МОДУЛЬ СИНТАКСИЧЕСКОГО АНАЛИЗА ДЛЯ ЛИТЕРАТУРНОГО РУССКОГО ЯЗЫКА

И. А. Кагиров[1*], Ан. Б. Леонтьева[2]

[1,2]Санкт-Петербургский институт информатики и автоматизации РАН

[1,2]СПИИРАН, 14-я линия ВО, д. 39, Санкт-Петербург, 199178

[1]<kagirov@iias.spb.su>, [2]<an_leo@iias.spb.su>

УДК 681.3

Кагиров И.А., Леонтьева Ан. Б. **Модуль синтаксического анализа для литературного русского языка** // Труды СПИИРАН. Вып. 6. — СПб.: Наука, 2008.

Аннотация. *Заявленная в заглавии данной статьи тема подразумевает описание концепции и способов реализации программного модуля синтаксического анализа для современного литературного русского языка. Описание организовано следующим образом: сначала излагаются теоретические принципы автоматического синтаксического анализа, а затем представляется способ программной реализации модуля.* — Библ. 10 назв.

UDC 681.3

Kagirov I. A., Leontyeva An. B. **Module for syntax parsing of the literary Russian language** // SPIIRAS Proceedings. Issue 6. — SPb.: Nauka, 2008.

Abstract. *The topic presented in the headline considers describing of conception and ways of elaborating a program module of syntax parsing for the contemporary literary Russian language. The description runs as follow: in the beginning the theory of syntax analyze are presented, then the program realization is introduced.* — Bibl. 10 items.

1. Введение

Модуль синтаксического анализа (МСА) — это программа или часть программы, выполняющая синтаксический анализ. На сегодня создание автоматического МСА является одной из самых актуальных задач в компьютерной лингвистике, решение которой позволило бы достичь высокого уровня формализации языковых структур в разнообразных прикладных целях: от создания систем автоматического распознавания речи до поисковых систем в Интернете.

Под целью синтаксического анализа в настоящей статье понимается вычленение базовых синтаксических структур (см. разд. 3) и установление синтаксических связей между ними. На выход МСА поступает цепочка слов, разбитая на группы, причем каждая группа имеет связь с другими группами. Кроме того, МСА идентифицирует такие синтаксические категории, как подлежащее и сказуемое.

Однако создание МСА для русского языка упирается в большое количество сложностей, связанных с недостаточно разработанной теоретической базой в общем и прикладном языкознании; структуры человеческого языка отличаются разнообразием и часто высоким уровнем сложности, предусмотреть который чрезвычайно тяжело. В связи с этим в настоящей статье предлагается структура МСА, работающего с простыми синтаксическими структурами (см. разд. 6); создание МСА, справляющегося с текстом на русском языке любой сложности, представляется на настоящем этапе невозможным.

Предложенный нами модуль синтаксического анализа является частью анализатора текста русского языка SMART (на нынешний момент разрабатыва-

[*] Данное исследование проводится в рамках гранта РФФИ № 07-07-00073-а.

ется), который может работать в двух режимах. Первый режим заключается в морфологическом анализе словоформ, второй — в синтаксическом разборе предложения. Впрочем, оба режима взаимосвязаны и дополняют друг друга. Также SMART может входить в структуру декодера русской речи. Модуль синтаксического анализа предназначен для синтаксического разбора предложений русского языка. Он может выступать в качестве самостоятельной программы или же работать в составе декодера русского языка. Поскольку разработанный декодер построен на распознавании основы слова, то могут возникнуть проблемы с корректным подбором окончаний (учитывая морфологические особенности русского языка). Синтаксический анализатор, определяя возможные связи членов предложения, позволяет скорректировать окончания слов.

Синтаксический анализ — это анализ синтаксических структур некоторого языка по критериям, предусмотренным синтаксисом (грамматикой) этого языка. Следовательно, формальное описание синтаксиса языка является по сути и основой для синтаксического анализа. Раз так, описание модуля синтаксического анализа для русского языка лучше начинать с описания синтаксиса русского языка.

2. Грамматика зависимостей

Под синтаксисом понимается такой уровень языка, максимальными и основными единицами которого являются предложения, а минимальными — грамматические слова (словоформы). Ниже предложением называется грамматически связная цепочка слов, выражающая некоторое суждение. Грамматически связная цепочка — такая цепочка, в которой словоформы находятся в определенных грамматических отношениях между собой. Словоформа, или грамматическое слово, это слово (лексема) в одной из своих грамматических форм (характеризующейся определенными для каждого языка грамматическими признаками; так, для русского существительного это падеж и число).

Получается, что синтаксическая структура предложения, в сильном упрощении, представляет собой цепочку, которая состоит из конечного множества словоформ, связанных между собой некоторыми синтаксическими отношениями. С математической точки зрения любое предложение может быть представлено как направленный граф, а именно дерево. Вершины дерева связываются подчинительными связями: если между вершинами (т.е. словоформами) налицо отношение зависимости $X \to Y$, то будем говорить, что X подчиняет Y, а Y зависит от X: X называется вершиной, а Y — зависимым. Существуют три типа подчинительной связи между словоформами: *управление*, *согласование* и *примыкание*.

1. Управление — такой тип связи, при котором главенствующий компонент (вершина) словосочетания требует постановки зависимого компонента (зависимого) в определенной грамматической форме, причем изменение формы главенствующего слова не вызывает изменения формы управляемого слова. Выбор формы зависит от лексических свойств главенствующего компонента: *убить волк-**а*** (а не *волк-**е*** или *волк-**Ø***), *перед домом* (*а не дом-**у*** или *дом-**ов***) и т.п.
2. Согласование — такой тип связи, при котором в зависимом слове повторяются морфологические показатели (или их часть) главного слова: *красив-**ая** девушка* (а не *красив-**ый** девушка*), *больш-**ой** мир* (а не *больш-**ие** мир*), и т.п.

3. Примыкание — такой тип связи, при котором вершина и зависимое находятся в простом синтаксическом соположении, причем форма зависимого не определяется формой вершинного компонента: *очень рано*, *слишком быстро*, *совсем большой*.

Таким образом, наличие синтаксической подчинительной связи не доказывается, а постулируется априорно. Достаточно важным представляется тот факт, что подчинительной связью связаны все элементы цепочки-предложения (в этом легко убедиться, взглянув на любое синтаксическое дерево). Можно сказать, что наличие подчинительных синтаксических связей является признаком вхождения единиц в систему одного предложения, в котором всегда есть одна вершина, независимая ни от чего, все остальные элементы от чего-то зависят и каждый элемент может зависеть только от одного другого элемента.

Важным эмпирическим обобщением является следующее: для большинства предложений естественного языка отношения зависимости образуют дерево. Другими словами, не наблюдаются структуры, образующие замкнутый контур. Основы такого представления предложения были заложены Л. Теньером в [1] с той особенностью, что минимальной таксономической единицей анализа являлись «ядра» (т.е. члены предложения), а не словоформы. Согласно Л. Теньеру, построить стемму (т.е. дерево зависимостей) предложения — значит преобразовать линейный порядок в структурный. Этот принцип используется и в настоящей статье при описании грамматики зависимостей.

3. Фразовые категории и необходимость их введения

Множество синтаксических явлений в пределах предложения трудно описать, опираясь исключительно на взаимоотношения между терминальными элементами — минимальными синтаксическими единицами (словоформами) [2]. Поэтому в синтаксический анализ вовлекаются иерархически организованные единицы более высокого уровня — фразовые категории (ФК, англ. *Phrasal Category*). ФК — это группа, в которой имеется одна вершина, а также может быть одно или несколько зависимых от этой вершины.

На сегодня ФК широко применяются при описании грамматик языков самого разного типа. ФК чрезвычайно удобны при формальном описании синтаксических структур, ибо характеризуются относительной самостоятельностью в предложении, что позволяет рассматривать их фактически как минимальные единицы синтаксического уровня. Кроме того, в каждом языке ФК имеют обычно прозрачную, жестко иерархизированную структуру, что позволяет описать синтаксис языка сравнительно небольшими усилиями.

На рис. 1 представлена фразовая категория A, состоящая из терминальных элементов a, b, c, d и f, причем вершиной является элемент a. В то же время группа из элементов d, e и f также составляет фразовую категорию (назовем ее D) с вершиной d. Как видно из рисунка, фразовая категория D входит во фразовую категорию A.

Труды СПИИРАН. 2008. Вып. 6. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2008. Issue 6. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

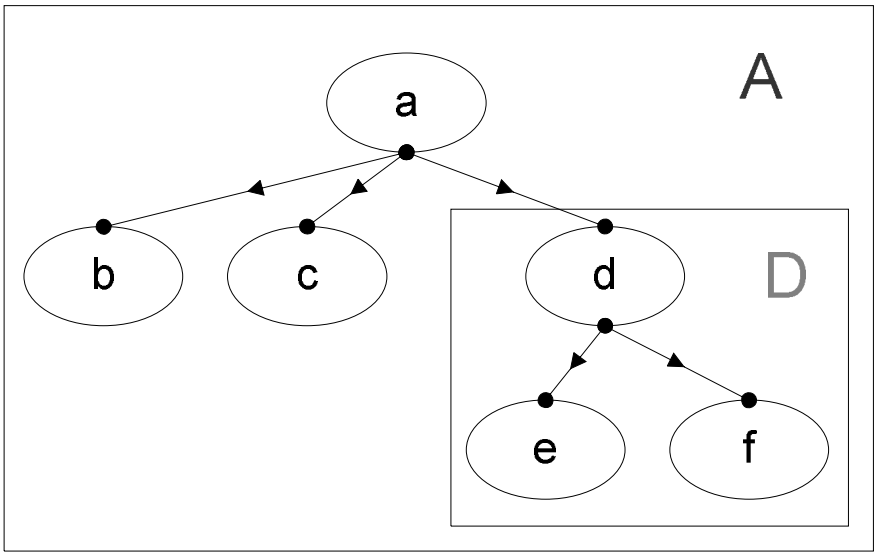

Рис. 1. Структура ФК.

Фразовые категории образуют ***систему составляющих*** некоторого предложения — множество отрезков предложения, в котором:

1. Есть отрезок, совпадающий со всем предложением;
2. Минимальные отрезки структуры составляющих — словоформы предложения;
3. Если составляющая A входит в составляющие B и C, то либо B входит в C, либо C входит в B. (Составляющие не могут частично пересекаться.)

Или, формулируя в виде дефиниции: «Системой составляющих на S (такая цепочка слов. — *И.К.*) является такое множество C отрезков S, которое содержит в качестве элементов само C и каждое слово, входящее в S, причем любые два отрезка, входящие в C, либо не пересекаются, либо один их них содержится в другом. Элементы C называются составляющими» [3].

Основанием для номенклатуры составляющих в основном является классификация вершин составляющих, прежде всего — их частеречная принадлежность. Таким образом, например, глагольная группа — это группа, вершиной которой является глагол, предложная группа — группа, вершиной которой является предлог, именная группа — группа, вершиной которой является существительное, и т.д. Эта классификация отражает хорошо известный факт: «выбор распространяющих средств определяется главным образом частью речи, к которой принадлежит распространяемое слово» [3]. Иными словами, свойства группы определяются свойствами ее вершины. Например, в предложении:

> Мама мыла [недавно окрашенную строителями раму из черного дерева], которая сильно скрипела.

ИГ [недавно окрашенную строителями раму из черного дерева] может быть заменена своей вершиной [раму] без изменения синтаксической структуры всего предложения в терминах НС.

Однако не все части речи, рассматриваемые в традиционной грамматике, задают соответствующие фразовые категории. Основное отклонение для русского языка — это местоимения. Оснований выделять особые «местоименные группы» нет; местоимения ведут себя так же, как фразовые категории. Напри-

Труды СПИИРАН. 2008. Вып. 6. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2008. Issue 6. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

мер, личные местоимения и местоимения-существительные обычно имеют дистрибуцию, сходную с именными группами, но не с терминальной категорией, т.е. существительным.

4. Структура клаузы

Важнейшей единицей синтаксического уровня языка является так называемая клауза («элементарное предложение», «предикация»). Под клаузой в настоящей статье понимается любая синтаксическая группа, распадающаяся на ГГ (вершиной является финитный глагол) и ИГ. От собственно предложения, соответствующего отдельному высказыванию, клауза отличается тем, что соответствует отдельной группе типа $C = ИГ; ГГ$, где C — это клауза, причем в случае с эллиптической конструкцией один из элементов в правой части выражения может быть опущен:

Начал писать статью. $C = ГГ$.

Пример клаузы на рис. 2:

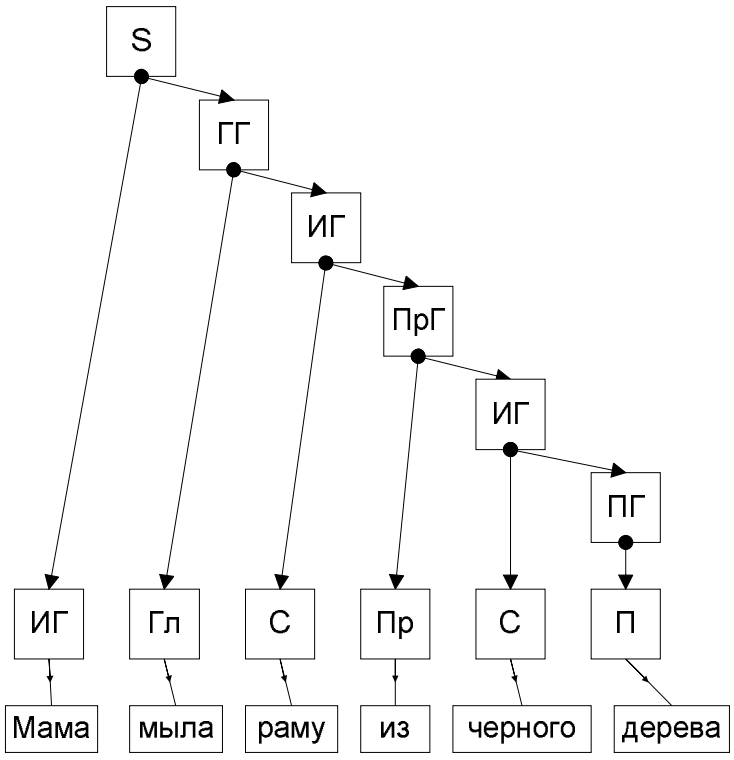

Рис. 2. Пример синтаксической структуры простой клаузы.

Когда глагольное сказуемое состоит из двух частей — полнозначного и вспомогательного глаголов, вершиной считается вспомогательный глагол — с лингвистической точки зрения носитель финитных категорий, которые и оформляют предложение: *он **начал** действовать*, *он **стал** много курить в последнее время*. От вспомогательного глагола следует отличать глагол-связку «быть»: *ты **будешь** врачом*, *он **был** приятным собеседником*. В настоящем времени глагол-связка «быть» опускается: *я (есть) врач*. Вслед за [3] в настоящей статье различаются три типа глагола «быть»:

Труды СПИИРАН. 2008. Вып. 6. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2008. Issue 6. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

1) вспомогательный глагол: *он **был** убит*;
2) глагол-связка: *день **был** хорош*;
3) полнозначный глагол: *часовой **был** на месте*.

Такие конструкции рассматриваются как фразовые категории типа ФГ (финитная группа, англ. *Inflection Phrase* (IP)), вершиной которых является глагол «быть», см. рис. 3:

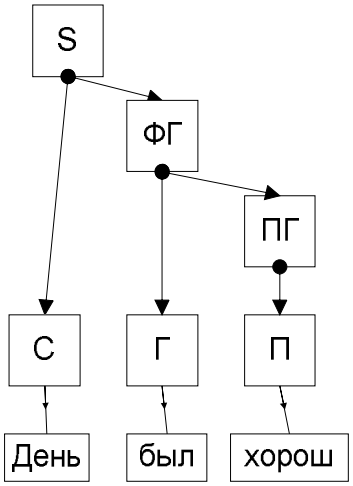

Рис. 3. Структура клаузы с ФГ.

5. Синтаксический анализ предложения

Сочетание различных ФГ порождает теоретически бесконечное число клауз, или предложений (см. подробнее [4] и [5]). Иными словами, число элементов синтаксических структур конечно, а сами структуры бесконечны, в том числе и из-за важного свойства ФГ — рекурсивности (т.е. способности включать в себя составляющую того же типа):

[младший сын [второй жены [моего отца]$_{ИГ}$] $_{ИГ}$] $_{ИГ}$… и т.д.

Поскольку число предложений бесконечно, при синтаксическом разборе имеет смысл ориентироваться на более мелкие единицы — ФК, введенные нами в разд. 2. Таким образом, алгоритм автоматического анализа сводится к вычленению ФК в составе предложения и поиску связей между ними.

Для разработки модуля автоматического синтаксического анализа был использован корпус текстов, состоящий из клауз с нераспространенной синтаксической структурой из [6]. Клаузы составлены в соответствии с нормами литературного русского языка. Этот корпус безусловно нуждается в расширении и усложнении, но на нынешнем этапе разработки модуля синтаксического анализа он отвечает основному поставленному требованию — идентификации отдельных ФК в структуре клаузы и определению связей между ними.

На основании анализа используемого корпуса были выделены пять основных синтаксических групп: именная группа (ИГ), глагольная группа (ГГ), группа прилагательного (ПГ), предложная группа (ПрГ), инфинитивная группа (ИнфГ). Для удобства за каждой группой был закреплен порядковый номер. Каждая синтаксическая группа имеет вершину, то есть слово, от которого зависят все остальные слова в группе. Вершиной ИГ является имя существительное или

Труды СПИИРАН. 2008. Вып. 6. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2008. Issue 6. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

личное местоимение; вершиной ГГ — личные формы глагола; вершиной ПГ выступает краткое прилагательное; вершиной ПрГ является предлог; вершиной ИнфГ — инфинитив. Связи и соотношения слов внутри групп представлены в табл. 1.

Символ «*» означает, что элементы группы могут стоять также и в обратном порядке. При этом тип связи между ними сохраняется неизменным. Например, зависящая от глагольной группы именная группа с типом связи «управление» может находиться как справа, так и слева от глагола, но в любом случае вершиной группы будет выступать глагол.

Следует отметить, что глагольные группы представлены двумя уровнями. Глагольная группа второго уровня ГГ” включает в себя, помимо глагольной группы первого уровня ГГ’, другие элементы. В данных группах определены следующие виды связи слов: управление, согласование и примыкание. Вид связи «управление» образуется, когда существительное стоит в определенном косвенном падеже. Вид связи «согласование» образуется, когда прилагательное принимает те же значения рода, числа и падежа, что и существительное-вершина. Вид связи «примыкание» состоит в синтаксическом соположении вершины и неизменяемого зависимого слова.

Таблица 1.

Типы ФГ, используемых в модуле синтаксического анализа

<table>
<tr><th>ИГ {Сущ}/{М}</th><th>ГГ’ {ГГ”}</th><th>ПГ {КрПрил}</th><th>ПрГ {Пр}</th><th>ГГ” {Глаг}</th><th>ИнфГ {Инф}</th></tr>
<tr><td>—</td><td>ИГ* управление</td><td>/ИГ управление</td><td rowspan="6">/ИГ управление</td><td>—</td><td rowspan="6">ИГ* управление</td></tr>
<tr><td>/{ИГ} управление</td><td>ИГ/ управление /ИГ управление</td><td rowspan="5">{Нар}/ примыкание</td><td>({AUX} → {КрПрил})</td></tr>
<tr><td>{Прил}/ согласование</td><td>ПрГ* примыкание</td><td>{Нар}* примыкание</td></tr>
<tr><td rowspan="3">/CON-{Сущ}</td><td>ИГ/ управление /ПрГ примыкание</td><td rowspan="3">{Нар}/ примыкание /{Нар} примыкание</td></tr>
<tr><td>/ИГ управление //ПрГ) примыкание</td></tr>
<tr><td>/ИнфГ примыкание</td></tr>
</table>

ИГ — вершиной является Имя существительное Сущ или местоимение М
ГГ” — вершиной является финитный глагол Глаг
ГГ’ — вершиной является группа ГГ”
ПГ — вершиной является краткое прилагательное КрПрил или прилагательное П
ПрГ — вершиной является предлог Пр
ИнфГ — вершиной является инфинитив Инф
AUX — вспомогательный глагол

Труды СПИИРАН. 2008. Вып. 6. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2008. Issue 6. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

Настоящая таблица представляет систематизацию ФК, встречающихся в 500 тестовых предложениях. Каждый столбец — ФК с вершиной одного типа. В строках таблицы расположены элементы, зависимые от вершины. Фактически, каждая строка — это один из вариантов ФК одного и того же типа. Знак «слеш» (/) обозначает расположение вершины, от которой зависит элемент; астериск (*) значит, что вершина может находиться как слева, так и справа от элемента. Стрелка (→) показывает направление синтаксической связи подчинения. В фигурных скобках стоят вершины каждой группы.

В каждой ФК действуют подчинительные связи одного из трех типов; на уровне морфологии это находит отражение в том, что при согласовании зависимое слово (Прил) принимает те же показатели рода, числа и падежа, что и вершина (Сущ или М); при примыкании наблюдается простое синтаксическое соположение вершины и неизменяемого слова-зависимого без дополнительного маркирования на морфологическом уровне, а при управлении зависимое слово (Сущ или М) стоит в определенном косвенном падеже, причем выбор падежа определяется по словарю, в характеристиках слова-вершины. Для определения падежа, в котором стоит зависимое слово при подчинительной связи, используется словарь [7]. Предполагается со временем создать свой словарь, специально приспособленный для нужд автоматического синтаксического анализа.

6. Описание алгоритма автоматического анализа синтаксической структуры предложения

Работа анализатора основывается на базе данных по ФК (см. табл. 1). В тексте ищутся только такие ФК, которые внесены в базу. Предложением (в случае с нераспространенными предложениями это клауза) называется отрезок текста между двумя показателями конца предложения — точкой/восклицательным знаком/вопросительным знаком + пробел и точкой/восклицательным знаком/вопросительным знаком.

Анализ начинается с того, что модуль морфологического анализа определяет морфологические характеристики и частеречную принадлежность анализируемого слова. Дальше начинается формирование гипотез о текущей ФК. Любая ФК может быть представлена в виде:

$$XГ = (x; Г),$$

где $XГ$ — название ФГ; x — вершина; $YГ$ — зависимое. $YГ$ может принимать значения $YГ = 0$, где 0 — это пустое множество:

$$XГ = волк;$$
$$x = волк;$$
$$YГ = 0.$$

или

$$YГ = y; ZГ,$$

где $ZГ$ — ФК, идентичная по структуре $YГ$:

Труды СПИИРАН. 2008. Вып. 6. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2008. Issue 6. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

$XГ = убил\ волка;$	$XГ = убил\ серого\ волка;$
$x = убил;$	$x = убил;$
$YГ = волка;$	$YГ = серого\ волка;$
$ZГ = 0$	$ZГ = серого$

Теоретически разложение $YГ$ на составляющие может быть бесконечным:

$_1$[кот, $_2$[который пугает и ловит синицу, $_3$[которая часто ворует пшеницу, $_4$[которая $_5$[в тёмном чулане]$_5$ хранится $_6$[в доме, $_7$[который построил (Джек)]$_1$]$_2$]$_3$]$_4$]$_5$]$_7$

На рис. 4 треугольником обозначена ФК, в левом углу расположена вершина (x), в сам треугольник вписано зависимое (YP), у правого угла раскрывается структура зависимого. В конце цепочки, у правого угла, пишется последнее зависимое. Цифрами помечены ФК.

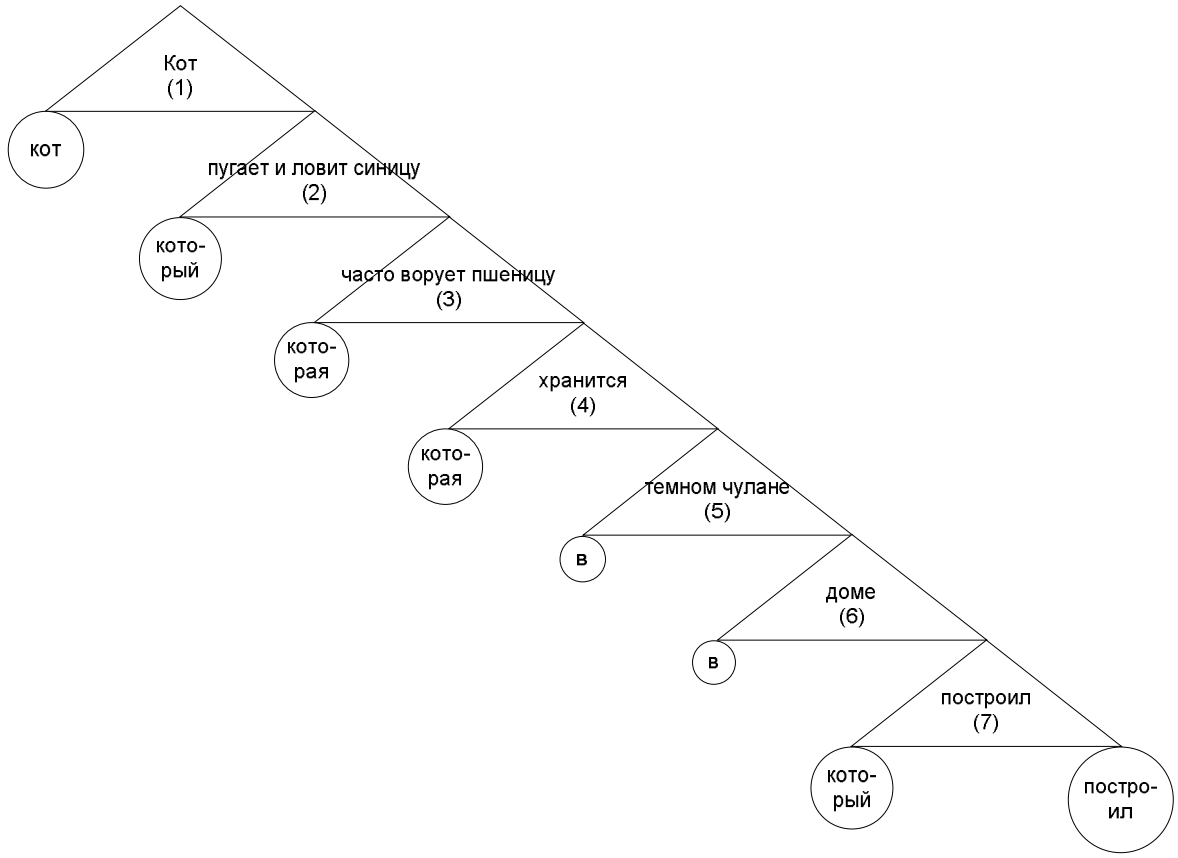

Рис. 4. Пример рекурсивного предложения.

Следующая задача — это определить по базе данных типы ФК, в которые может входить анализируемая словоформа. Это решается за счет простого перебора НС, которые возможны в текущем контексте при данных морфологических характеристиках слова.

После формирования гипотезы относительно ФК анализатор переходит к поиску и анализу следующей словоформы; далее аналогичным методом ищутся вершины и зависимые (см. разд. 7). По сути, все предложение разлагается

Труды СПИИРАН. 2008. Вып. 6. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2008. Issue 6. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

на две группы — группу подлежащего (ИГ) и группу сказуемого (ГГ), построенных по модели $XГ = (x; YГ)$. Причем если словоформа A — существительное в именительном падеже, то у него нет вершины, оно объявляется подлежащим; если словоформа B — глагол в личной форме, у него нет вершины, он объявляется сказуемым и согласуется в числе и лице с подлежащим.

При определении связи между словоформами используются введенные нами в разд. 1 понятия согласования, управления и примыкания (в общем случае, прилагательное согласуется с существительным в роде, числе и падеже, глагол с подлежащим — в числе и лице, зависимое существительное стоит в одном из косвенных падежей).

Анализ идет до тех пор, пока все словоформы в предложении не будут связаны друг с другом.

7. Программная реализация МСА

Структурная схема анализатора представлена на рис. 5:

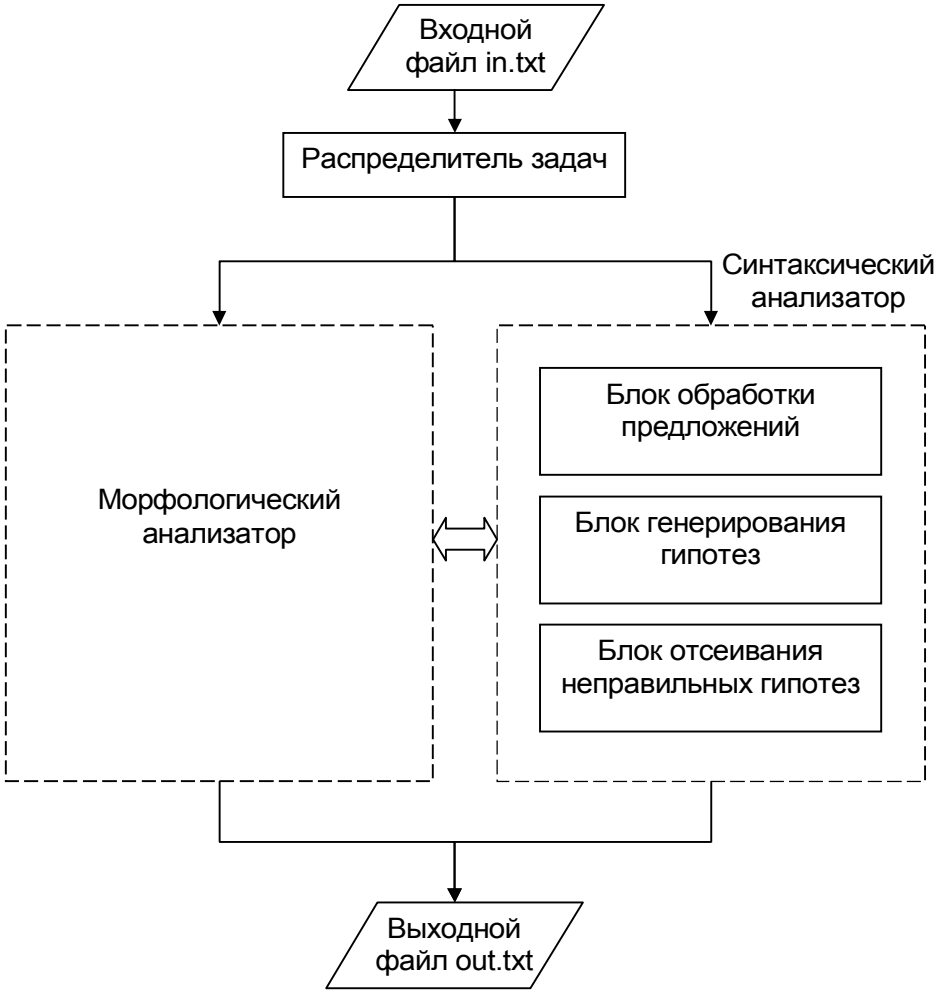

Рис. 5. Структурная схема модуля синтаксического анализа.

Как было сказано выше, входные данные представляют собой список простых предложений. Распределитель задач передает эти данные в блок синтаксического анализа.

В блоке обработки предложения обрабатываются пословно. Исходная словоформа предается в блок морфологического анализа [8], в котором для нее подбираются все возможные варианты основ и соответствующие грамматические показатели. Каждое предложение считывается пословно, после чего словоформа поступает на вход морфоанализатора. В результате определяются все возможные основы и соответствующие грамматические показатели. Если данной словоформе соответствует только одна основа, она поступает в процедуру построения гипотез. В зависимости от части речи и грамматических показателей выделяется соответствующая синтаксическая группа. В том случае, когда одной словоформе соответствует несколько основ, существующие гипотезы копируются в свободные ячейки памяти и построение гипотез осуществляется для всех вариантов основ. Если словоформа не найдена в словаре, то в файл печатается неразобранное предложение.

После предварительной обработки словоформа поступает в блок генерирования гипотез. Этот блок является основным. На его вход поступает словоформа. Если это первое слово в предложении, то в соответствии с частью речи определяется синтаксическая группа. На основании исходного текста был составлен список частей речи, которые в нем встречаются. Он включает в себя: имя существительное, финитный глагол, прилагательное, краткое прилагательное, инфинитив, наречие, предлог, сочинительный союз «и», вспомогательный глагол «быть» и другие части речи. Таким образом, было выделено десять структурных элементов, из которых состоят синтаксические группы (см. табл. 1).

Как было сказано ранее, первое слово в предложении определяет синтаксическую группу. В зависимости от частей речи группа может определяться однозначно, либо могут быть варианты. Например, если на вход поступило имя существительное, то первая группа в предложении будет именной. В этом случае запоминается порядковый номер группы и обрабатывается следующее слово. Если же первым словом в предложении является наречие, то оно может относиться как к глагольной группе, так и к группе прилагательного. В этом случае для однозначного определения группы требуется следующее слово.

Начиная со второго слова в предложении, важную роль играет не только часть речи данной словоформы, но и информация о группе или группах, которые выделены на данный момент. Поступившая на вход словоформа может как принадлежать текущей синтаксической группе, так и выделяться в другую синтаксическую группу. В этом случае формируется дополнительная гипотеза, и рассматриваются оба варианта. В конечном счете, в предложении выделяется группа подлежащего и группа сказуемого. Важно отметить, что некоторые группы, например глагольная, могут содержать в себе другие группы. История выделения слов в группы сохраняется в виде индекса.

Таким образом, из-за морфологической и синтаксической неоднозначности для одного предложения может быть сформировано несколько гипотез, ср. [9], [10]. Далее эти гипотезы поступают в блок отсеивания неправильных гипотез. Данный блок имеет два уровня проверки. На первом уровне проверяется согласование синтаксических групп в рамках одного предложения. Это согласование определяется исходя из грамматических характеристик вершин групп. То есть группа подлежащего и группа сказуемого должны согласоваться по роду, лицу и числу. Если согласование подтверждается, то гипотеза поступает на второй

уровень проверки. В противном случае — гипотеза отсеивается. На втором уровне проверяются связи слов внутри каждой группы. В зависимости от типа связи слов внутри группы (управление, согласование или примыкание) проверяются соответствующие грамматические характеристики элементов группы. Если связь не подтверждается — гипотеза отсеивается.

Выходной файл представляет собой список предложений, каждое из которых разбито на синтаксические группы. Если предложение содержит слово, отсутствующее в словаре, то оно выводится без разбора.

Теперь на примере рассмотрим работу программы. Пусть на вход поступило предложение:

Химия и физика — интересные науки.

Первое слово «химия» является существительным, таким образом, однозначно определяется именная группа:

Химия — Сущ

И — CON

Физика — Сущ

Поскольку в данном случае никаких неоднозначностей не возникает, то группа закрывается и следующее слово будет относиться к новой группе. Следующим словом является прилагательное «интересные», которое также однозначно определяет именную группу:

Интересные — Прил

Науки — Сущ

Слово «науки» является последним в группе и в предложении. Таким образом, имеется группа подлежащего и группа сказуемого:

[Химия и физика]иг — группа подлежащего;

[интересные науки]иг — группа сказуемого.

Принадлежность к группе подлежащего или группе сказуемого определяется частью речи вершины отдельно выделенной группы (т.е. группа не должна быть вложенной в другую группу). В данном примере вершинами первой группы являются имена существительные в именительном падеже (химия, физика), следовательно, они являются подлежащими в этом предложении. Вершиной второй группы является существительное в именительном падеже (науки), оно же выделяется в качестве сказуемого. Во второй группе тип связи — согласование, следовательно, нужно проверить род, число и падеж. В данном случаем оба слова имеют одинаковые грамматические характеристики: множественное число, именительный падеж. Так как у прилагательных род во множественном числе совпадет, то предложение является согласованным.

Таким образом, выходной файл будет содержать разобранное предложение, записанное в виде:

[Химия и физика]иг — [интересные науки]иг.

В дальнейшем планируется добавить разбор сложных предложений, а также сделать выходные данные более информативными, представляя структуру предложения в виде дерева и с указанием типов связей.

Труды СПИИРАН. 2008. Вып. 6. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2008. Issue 6. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

8. Заключение

В результате описанных в настоящей статье разработок был создан модуль синтаксического анализа текста русского языка. Данный модуль включает в себя полноценный морфоанализатор, который может работать самостоятельно. Синтаксический модуль позволяет производить разбор предложения. В ходе разбора выделяются группы подлежащего и сказуемого, а также устанавливаются связи слов в предложении. В качестве исходного материала использовался ГОСТ 16600-72. На основании списка из 500 фраз, содержащихся в нем, была разработана основная структура синтаксического анализатора, определены зависимости между членами предложения и критерии выделения синтаксических групп. Данная работа является только первым шагом при разработке полноценного синтаксического анализатора. В будущем планируется сделать более информативный вывод, представляя предложения в виде дерева. Таким образом, будут наглядно показаны синтаксические зависимости внутри предложений и указаны типы связи. Также планируется провести тестирование анализатора на произвольном тексте, взятом из художественной или научной литературы. На данном этапе разбор производится только для простых предложений, так что планируется разработать и реализовать алгоритм разбора сложных предложений. Модуль синтаксического анализа может работать как самостоятельная программа, а может выступать в качестве синтаксической составляющей декодера для русской слитной речи. Поскольку словарь распознавателя содержит в себе отдельно основы и окончания, то в связи с вариативностью русского языка могут быть ошибки при подборе окончаний. Разрабатываемый синтаксический анализатор позволит устранить подобные ошибки, таким образом, повысив качество распознавания русской слитной речи.

Литература

1. *Tesniére L.* Eléments de syntaxe structurale. Paris: Librairie Klincksieck, 1959. xxvi, 670 p.
2. *Фитиалов С. Я.* Об эквивалентности грамматик НС и грамматик зависимости // Проблемы структурной лингвистики, 1967. М.: Наука, 1967. 145 с. С. 16–43.
3. *Тестелец Я. Г.* Введение в общий синтаксис. М.: Российский государственный гуманитарный университет, 2001. 800 с.
4. Современная американская лингвистика: Фундаментальные направления. М.: Едиториал УРСС, 2006. 480 с.
5. *Мельчук И. А.* Автоматический синтаксический анализ. Т.1. Новосибирск: Наука, 1964. 357 с.
6. Передача речи по трактам радиотелефонной связи: Требования к разборчивости речи и методы артикуляционных изменений. ГОСТ 16600 — 72. М.: Изд=во стандартов, 1973. 90 с.
7. Большой толковый словарь русского языка / Под ред. Д. Н. Ушакова. М.: Альта-принт, 2005. 1239 с.
8. *Kagirov I. A., Leontyeva An. B.* Grammar-Based Speech- and Word-splitting // Proceedings of 3rd Language & Technology Conference. October 5-7, Poznań, Poland. Poznań: Fundacja Uniwersytetu im. A. Mickiewicza, 2007. 578 s. P. 413–417.
9. *Попов Э. В.* Общение с ЭВМ на естественном языке. М.: Едиториал УРСС, 2004. 260 с.
10. *Mel'čuk I. A.* Dependency syntax: Theory and practice. Albany, NY: SUNY Press, 1988. 428 p.