	Adjective		
_K3_6	x	Copula	Any	$N_x=N_y$ & C_y= ‘in.’
	y	Noun		
_KCN	x	Copula	!Verb	wOrder=‘direct’&$N_x=N_y$&C_y=‘nom.’& ((T_x=‘past’∨F_x=3)&(N_x=‘sing.’&$G_x=G_y$ ∨N_x= ‘pl’) ∨ T_x=‘fut.’∨ T_x=‘pres.’))
	y	Noun		
_KNC_L	x	Copula	!Verb	wOrder=‘indirect’&$N_x=N_y$&C_y=‘nom.’ &((T_x=‘past’∨F_x=3)&(N_x=‘sing.’& $G_x=G_y$∨N_x=‘pl’)∨T_x=‘fut.’∨T_x=‘pres.’))
	y	Noun		
_KC_Pr	x	Copula	!Verb	C_y=‘gen.’ ∨ C_y=‘dat.’ ∨ C_y=‘acc.’
	y	Noun		
_K2	x	Noun	!Verb	C_x=‘nom.’&($N_x=N_y$&$G_x=G_y$& (C_y=‘nom.’ ∨ $AdjF_y$=‘short’)∨$AdjF_y$=‘comp.’)
	y	Adjective		
_K3	x	Noun	!Verb	C_x=‘nom.’ & C_y=‘nom.’ & *Dash*
	y	Noun		
_KN1_Prep	x	Noun	any	C_x=‘nom.’ & *Dash*
	y	Preposition		
_K6	x	Verb (infinitive)	!Verb	C_y=‘nom.’ &*Dash*
	y	Noun		
_K17	x	Copula	!Verb	$C_y \neq$‘nom.’
	y	Noun		

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
t_{a1}	x	Verb	Any	Any
	y	Adverb		
t_{a2}	x	Verb	!Verb	Any
	y	Adverbial participle		
t_{a3}	x	Verb	!Verb	Any
	y	Verb (infinive)		
t_{a4}	x	Verb	!Verb	Any
	y	Preposition		
t_{a5}	x	Noun	Adjective ∨ Participle	Any
	y	Preposition		
t_{a6}	x	Noun	Adjective ∨ Participle	Direct
	y	Verb (infinive)		
t_{a7}	x	Adjective	None	Any
	y	Adverb		
t_{a8}	x	Adjective	None	Direct
	y	Preposition		
t_{a9}	x	Adjective	Adverb	Direct
	y	Verb (infinive)		
t_{a10}*	x	Preposition	Adjective ∨ Participle	Direct
	y	Noun		
t_{a11}	x	Adverb	None	Direct
	y	Adverb		
t_{a12}	x	Adverb	None	Direct
	y	Preposition		

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}	x	Verb	Any	Any	$C_y \neq$ ‘nom.’
	y	Noun			
t_{a14}	x	Noun	Adverb	Any	$(C_x=C_y \vee AdjF_y=$‘comp.’)& $N_1=N_2$&($N_1=$‘pl.’$\vee G_1=G_2$)
	y	Adjective			
t_{a15}	x	Noun	Adjective ∨ Participle	Direct	$C_y \neq$ ‘nom.’
	y	Noun			
t_{a16}	x	Adjective	Adjective ∨ Participle	Direct	$C_y=$‘gen.’ $\vee$ $C_y=$‘dat.’ $\vee$ $C_y=$‘in.’
	y	Noun			

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):

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

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).

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

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."

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

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.

$$S' = (s'_1, \ldots s'_i, \ldots s'_N), \quad s'_i \subseteq s_i : \forall z \in s'_i \ z \in (A \cup B). \tag{7}$$

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

$$D = s'_1 \times \ldots \times s'_i \times \ldots \times s'_N, \quad D = \{d_k : d_k = (d_1^k, \ldots, d_i^k, \ldots d_N^k)\} \tag{8}$$

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.

$$\begin{aligned} F_k &= \{d_i^k\}, i = \overline{1,N} \\ R_k &= \{(x,y,t) : (x,y,t) \in R, x \in F_k, y \in F_k\} \\ A_k &= \{x : (x,y,t) \in R_k\} \\ B_k &= \{y : (x,y,t) \in R_k\} \end{aligned} \quad (9)$$

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

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

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."

$$| A_k \setminus B_k | \leq 1 \quad (11)$$

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."

$$\forall z \in P \cap F_k \; z \in B_k \cap A_k \quad (12)$$

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

$$Rm_i = \{(x,y,t) : (x,y,t) \in R_k, t = t_i, t_i \in h\} \quad (12)$$

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

Let we put $RM=\{rm_V\}$, where $\subseteq Rm1\times...\times Rmi\times...\times Rml$ and $rm_V=((x1,y1,t1), ...,(xl,yl,tl))$: if $l>1$ $x1=x2, \forall i>1$ $xi+1=yl$.

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.

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

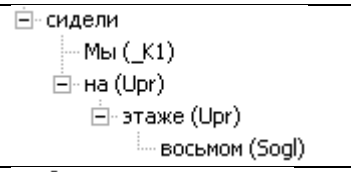
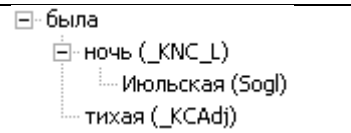
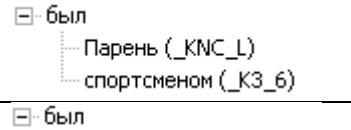
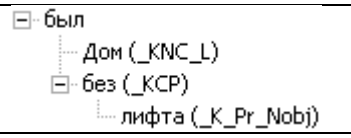
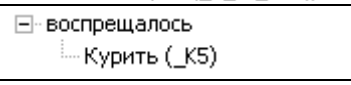
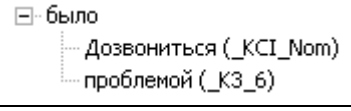
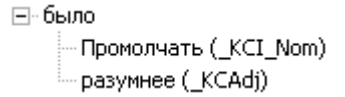
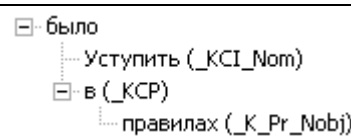
Let's mark $g' = g \cup c$. If the digraph (g', R_k) is not single-connected, so it's impossible to create the correct TSS. Otherwise it's necessary to solve a peaks problem with an in-degree more than 1. For each such peak only one connection is left, according to the requirement, that the way length 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
	x	*y*	*t*	
Мы сидели на восьмом этаже	сидели	Мы	_K1	
	сидели	на	Upr	
	на	этаже	Upr	
	этаже	восьмом	Sogl	
Июльская ночь была тихая	ночь	Июльская	Sogl	
	была	ночь	_KNC_L	
	была	тихая	_KCAdj	
Парень был спортсменом	был	Парень	_KNC_L	
	был	спортсменом	_K3_6	
Дом будет без лифта	был	Дом	_KNC_L	
	был	без	_KCP	
	без	лифта	_K_Pr_Nobj	
Курить воспрещалось	воспрещалось	Курить	_K5	
Дозвониться было проблемой	Дозвониться	проблемой	_KCI_Nom	
	было	Дозвониться	_KC_Pr	
Промолчать было разумнее	было	Промолчать	_KCI_Nom	
	было	разумнее	_KCAdj	
Уступить было в правилах	было	Уступить	_KCI_Nom	
	было	в	_KCP	
	в	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.