

ДИСКРЕТНЫЕ МОДЕЛИ РЕАЛЬНЫХ ПРОЦЕССОВ

УДК 519.17

ФОРМАЛЬНАЯ ГРАММАТИКА РУССКОГО ЯЗЫКА В ПРИМЕРАХ

Я. М. Мирзабеков, Ш. Б. Шихиев

Дагестанский государственный университет, г. Махачкала, Россия

На примере одного предложения показана возможность построения конструктивной теории естественного языка и реализации морфологии и синтаксиса на компьютере путём формализации соответствующих правил. В лексическом составе рассматриваемого примера несколько слов. Грамматика русского языка упрощена. Модель морфологии определяется лексическим наполнением предложений и правилами морфологии, которые хранятся в словообразовательном (морфологическом) словаре. В компьютерной модели морфологии реализованы два алгоритма: синтез и анализ словоформ. Модель синтаксиса языка определяется его морфологией и правилами его синтаксиса. Правила синтаксиса порождают словосочетания, используемые в языке. Правила синтаксиса представлены в виде пар (прямых производных слов) разных лексических групп. Лексическая группа есть множество словоформ, имеющих одинаковые морфологические параметры. Словосочетания «оставляют» на лексическом наполнении предложений некоторый граф, который и представляет собой синтаксис. В терминах теории графов определяются понятия синтаксиса, в частности понятие предложения. Формулируются две задачи: синтез и анализ предложения; описаны и реализованы алгоритмы их решения в среде программирования Delphi 7.0.

Ключевые слова: *естественный язык, морфология, синтаксис, словоформа, правила морфологии, правила синтаксиса, словосочетание, реализация алгоритма на компьютере, компьютерная модель, граф, дерево, анализ предложения.*

DOI 10.17223/20710410/40/10

FORMAL GRAMMAR OF RUSSIAN LANGUAGE IN EXAMPLES

Ya. M. Mirzabekov, Sh. B. Shihiev

*Dagestan State University, Makhachkala, Russia***E-mail:** yash831@mail.ru

The possibility of a constructive theory creation for a natural language and of its morphology and syntax implementation on a computer by formalizing the appropriate rules is shown in the article. There are a few words in the lexicon of the example under consideration. The grammar of the Russian language is simplified. The morphology model is determined by the vocabulary and the rules of morphology which are stored in the word-morphological (morphological) dictionary. In the computer model of morphology, two algorithms are implemented: synthesis and analysis of word forms. The syntax model is defined by the morphology and syntax rules. Syntax rules generate word combinations used in the language. The rules of syntax are presented in

the form of pairs (direct products) of lexical groups. A lexical group is a set of word forms that have the same morphological parameters. Phrases are defined in the lexicon of some graph which is a syntax. The concepts of syntax are defined in terms of graph theory. In particular, the concept of the sentence in a syntax is defined. Then, two problems are formulated: synthesis and analysis of the sentence. Algorithms for solution of them are described and implemented in the Delphi 7.0 programming environment.

Keywords: *natural language, morphology, syntax, word form, morphology rules, syntax rules, the combinations of two words, the implementation of the algorithm on the computer, the computer model, graph, tree, analysis of the sentence.*

Формулировка задачи

Если сочетание слов «компьютерная лингвистика» обозначает раздел лингвистики, предмет, методы и задачи которого представляются и реализуются на компьютере, то, разумеется, надо иметь описание естественного языкового явления в логико-технических терминах вычислительной машины. Такое описание принято называть компьютерной моделью языка или программой, реализующей на компьютере алгоритмы решения задач лингвистики. Из известных формулировок задачи моделирования естественного языка наиболее корректной является следующая.

«Основным орудием исследования и описания» естественного языка, как утверждает И. А. Мельчук [1], является модель языка как «сложно организованной совокупности правил, чисто механическое применение которых должно» реализовать отображение (желательно, биекцию) двух множеств (субстанций, по Ф. де Соссюру): «планов содержания» и «планов выражения» (по Л. Ельмслеву) или «смыслов» и «текстов» (по Мельчуку). Но ни одна такая модель, продолжает И. А. Мельчук, «хотя бы в виде чисто экспериментального варианта» ему не известна. «Можно представить себе два пути упрощения этой задачи . . . ограничившись её достаточно точно разработанными фрагментами. . . Второй путь — чтобы как-то были намечены обязательно все основные компоненты модели (пусть даже приблизительно) . . . и необходимые связи между ними» [1].

Однако когда речь идёт о компьютерной модели такого сложного и ещё не изученного явления, как язык, общими и приблизительными описаниями задача не решается. В данной работе делается попытка показать один из возможных приёмов моделирования языка: построены синтаксические примитивы и продемонстрированы их возможности на примере.

Имеет смысл вспомнить и тот факт, что в начале прошлого века известный исследователь языка Л. Витгенштейн для демонстрации различных функций и возможностей языка рассматривал языки с небольшим числом слов. Такие примеры слов он назвал языковыми играми. Содержание данной работы можно называть и языковой игрой, подчеркивая значительные перспективы моделирования отдельных фрагментов естественного языка.

1. Морфология

Рассмотрим пример грамматики, лексика которой включает следующие слова и их формы:

- 1) **пол, дом, угол, топор, нож** (обозначение — ИС, код — 01);
- 2) **серый, белый, легкий, быстрый** (обозначение — ИП, код — 02);
- 3) **находиться, лежать, спать, строить** (обозначение — Гл, код — 07);

4) **у/2, с/2, с/5** (код предлога — 08).

В каждом пункте напротив слов указано сокращённое обозначение части речи и код. Номер падежа, которым он управляет, указан за предлогом.

В примере морфология включает две категории:

числа — код: 20 = (21, 22);
падежа — код: 30 = (31, 32, 33, 34, 35, 36).

Рассмотрим правила морфологии для этих слов. Категории числа и падежа являются словоизменительными для ИС и ИП. Следовательно, каждое из имён имеет по 12 беспредложных форм (ИП имеет только форму мужского рода). Каждое ИС имеет ещё шесть предложных форм, образованных в сочетании с тремя предлогами в единственном и множественном числах, например: **у дома, с домами, с пола** и т. д.

Глагол имеет форму настоящего времени третьего лица в единственном и множественном числах, например: **лежит, лежат** и т. д.

Чтобы иметь исчерпывающие сведения о морфологии для лексического набора из 16 слов и двух категорий, нужны описания морфологических правил, по которым образуются формы слов. Эти правила известны из грамматики русского языка, они же описаны в грамматических словарях, например в словаре А. А. Зализняка [2]. Там слово **пол** имеет следующий формат: «**пол м 1с**». Буква «м» означает «ИС мужского рода», а «1с» есть код морфологического правила, по которому образуются формы слова **пол**. Морфологические правила и их кодировки описаны в начале словаря.

В данной работе используется иная кодировка морфологических правил. ИС **пол**, например, имеет следующий формат:

пол01:2130:0,а,у,ом,е,:2230:0ы,ов,ам,ы,ами,ах.

Данная конструкция (словарная статья) состоит из трёх частей, разделённых двоеточиями. В первой части — слово (**пол**) и код его лексической группы (**01**). Далее идут два морфологических правила.

Морфологические правила имеют унифицированный формат: состоят из пятизначного числа и последовательности постфиксов (окончаний). Рассмотрим подробно формат и содержание правила «**2130:0,а,у,ом,е,**», в котором упакованы следующие четыре параметра данного правила:

- 1) 21 — формы единственного числа;
- 2) 30 — все падежные формы;
- 3) 0 — число букв, отбрасываемых справа от данного слова, чтобы получить его основу;
- 4) «а,у,ом,е,» — окончания, разделённые запятыми, приписываются к основе слова для получения форм данного слова.

Аналогично читается и второе правило: «**2230:0 ы,ов,ам,ы,ами,ах,**». Окончания слов имеет смысл хранить в отдельном списке, а в словаре вместо окончаний указывать их порядковый номер в списке.

Унифицированный формат словарной статьи даёт возможность описать процедуру `MorfForms(s: string; var forms: smas12)`, которая по словарной статье строит все двенадцать форм любого из девяти имён. ИП имеет такой же формат, как ИС. Например:

белый 02:2130:2ый,ого,ому,ый,ым,ом,:2230:2ые,ых,ым,ые,ыми,ых.

Словарь (словообразовательный, грамматический или морфологический) из 16 слов приводится ниже:

- 1) пол 01:2130:0,а,у,,ом,е,:2230:0ы,ов,ам,ы,ами,ах;
- 2) дом 01:2130:0,а,у,,ом,е,:2230:0а,ов,ам,а,ами,ах;
- 3) угол 01:2130:0,а,у,,ом,е,:2230:2лы,лов,лам,лы,лами,лах;
- 4) топор 01:2130:0,а,у,,ом,е,:2230:0ы,ов,ам,а,ами,ах;
- 5) нож 01:2130:0,а,у,,ом,е,:2230:0и,ей,ам,и,ами,ах;
- 6) серый 02:2130:2ый,ого,ому,ый,ым,ом,:2230:2ые,ых,ым,ые,ыми,ых;
- 7) белый 02:2130:2ый,ого,ому,ый,ым,ом,:2230:2ые,ых,ым,ые,ыми,ых;
- 8) легкий 02:2130:2ий,ого,ому,ий,им,ом,:2230:2ие,их,им,ие,ими,их;
- 9) быстрый 02:2130:2ый,ого,ому,ый,ым,ом,:2230:2ые,ых,ым,ые,ыми,ых;
- 10) находиться 07: 21-находится 22-находятся;
- 11) лежать 07: 21-лежит 22-лежат;
- 12) спать 07: 21-спит 22-спят;
- 13) строить 07: 21-строит 22-строят;
- 14) у08:2;
- 15) с08:2;
- 16) с08:5.

Извлечение форм глагола и предлога из словарной статьи не представляет труда. Процедура MorfForms используется для синтеза форм всех слов лексики.

Анализ или распознавание словоформы является еще одной задачей морфологии. Она заключается в определении исходной формы слова и всех её морфологических параметров. Например, результатом анализа словоформ

полом, у пола, белым, белыми, лежит, у

являются следующие строки (исходные формы слов и их параметры):

«пол 01 21 35», «пол 01 21 32 14», «белый 02 22 35»,
«белым 02 21 35; 02 22 33», «лежит 07 21», «у08:2».

Привычные термины, например «ИП, множественное число, винительный падеж», записаны кодами «02 22 35». Четвёртый параметр в форме ИС указывает на порядковый номер предлога в словаре, более того, наличие этого параметра в форме ИС является признаком предложной формы имени. Словоформа **белым** имеет два набора морфологических параметров.

Процедура анализа словоформы — довольно трудоёмкий процесс из-за того, что словарная статья начинается с исходной формы слова. Например, чтобы обнаружить в словаре словоформу **белым**, потребуется построить все двенадцать форм слова **белый** и среди них найти данное слово. Сложность поиска (в общем случае) заключается в отсутствии информации о том, что слово **белым** окажется формой слова **белый**, а не **бельмо**, **белить** и т. д. Поэтому требуется искать слово по всему словарю: распаковать каждую словарную статью с помощью процедуры MorfForms и сравнить искомое слово со всеми элементами массива forms.

Словарь, содержащий в себе информацию для образования всех словоформ лексики, будем называть математической моделью морфологии (МММ).

2. Лексические группы

Лексическая группа — это множество, состоящее из словоформ с одинаковыми морфологическими параметрами. В синтаксисе, который строится в этом примере, участвуют следующие лексические группы.

Через ИС31 обозначим множество ИС в именительном падеже, оно образует лексическую группу 012031 и представляет собой объединение $012031 = 012131 \cup 012231$, где

- 1) $012131 = \{\text{пол, дом, угол, топор, нож}\};$
- 2) $012231 = \{\text{полы, дома, углы, топоры, ножи}\}.$

Через ИС/рг обозначено множество ИС в предложной форме. Каждое ИС имеет по шесть предложных форм. Например: **у дома** и **у домов**, **с дома** и **с домов**, **с домом** и **с домами**. То есть предлог ведёт себя как элемент падежной информации. ИС/рг из 30 словоформ разбито на шесть лексических групп:

- 1) $01213201 = \{\text{у пола, у дома, у угла, у топора, у ножа}\};$
- 2) $01213202 = \{\text{с пола, с дома, с угла, с топора, с ножа}\};$
- 3) $01213503 = \{\text{с полом, с домом, с углом, с топором, с ножом}\};$
- 4) $01223201 = \{\text{у полов, у домов, у углов, у топоров, у ножа}\};$
- 5) $01223202 = \{\text{с полов, с домов, с углов, с топоров, с ножа}\};$
- 6) $01223503 = \{\text{с полами, с домами, с углами, с топорами, с ножами}\}.$

Имена прилагательные имеют только полные формы и в словосочетаниях рассматриваемого синтаксиса участвуют только ИП из группы $\text{ИП125} = 022131 \cup 022132 \cup 022135 \cup 022231 \cup 022232$, где

- 1) $022131 = \{\text{серый, белый, легкий, быстрый}\};$
- 2) $022132 = \{\text{серого, белого, легкого, быстрого}\};$
- 3) $022135 = \{\text{серым, белым, легким, быстрым}\};$
- 4) $022231 = \{\text{серые, белые, легкие, быстрые}\};$
- 5) $022232 = \{\text{серых, белых, легких, быстрых}\};$
- 6) $022135 = \{\text{серыми, белыми, легкими, быстрыми}\}.$

Лексическая группа Гл12, или $0720 = 0721 \cup 0722$, где

- 1) $0721 = \{\text{находится, лежит, спит, строит}\};$
- 2) $0722 = \{\text{находятся, лежат, спят, строят}\}.$

Объединение лексических групп ИС31, ИС/рг, ИП125 и Гл12 обозначим через $LG1$. В нём всего $10 + 30 + 24 + 8 = 72$ словоформы.

Далее под лексической группой, если не уточнено её определение, подразумевается любая из перечисленных шестнадцати групп.

3. Синтаксис

Одним из основных понятий синтаксиса является понятие словосочетания. Формально, словосочетание есть пара словоформ. Но не каждая пара словоформ образует словосочетание. Синтаксис языка устроен так, что если пара словоформ w_1 и w_2 из лексических групп L_1 и L_2 соответственно образует словосочетание, то любая пара словоформ из этих групп образует словосочетание. В таких случаях будем говорить, что пара лексических групп L_1 и L_2 образует словосочетание; множество словосочетаний, образованных группами L_1 и L_2 , обозначается $L_1 \times L_2$.

Словосочетания — неделимые и базовые единицы синтаксиса — задаются как элементы прямого произведения различных лексических групп. Описание и именование лексических групп необходимы для определения синтаксиса. Например: $\{\text{дом, пол}\} \times \{\text{белый, крепкий}\} = \{(\text{дом, белый}), (\text{дом, крепкий}), (\text{пол, белый}), (\text{пол, крепкий})\}$. Переписывая пары словоформ в виде «дом белый», «дом крепкий», «пол белый», «пол крепкий» или, как принято в синтаксисе русского языка, «бе-

лый дом», «крепкий дом», «белый пол», «крепкий пол», получим словосочетания.

В таблице перечислены 34 синтаксических отношения (СО) (или 34 произведения лексических групп), порождающих словосочетания русского языка из словоформ $LG1$.

СО, порождающие словосочетания

№	СО	Комментарии
1	012131×0721	Подлежащее и сказуемое
2	012231×0722	
3	012131×022131	Согласованное определение
4	012231×022231	
5	01213201×022132	
6	01213202×022132	
7	01213503×022135	
8	01223201×022232	
9	01223202×022232	
10	01223503×022235	
11	012131×01213201	Отношение беспредложного ИС с предложным ИС
12	012131×01213202	
13	012131×01213503	
14	012131×01223201	
15	012131×01223202	
16	012131×01223503	
17	012231×01213201	
18	012231×01213202	Отношение беспредложного ИС с предложным ИС
19	012231×01213503	
20	012231×01223201	
21	012231×01223202	
22	012231×01223503	
23	0721×01213201	Отношение глагола с предложным ИС
24	0721×01213202	
25	0721×01213503	
26	0721×01223201	
27	0721×01223202	
28	0721×01223503	
29	0722×01213201	
30	0722×01213202	
31	0722×01213503	
32	0722×01223201	
33	0722×01223202	
34	0722×01223503	

Обозначим через $R1$ множество словосочетаний, содержащихся в СО из таблицы. В этих словосочетаниях присутствуют только 72 словоформы из $LG1$. Мы имеем граф $G = (LG1, R1)$, который называется синтаксисом нижнего уровня. Параллельно с G имеется граф $CodeG = (CodeLG1, CodeR1)$, который называется синтаксисом верхнего уровня или просто синтаксисом. Через $CodeR1$ обозначено множество из СО, перечисленных в таблице. Все применённые без определения понятия из теории графов можно найти в [3].

Графы G и $CodeG$ оправдывают свои названия в том смысле, что все понятия синтаксиса будут определены в терминах этих графов, в частности понятие предложения. Граф $CodeG$ как математическая структура представляет собой математическую

модель синтаксиса (ММС). Лексика $LG1$, МММ и ММС образуют математическую модель грамматики.

4. Предложение

Предложением в синтаксисе G считается любое корневое дерево с корнем из ИС1, если в нём присутствует дуга из $012131 \times 0721 \cup 012231 \times 0722$ [4]. Иначе говоря, в предложении должны быть подлежащее (из ИС1) и согласованное с ним по тем или иным морфологическим характеристикам сказуемое (из Гл1).

Корневое дерево в $CodeG$ с корнем в 012131 или 012231 называется синтаксической формой (СФ), если в нём присутствует дуга 012131×0721 или 012231×0722 .

Между предложениями и СФ имеют место следующие отношения.

Если в СФ ($CodeV$, $CodeA$) код каждой лексической группы заменить некоторым элементом этой группы, то получится некоторое предложение (V, A) . В этом случае будем говорить, что предложение (V, A) порождено СФ ($CodeV$, $CodeA$).

Два корневых дерева $RT1 = (V1, A1)$ и $RT2 = (V2, A2)$ из G имеют одинаковую синтаксическую форму, если они изоморфны и в изоморфном отображении $f: V1 \rightarrow V2$ вершины $v1$ из $V1$ и $f(v1)$ из $V2$ принадлежат одной и той же лексической группе. Если два предложения имеют одинаковые синтаксические формы, то они порождены одной и той же СФ. Все предложения, порождённые данной СФ, имеют одинаковые синтаксические формы. Таким образом, каждая СФ представляет собой правило, порождающее предложения определённого класса.

Размер графа G зависит от объёма лексики, размер графа $CodeG$ никак не зависит от этого. Граф $CodeG$ на самом деле есть редукция графа G . Вершинам графа $CodeG$ соответствуют лексические группы, а вершинам графа G — словоформы из этих лексических групп.

Граф $CodeG$ показан на рис. 1. В $CodeG$ клонированы (дублированы) вершины ИП125\16-1 (A1) и ИС2\30-1 (S1) для упразднения циклов. При редукции графа G в граф $CodeG$ были стянуты в одну дугу (графа $CodeG$) несколько параллельных дуг (графа G); их число указано у каждой дуги графа $CodeG$. В верхнем правом углу записаны условные обозначения вершин, в каждой вершине указано имя соответствующего ей множества и коды групп, входящих в это множество. После имени множества через «\» указано число элементов в нём. Структуры множеств ИП125\16-1 и ИП125\16-2, ИС2\30-1 и ИС2\30-2 идентичны.

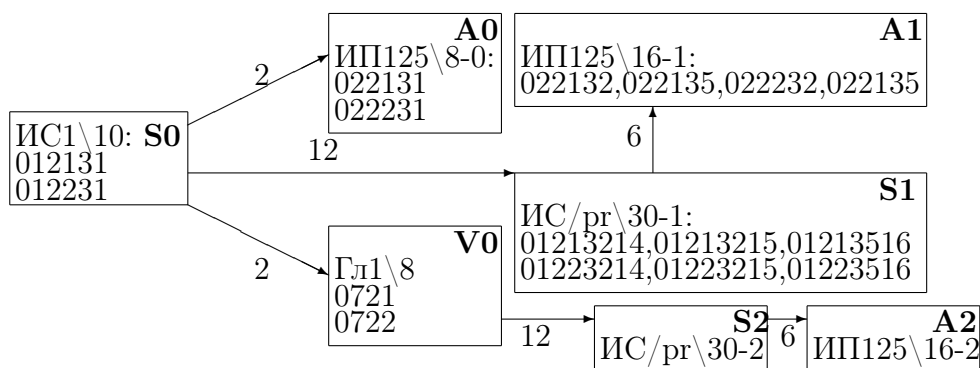


Рис. 1. Граф $CodeG$ без циклов

Нетрудно посчитать, что в графе на рис. 1 имеется 18 СФ. Выпишем их в скобочной форме, используя обозначения вершин, после записи дерева указано число СФ, содержащихся в этом дереве:

- 1) $S_0(V_0)$ содержит 2 СФ;
- 2) $S_0(S_1, V_0)$ содержит 12 СФ;
- 3) $S_0(S_1(A_1), V_0)$ содержит 12 СФ;
- 4) $S_0(A_0, V_0)$ содержит 2 СФ;
- 5) $S_0(A_0, S_1, V_0)$ содержит 12 СФ;
- 6) $S_0(A_0, S_1(A_1), V_0)$ содержит 12 СФ;
- 7) $S_0(V_0(S_2))$ содержит 12 СФ;
- 8) $S_0(S_1, V_0(S_2))$ содержит 12 СФ;
- 9) $S_0(S_1(A_1), V_0(S_2))$ содержит 72 СФ;
- 10) $S_0(A_0, V_0(S_2))$ содержит 12 СФ;
- 11) $S_0(A_0, S_1, V_0(S_2))$ содержит 72 СФ;
- 12) $S_0(A_0, S_1(A_1), V_0(S_2))$ содержит 72 СФ;
- 13) $S_0(V_0(S_2(A_2)))$ содержит 12 СФ;
- 14) $S_0(S_1, V_0(S_2(A_2)))$ содержит 12 СФ;
- 15) $S_0(S_1(A_1), V_0(S_2(A_2)))$ содержит 72 СФ;
- 16) $S_0(A_0, V_0(S_2(A_2)))$ содержит 12 СФ;
- 17) $S_0(A_0, S_1, V_0(S_2(A_2)))$ содержит 72 СФ;
- 18) $S_0(A_0, S_1(A_1), V_0(S_2(A_2)))$ содержит 72 СФ.

Получится внушительное число: 556 СФ в небольшом графе $(LG1, R1)$. Последнему дереву с номером 18 соответствует сам граф *CodeG*. Выпишем одну из 72 СФ, соответствующих этому дереву:

$$012131(022131, 01213516(022135), 0721(01213214(022132)). \quad (1)$$

Каждый из семи кодов групп, входящих в (1), может быть заменён любой словоформой из этой группы. Следовательно, СФ (1) порождает $5 \cdot 4 \cdot 5 \cdot 4 \cdot 4 \cdot 5 \cdot 4 = 32000$ предложений.

Замечание 1. Если к лексике *LG1* добавить тысячи новых имён и глаголов, размер графа *CodeG* не изменится.

Тем не менее 556 — большое число, когда речь идет о правилах синтаксиса, порождающих предложения. Однако видно, что все 18 деревьев — поддеревья дерева 18, а все 556 правил — производные от 72 правил. Поэтому можно считать, что правило в рассматриваемом синтаксисе одно: задано оно деревом 18 и допустимым набором лексических групп, прикреплённых к его вершинам. Это важное для синтаксиса утверждение демонстрируется ниже на одном примере синтеза и анализа предложения.

Представим дерево (1) в привычном виде (рис. 2) и назовём его «Правило 1». В вершинах графа за кодом лексической группы указана переменная, определённая в этой группе. У каждой дуги дерева указано соответствующее ей СО:

$$\mathbf{A} = 01112131 \times 02112131,$$

$$\mathbf{B} = 01112131 \times 02112131,$$

$$\mathbf{D} = 01112131 \times 02112131,$$

$$\mathbf{C} = 01112131 \times 02112131,$$

$$\mathbf{F} = 01112131 \times 02112131,$$

$$\mathbf{G} = 01112131 \times 02112131.$$

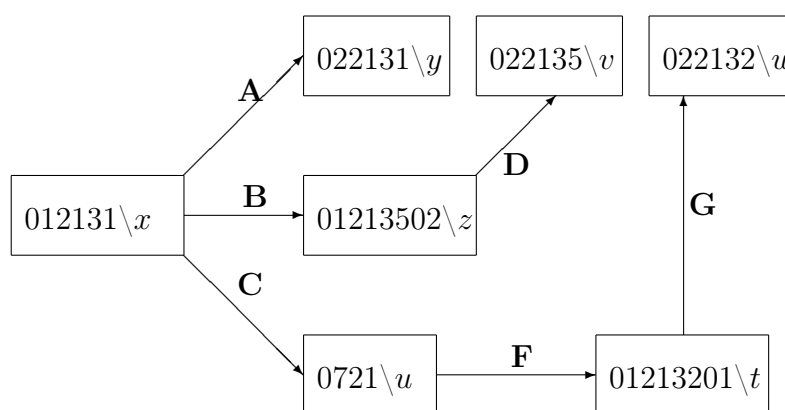


Рис. 2. Пример СФ. Правило 1

Дерево можно записать через его дуги: $0 (A, B(D), C(F(G)))$. Сделано допущение, что корень дерева является концом фиктивной дуги «0».

Присваивая переменным x, y, z, u, v, w, t из Правила 1 значения из соответствующих групп, можно получить различные предложения. Этот процесс называется синтезом предложений. На рис. 3 показан пример предложения, синтезированного по Правилу 1.

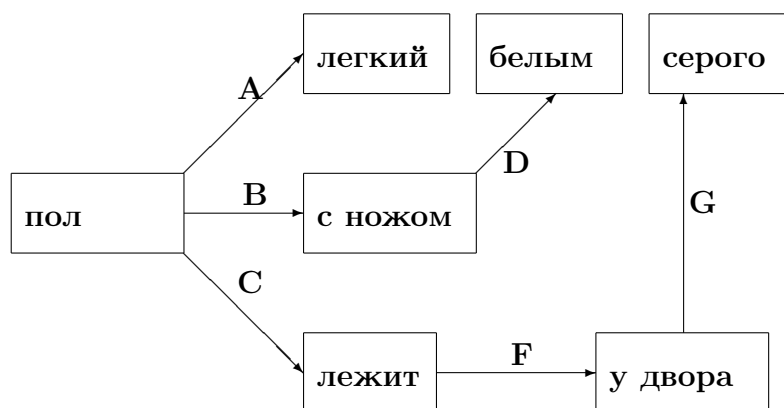


Рис. 3. Пример предложения, синтезированного по Правилу 1

Предложение, соответствующее рис. 3, в скобочной форме выглядит так:

пол (легкий, с ножом(белым), лежал (на дворе (сером))). (2)

Предложение-дерево (2) трансформируется в предложение-последовательность

легкий пол с белым ножом лежал на сером дворе

по определённому правилу: односторонним обходом предложения-дерева и перестановкой местами ИС и ИП.

Задача анализа обратна задаче синтеза: для заданного предложения требуется найти порождающее его правило. Обобщением задачи анализа является задача построения предложения из элементов заданной последовательности словоформ.

Эти две задачи отличаются по той причине, что между фрагментами предложения-последовательности (отрезком из соседних словоформ в последовательности) и сегментами предложения-дерева (членами предложения, принадлежащих одной ветви предложения-дерева) имеют место известные в грамматике закономерности. Например, сегмент сказуемого (члены предложения, принадлежащие ветви, корнем которого является сказуемое) и сегмент подлежащего делят предложение на два непересекающихся фрагмента.

Рассмотрим следующий пример задачи анализа предложения. Дана строка WFs , состоящая из словоформ, разделенных пробелами:

легкий пол с белым ножом лежал на сером дворе.

Требуется выяснить: образуют ли предложение словоформы из WFs ? Решается эта задача следующим образом.

Шаг 1. Из строки WFs выделяются все его словоформы и подвергаются морфологическому анализу. Для нашего примера имеем:

- 1) **легкий** — легкий 022131, 022134;
- 2) **пол** — пол 012131, 012134;
- 3) **белым** — белым 022135, 022233;
- 4) **с** — с 0832, 0835;
- 5) **ножом** — ножом 011135;
- 6) **лежит** — лежит 0721;
- 7) **серого** — сером 022132;
- 8) **у** — у 0832;
- 9) **двора** — дворе 012132.

Шаг 2. Предлоги прикрепляются к соответствующим ИС, и таким образом образуются предложные формы ИС. В нашем примере они таковы:

- 1) **легкий** — легкий 022131, 022134;
- 2) **пол** — пол 012131, 012134;
- 3) **белым** — белым 022135, 022233;
- 4) **с ножом** — ножом 01113502;
- 5) **лежит** — лежит 0721;
- 6) **серого** — сером 022132;
- 7) **у двора** — дворе 01213201.

Дальнейшие действия зависят от решаемой задачи. Для построения предложения из заданных словоформ строится граф на множестве из семи лексических групп, к которым принадлежат данные слова. Если слово принадлежит нескольким группам (как в случае слов **легкий**, **пол** и **белым**), то следует рассмотреть всевозможные варианты. Строится граф следующим образом. Если $WFsCode$ — множество кодов групп, которым принадлежат данные слова, а $Codes$ — множество кодов из таблицы (с. 119), то множество дуг графа будет $(WFsCode \times WFsCode) \cap Codes$. Любая СФ (корневое дерево) в этом графе будет структурой предложения из заданных слов.

Рассматриваемый пример задачи анализа предложения проще: образуют ли предложение словоформы из WFs (в том порядке, в котором они находятся в WFs)? Это значит, что есть возможность воспользоваться следующими правилами расположения членов предложения в Правиле 1:

- 1) предлог прикреплён к первому ИС, которое находится после него;
- 2) ИП связано с первым ИС, которое находится после него;
- 3) подлежащее предшествует сказуемому;
- 4) группа (сегмент) сказуемого находится за группой (сегментом) подлежащего;
- 5) синтаксической формой предложения может быть одна из рассмотренных 556 СФ.

Первые четыре требования обязательны для предложений русского языка. Пятое требование реализовано в алгоритме прикрепления словоформ к вершинам дерева СФ (структура заранее известна): если в заданной последовательности имеется слово, соответствующее данной вершине, то она включается в искомое дерево СФ, иначе её не будет в СФ.

В ходе анализа словоформ из примера уточняются их группы:

- 1) **легкий** — **легкий** 022131;
- 2) **пол** — **пол** 012131;
- 3) **белым** — **белым** 022135;
- 4) **с ножом** — **ножом** 01113516;
- 5) **лежит** — **лежит** 0721;
- 6) **у двора** — **дворе** 01213214.

В результате построения графа получится дерево с множеством вершин Vrt и дуг Arc :

$$Vrt = \{022131, 012131, 022135, 01113515, 020721, 022132, 01213214\};$$

$$Arc = \{(012131, 022131), (012131, 01113516), (01113516, 022135),$$

$$(012131, 0721), (0721, 01213214), (01213214, 022132)\}.$$

Данное дерево (Vrt, Arc) представляет собой СФ, порождающее предложение-дерево на рис. 3.

Выводы

На рассмотренном примере показана возможность конструктивной теории грамматики естественного языка, без которой компьютерная лингвистика будет довольствоваться только эвристическими методами индуктивного поиска и перебора предполагаемых решений в самых актуальных задачах лингвистики.

- 1) Построение (синтез) и распознавание (анализ) морфологических форм слов (словоформ) русского языка на компьютере не нуждается в поиске и теоретическом обосновании соответствующих алгоритмов. Они известны и формализованы [2, 4].
- 2) Форма слова задаётся и однозначно определяется исходной формой этого слова и его морфологическими параметрами, которые могут быть кодированы двухразрядными числами. Словоформы распределены по лексическим группам; лексическая группа — множество, состоящее из словоформ с одинаковым набором морфологических параметров; лексическая группа имеет такой же код, что и её элементы.
- 3) Основу синтаксиса составляет конечное множество словосочетаний; словосочетание — связанная пара словоформ; признаком связности являются отношения между морфологическими параметрами словоформ. Связность двух словоформ не зависит от их значений.
- 4) Из п. 2 следует, что понятие связности двух словоформ механически (по способу их кодирования) переносится на лексические группы.
- 5) Если две лексические группы образуют связанную пару, то их прямое произведение состоит из связанных словоформ (словосочетаний). Связанных лексических

групп в грамматике русского языка имеется около двухсот [5, 6]. В нашем примере были рассмотрены только пять из них, которые были разбиты на 34 группы для уточнения и удобства представления словосочетаний. Мы их назвали СО.

- 6) Два множества — одно состоит из словоформ, другое — из словосочетаний — образуют структуру данных, известную в математике как ориентированный граф. Обозначим его $G1$.
- 7) Введённое в п. 4 понятие связности двух словоформ задаёт ещё один граф (обозначим его через $G2$); вершинами $G2$ являются лексические группы; две вершины связаны дугой, если они — связанные группы. Граф $G2$ представляет редукцию графа $G1$, размеры (число вершин и дуг) которого в тысячи раз больше размеров $G2$ (если в $G1$ включить всю лексику русского языка).
- 8) Все понятия синтаксиса (подлежащее, сказуемое, однородные члены, дополнения и т. д.) определяются в терминах теории графов. В частности, предложением является любое корневое дерево в $G1$, содержащее дугу, соединяющую подлежащее (корень дерева) со сказуемым [4, 6]. Поэтому граф $G1$, образованный из словоформ и словосочетаний, назван синтаксисом $G1$.
- 9) Понятие предложения естественным образом переносится из графа $G1$ в граф $G2$: корневое дерево в $G2$, содержащее дугу, соединяющую лексическую группу из подлежащих (корень дерева) с лексической группой из сказуемых, называется синтаксической формой (предложения). Синтаксическая форма представляет собой правило, порождающее определённый класс предложений.
- 10) По заданной синтаксической форме легко строятся соответствующие ей предложения. Этот процесс называется синтезом предложений.
- 11) Для заданного предложения нетрудно найти порождающую его синтаксическую форму. Этот процесс называется анализом предложения.
- 12) Процедура нахождения синтаксической формы из п. 11 необходима для экспериментального обоснования корректности описанной грамматики: предложения русского языка (из литературных источников) являются предложениями в формальной грамматике.
- 13) Процедура нахождения синтаксической формы может быть использована для экспериментального изучения статистики разнообразия синтаксических форм, которыми пользуются различные авторы — их не более двух десятков в арсенале известных писателей.
- 14) В дальнейшем при построении конструктивной теории языка фрагменты семантики инкапсулируются в синтаксические формы, из которых состоит семантический словарь языка.
- 15) Если носители русского языка пользуются двумя десятками синтаксических форм и программы, анализирующие предложения, порождённые этими формами, невелики, то имеет смысл разработать соответствующие программы (синтаксические примитивы, по аналогии с примитивами в графических редакторах) для реализации на компьютере грамматики русского языка, а далее и языка в целом.

ЛИТЕРАТУРА

1. Мельчук И. А. Опыт теории лингвистических моделей «СМЫСЛ — ТЕКСТ». М.: Школа «Языки русской культуры», 1999. 346 с.
2. Зализняк А. А. Грамматический словарь русского языка. М.: Русский язык, 1977. 880 с.
3. Харари Ф. Теория графов. М.: Мир, 1973. 300 с.
4. Тестелец Я. Г. Введение в общий синтаксис. М.: РГГУ, 2001. 796 с.

5. *Виноградов В. В.* Грамматика русского языка. В 2 т. М.: АН СССР, 1960. 1864 с.
6. *Шихиев Ф. Ш.* Формализация и сетевая формулировка задачи синтаксического анализа: дис. ... канд. физ.-мат. наук. СПб.: СПбГУ, 2006. 171 с.

REFERENCES

1. *Mel'chuk I. A.* Opyt teorii lingvisticheskikh modeley «SMYSL — TEKST» [The Experience of the Theory of Linguistic Models "SENSE — TEXT"]. Moscow, Shkola "Yazyki russkoy kul'tury", 1999. 346 p. (in Russian)
2. *Zaliznyak A. A.* Grammaticheskiy slovar' russkogo yazyka [Russian Grammar Dictionary]. Moscow, Russkiy Yazyk Rubl., 1977. 880 p. (in Russian)
3. *Harary F.* Graph Theory. Addison-Wesley Pub. Co., 1969. 274 p.
4. *Testelets Ya. G.* Vvedenie v obshchiy sintaksis [Introduction to the Common Syntax]. Moscow, RSUH Publ., 2001. 796 p. (in Russian)
5. *Vinogradov V. V.* Grammatika russkogo yazyka [Russian Grammar]. Moscow, USSR AS Publ., 1960. 1864 p. (in Russian)
6. *Shikhiev F. Sh.* Formalizatsiya i setevaya formulirovka zadachi sintaksicheskogo analiza [Formalization and Network Formulation of the Task of Parsing]. PhD Thesis, St. Petersburg, SPBU, 2006. 171 p. (in Russian)