УДК 519.7

Н.А. Валенда, Л.Д. Самофалов, Н.А. Павленко

Харьковский национальный университет радиоэлектроники, Харьков

АНАЛИЗ СЛОВОСОЧЕТАНИЙ РУССКОГО ЯЗЫКА НА ОСНОВЕ LR-ГРАММАТИК

В работе рассматривается поэтапный анализ словосочетаний русского языка и получение их формального представления. Рассмотрены различные виды связей в словосочетаниях. В статье рассмотрено применение аппарата формальных грамматик для описания синтаксиса словосочетаний русского языка и разбор словосочетаний с помощью технологии LR-анализа.

Ключевые слова: *синтаксический анализ, атрибутная грамматика, LR-анализ, семантические правила, лингвистический процессор.*

Введение

При разработке современных интеллектуальных систем, выполняющих обработку естественно-языкового текста, основным этапом является реализация компоненты, обеспечивающей корректное преобразование языка в формальное представление. Обычно такой компонентой является лингвистический процессор. Лингвистический процессор – это программное обеспечение, выполняющее анализ и синтез естественно-языкового текста в некое формальное представление, позволяющее корректно его интерпретировать и обрабатывать [1]. Обычно лингвистический процессор выполняет следующие основные этапы анализа естественно-языкового текста: лексический, морфологический, синтаксический, семантический и прагматический [2]. Предметом рассмотрения данной статьи является этап синтаксического анализа словосочетаний русского языка. Такой анализ будет проще и точнее, чем анализ целого предложения. Последующая обработка должна производиться на уровне семантики.

1. Структура словосочетания

Синтаксис - это раздел грамматики, предметом которого является синтаксический строй языка – его синтаксические единицы, связи и отношения между ними [3]. В предложении между словами существуют синтаксические связи и отношения, которые либо идут от слова как лексико-грамматической единицы – они предопределены и не зависят от синтаксических функций, выполняемых в предложении, либо возникают в предложении и обусловлены синтаксическими позициями соединяющих слов. Синтаксическая связь, предопределенная не синтаксической позицией слова, а самим словом, называется присловной подчинительной связью. Способность слова соединяться с другими словами называется сочетаемостью слова. В современном русском языке выделяют три вида подчинительной связи: согласование, управление и примыкание. Согласование выражается уподоблением формы зависимого слова форме главного слова.

Управление выражается присоединением к главному слову существительного в форме косвенного падежа. При примыкании в роли зависимого слова выступают неизменяемые слова.

Словосочетание – это синтаксическая конструкция, сформированная на основе подчинительных связей [4]. В состав любого словосочетания входит одно главное слово, которое своими лексико-грамматическими свойствами предопределяет связь с одним или несколькими зависимыми словами, грамматически подчиненными главному. Основные типы словосочетаний, в зависимости от главного слова, представлены в табл. 1.

Таблица 1

Типы словосочетаний

Тип	Главное слово
Глагольное	Глагол
Субстантивное	Существительное, местоимение-существительное, числительное
Адъективное	Прилагательное
Наречное	Наречие, компаратив

Каждое главное слово обладает свойствами, предопределяющими его сочетаемость:

- принадлежность слова к какой-либо части речи;
- морфологические значения слова;
- морфемный состав слова;
- лексическая семантика слова.

Данные свойства называются лексико-грамматическими характеристиками слова.

Для словосочетания характерен грамматический порядок слов, определяемый видом подчинительной связи и лексико-грамматическими характеристиками главного слова.

Синтаксический анализ естественно-языкового текста – это выделение в тексте синтаксических единиц, определение отношений между ними и построение дерева разбора этого текста. Соответственно, основанная задача синтаксического анализа словосочетания заключается в выделении главного и зависимого слова и анализ отношений между ними на основе лексико-грамматических характеристик главного слова. Приведенный в данной статье метод синтаксического анализа учитывает первые два лексико-грамматические свойства анализируемых слов.

2. Формализация синтаксиса словосочетаний

Словосочетания строятся по определенным абстрактным образцам (схемам), определяющим их формальное устройство. Такие схемы строятся для каждого типа словосочетания, исходя из лексико-грамматических свойств главного слова. Они имеют следующий вид: главное слово + зависимое слово, и представлены в табл. 2.

Таблица 2

Схемы словосочетаний

Главное слово	Зависимое слово
Глагол	Существительное, прилагательное, числительное, местоимение, наречие, деепричастие, компаратив, инфинитив
Существительное	Прилагательное, существительное, причастие, числительное, инфинитив, наречие, компаратив
Местоимение-существительное	Местоимение-существительное
Числительное	Существительное, прилагательное, причастие
Прилагательное	Существительное, инфинитив, наречие
Наречие	Существительное, прилагательное, наречие
Компаратив	Компаратив

По своему составу словосочетания делятся на простые, сложные и комбинированные. Простое словосочетание образуется на основе одной подчинительной связи. Сложное словосочетание образуется на основе двух или более подчинительных связей, исходящих от одного главного слова. Комбинированное словосочетание образуется на основе связей, исходящих от разных главных слов. Зависимое слово в таком словосочетании одновременно может являться главным словом другого простого или сложного словосочетания. Из этого следует, что комбинированное словосочетание состоит из нескольких простых и/или сложных словосочетаний. Наличие такой иерархии позволяет представить любое словосочетание в виде дерева. Листьями такого дерева являются слова. Узлы и вершина – это словосочетания. В каждый узел и вершину входит два ребра – от главного и зависимого слова или словосочетания. Вершина определяет тип, главный и зависимый компонент в рамках всего словосочетания. Возможность построения такого дерева является необходимым и достаточным условием синтаксической корректности словосочетания. Для проведения синтаксического анализа словосочетаний необходимо наличие двух компонентов:

– модель, позволяющая в формальном виде представить словосочетание и описывающая возможные отношения между элементами этого формального представления;

– алгоритм анализа словосочетаний в соответствии с правила модели.

Так как существует конечное число схем формирования словосочетаний, то такую часть грамматической системы можно представить в виде формальной грамматики. Наиболее естественным способом такого описания является формализация с помощью атрибутной контекстно-свободной грамматики [5]. Продукции такой грамматики имеют следующий вид:

$$< Б >_{c_1, ..., c_k} \rightarrow б_{c_1, ..., c_k},$$

где А – символ нетерминального алфавита, α – строка грамматических символов.

Каждый грамматический символ в такой продукции имеет атрибуты $c_1, ..., c_k$. Продукции такого вида отображают структуру и грамматические свойства словосочетания: правая часть продукции описывает структурную схему словосочетания с помощью терминальных (отдельные слова) и/или нетерминальных (представляющих грамматические свойства словосочетания, входящего в текущее) символов, нетерминальный символ в левой части продукции характеризует грамматические свойства словосочетания, атрибуты описывают значения морфологических категорий слова.

Значения атрибутов символа в левой части продукции определяются значениями атрибутов главного грамматического символа – эти операции задаются с помощью семантических правил продукции.

Атрибутная КС-грамматика, описывающая синтаксис словосочетаний русского языка имеет следующий вид:

$$G = (V_t, V_n, P, S),$$

где V_t – алфавит терминальных символов. В данном случае – это множество частей речи, формирующих словосочетания;

V_t – {V (глагол), N (существительное), adj (прилагательное), Adv (наречие), Inf (инфинитив), Pron (местоимение), Part (причастие), numeral (числительное), compr (компаратив), gerund (деепричастие), prepos (предлог)};

V_n – алфавит нетерминальных символов. Он содержит типы словосочетаний в зависимости от главного слова;

V_n – {VP (глагол), NP (существительное), PronP (местоимение-существительное), NumeralP (числительное), AdjP (прилагательное), AdvP (наречие), CompP (компаратив), Prep (предлог), S (словосочетание)};

P – система продукций. Она содержит схемы формирования словосочетаний;

S – аксиома. Это символ нетерминального алфавита. Если текущая входная строка сворачивается в данный символ, то синтаксический анализ считается завершенным.

Введем следующие обозначения:

строчные буквы из начала алфавита, такие как a, b, c – терминальные символы,

прописные буквы из начала алфавита, такие как A, B, C – нетерминальные символы,

прописные буквы из конца алфавита, такие как X, Y, Z – грамматические символы,

строчные буквы греческого алфавита, такие как α, β, γ – строки грамматических символов.

Так как словосочетания образуют иерархию от простого к комбинированному, то продукция, описывающая такой вид зависимости, имеет следующий вид:

$$< Б >_{c_1, ..., c_k} \rightarrow б_{c_1,...,c_k} < Б >_{c_1, ...,c_k} в_{c_1, ..., c_k} | < Prep > a_{c_1, ..., c_k}.$$

При этом α и β не могут одновременно являться пустыми строками.

Данная продукция может иметь несколько альтернатив вида

$$б_{c_1,...,c_k} < Б >_{c_1, ...,c_k} в_{c_1, ..., c_k},$$

каждая из которых описывает определенную схему формирования словосочетания для данного главного слова.

Правая часть продукции вида

$$б_{c_1,...,c_k} < Б >_{c_1, ...,c_k} в_{c_1, ..., c_k}$$

представляет собой словосочетание, где в роли главного слова выступает слово $a_{c_1, ..., c_k}$ либо словосочетание с таким главным словом, грамматические характеристики которого представляет символ $< Б >_{c_1, ...,c_k}$, а в роли зависимого слова – другие слова или словосочетания, грамматические характеристики которых описывают остальные символы правой части продукции.

Правая часть вида

$$< Prep > a_{c_1, ..., c_k}$$

позволяет характеризовать грамматические свойства символа $< Б >_{c_1, ..., c_k}$ как свойства слова $a_{c_1, ..., c_k}$ при <Prep> → ε или падежную форму слова $a_{c_1, ..., c_k}$ с предлогом при <Prep> → prepos.

Таким образом, символ $< Б >_{c_1, ..., c_k}$ может представлять грамматические характеристики как отдельного символа, так и словосочетания на основе этого символа или другого словосочетания с данным главным символом.

Вид правых частей продукции

$$б_{c_1,...,c_k} < Б >_{c_1, ...,c_k} в_{c_1, ..., c_k},$$

описывающей словосочетание по его структурной схеме определяется видом подчинительной связи и лексико-грамматическими характеристиками главного слова, а также порядком слов.

Применяя приведенную выше продукцию и правила ее формирования для формализации синтаксиса словосочетаний, получаем следующую систему продукций.

P = {

<S> → <VP> | <NP> | <PronP> | <NumeralP> | <AdjP> | <AdvP> | <ComprP>,

<VP> → <VP> <NP> | <VP> <AdjP> | <VP> <QuantP> | <VP> <PronP>,

<VP> → <VP> <AdvP> | <VP> <CompP> | <VP> gerund | <VP> Inf | <Prep> V,

<NP> → <AdjP> <NP> | <NP> <NP> | <NP> <NumeralP> | <NP> <CompP>,

<NP> → <NP> <AdvP> | <NP> Inf | <NP> Part | <Prep> N,

<PronP> → <PronP> <PronP> | <PronP> <AdjP> | <PronP> <CompP>,

<PronP> → <Prep> Pron,

<NumeralP> → <NumeralP> <NP> | <NumeralP> <AdjP> | <NumeralP> Part, <NumeralP> → <Prep> numeral,

<AdjP> → <AdjP> <NP> | <AdjP> <AdvP> | <AdjP> Inf | <Prep> adj,

<AdvP> → <AdvP> <NP> | <AdvP> <AdjP> | <AdvP> <AdvP> | <Prep> Adv,

<CompP> → <CompP> <CompP> | <Prep> compr,

<Prep> → prepos | ε

}

Данная система продукций описывает все возможные схемы формирования словосочетаний, однако такая формализация основана только на принадлежности слова к той или иной части речи, и возможность существования таких связей ограничена другими лексико-грамматическими характери-

стиками. Следующим этапом формализации является описание подчинительных связей тех частей речи, слова которых выступают в роли главного слова словосочетания, то есть на данном этапе будут учитываться морфологические значения слова. Такое описание выполняется с помощью задания семантических правил для каждой продукции, описывающей структурную схему словосочетания. Семантическое правило выполняет проверку значений атрибутов соединяемых слов, представленных грамматическими символами, и присвоение значениям атрибутов символа в левой части продукции значений атрибутов главного слова. Вид семантического правила зависит от вида подчинительной связи между главным и зависимым словом и от частей речи, к которым принадлежат эти слова.

В общем случае, семантические правила, в зависимости от подчинительной связи, имеют вид, представленный в табл. 3.

Таблица 3

Виды семантических правил

Вид подчинительной связи	Вид семантического правила
Согласование	**if** $B.c_i = a.c_i$, $\forall a \in б \cup в$, $i = \overline{i,k}$ **then** acc(A, B); **else** unacc;
Управление	**if** (!$\exists$B.падеж **or** B.падеж = “именительный”) **and** a.падеж != “именительный”, $\forall a \in б \cup в$ **then** acc(A, B); **else** unacc;
Примыкание	acc(A, B);

Здесь B характеризует главное слово словосочетания, а A – символ в левой части продукции. Процедура acc выполняет присвоение значениям атрибутов символа A значений атрибутов символа B.

Процедура unacc – не допуск анализируемой продукции, то есть данное словосочетание является синтаксически некорректным.

При примыкании, с точки зрения формального выражения этого вида связи, все словосочетания будут считаться синтаксически корректными, так как зависимое слово является неизменяемым (нет морфологических категорий).

Выводы

Описание синтаксиса словосочетаний русского языка на основе атрибутных LR-грамматик позволяет использовать методы, разработанные для формальных грамматик, при построении синтаксического анализатора для текстов русского языка. Анализатор является универсальным, настройка на определенный язык производиться на основе таблицы анализа.

Использование методов, разработанных для формальных языков, при построении анализатора естественного языка, упрощает процесс создания такого приложения, делает его универсальным и легко модифицируемым, обеспечивает линейную зависимость скорости работы от объема входных данных.

Список литературы

1. Андреев А.М. Лингвистический процессор для информационно-поисковой системы [Электронный ресурс] / А.М. Андреев, Д.В.Березкин, А.В. Брик. – Интелтек издательство. – Режим доступа: http://inteltec.ru/ publish/articles /textan/art_21br.shtml – 3.06.2016 г. – Загл. с экрана.

2. Толпегин П.В. Информационные технологии анализа русских естественно-языковых текстов. Часть I [Текст] / П.В. Толпегин // Информационные технологии. – 2006. – № 8. – С. 41-50.

3. Валгина Н.С. Синтаксис современного русского языка [Текст] / Н.С. Валгина. – М.: Агар, 2000. – 416 с.

4. Русская грамматика [Текст] / Гл. ред. Н.Ю. Шведова. – М.: Наука, 1980. – Т. 2. – 709 с.

5. Ахо А. Компиляторы: принципы, технологии и инструменты [Текст]: Пер. с англ. / А. Ахо, Р. Сети, Д. Ульман – М.: Издательский дом “Вильямс”, 2003. – 768 с.

Поступила в редколлегию 18.05.2016

Рецензент: д-р физ.-мат. наук, проф. А.В. Грицунов, Харьковский национальный университет радиоэлектроники, Харьков.

АНАЛІЗ СЛОВОСПОЛУЧЕНЬ РОСІЙСЬКОЇ МОВИ НА ОСНОВІ LR-ГРАМАТИК

Н.А. Валенда, Л.Д. Самофалов, Н.А. Павленко

У статті розглядається використання формальних граматик для опису синтаксису словосполучень російської мови і розбір словосполучень за допомогою технології LR-аналізу.

Ключові слова: *семантичний аналіз, атрибутна граматика, LR-аналіз, семантичні правила, лінгвістичний процесор.*

ANALYSIS OF THE RUSSIAN LANGUAGE PHRASES BASED ON LR-GRAMMARS

N.A. Valenda, L.D. Samofalov, N.A. Pavlenko

This article is about the use of formal grammars to describe the syntax of the Russian language phrases, parsing phrases using LR-analysis.

Keywords: *semantic analysis, attribute grammars, LR-analysis, semantic rules, linguistic processor.*