*Доклады БГУИР*
*2007* *июль–сентябрь* *№ 3 (19)*

*УДК 519.688*

# КОМПЬЮТЕРНАЯ ОБРАБОТКА РУССКОГО ТЕКСТА

П.А. ВЕЙНИК

*Белорусский государственный университет информатики и радиоэлектроники*
*П. Бровки, 6, Минск, 220013, Беларусь*

*Поступила в редакцию 6 отября 2006*

Рассматривается проблема расстановки ударений в русском тексте как одна из проблем компьютерной обработки текста, предлагается ее решение для русского языка с помощью алгоритма автоматизированного морфологического анализа. Описывается алгоритм и структура словаря, процесс создания словарей для анализа и синтеза русских словоформ.

*Ключевые слова:* автоматическая обработка текста, русский язык, автоматизированный грамматический анализ, автоматизированный грамматический синтез, словарь.

## Введение

В русском тексте ударение является разноместным, т.е. оно не прикреплено к какому-либо определенному слогу или морфологической части слова. Ударение может падать на разные части слова и на разные слоги. Разноместность ударения является важным средством различения слов.

Правила расстановки ударений в текстах на русском языке нетривиальны и вызывают известные трудности в изучении и применении. Задача эффективной автоматической расстановки ударений относится к числу нетрадиционных, однако в настоящее время она возникает в некоторых областях обработки текстов на естественном русском языке — при синтезе русской речи, в издательском деле при допечатной подготовке изданий.

Наиболее очевидным решением проблемы является использование словаря всех возможных словоформ русского языка с указанием для каждой формы позиции ударения [1]. Достоинством этого подхода является его простота. Недостатком является большая трудоемкость построения словаря всех словоформ. Так как русский язык обладает развитым словоизменением, количество словоформ для каждого слова достаточно велико. Очевидно, что построение всех словоформ для значительного количества слов русского языка является трудоемкой задачей

Другим решением проблемы является расстановка ударения по определенным правилам на основе его грамматических характеристик. Грамматические характеристики выясняются в ходе морфологического анализа.

## Морфологический анализ русских слов

Задача морфологического анализа формулируется следующим образом: определить, при каких характеристиках и от каких слов может образовываться данная словоформа.

Для морфологического анализа О.С. Кулагиной были разработаны алгоритм и структура словаря [2, 3], которые позволили эффективно проводить анализ словоформ. В этом словаре слова разбиты на пять морфологических классов [2]. Каждый класс характеризуется своими

признаками и своим алгоритмом анализа. К первому классу относятся слова субстантивного склонения, ко второму — слова адъективного склонения, к третьему — глаголы, к четвертому — числительные. Так как невозможно ограничить многообразие морфологических форм русского языка какими-либо правилами, кроме указанных четырех морфологических классов выделен еще один класс для неизменяемых слов или таких форм изменяемых слов, которые не вкладываются в общую парадигму своего класса и являются исключениями. Эти формы названы особыми.

Внутри каждого морфологического класса, кроме последнего, слова разбиты по признаку формального сходства словоизменения на словоизменительные типы, соответствующие типам склонения и спряжения, а в классе глаголов, как в самом сложном, выделено несколько типов окончаний и суффиксов. Каждый тип, приписанный слову, является ссылкой на соответствующую таблицу, содержащую набор флексий, а при каждой флексии хранится набор морфологических признаков, при которых эта флексия может встречаться в словоформах данного слова.

Словарь состоит из словарных основ, при которых хранятся признаки — словоизменительные типы, типы окончаний и суффиксов. Во время анализа в словаре ищется основа, максимально вкладывающаяся слева в словоформу, и считываются признаки, хранящиеся при этой основе. После этого от словоформы отсекается основа, а в оставшейся части среди наборов флексий, определяемых признаками, ищутся флексии, входящие в словоформу. Морфологические характеристики, хранящиеся при флексиях в наборе, являются результатом морфологического анализа словоформы.

Этот алгоритм с некоторыми дополнительными признаками позволяет очень эффективно и с большой точностью (с точностью до морфологической омонимии) проводить анализ словоформы.

Парадигма слов 1-го класса насчитывает 12 позиций и может быть представлена 1–3 словарными статьями. В словаре в качестве словарной информации хранится только словоизменительный тип слова, всего 110 типов. Разбиение слов субстантивного склонения сделано таким образом, чтобы слова одного типа имели одинаковые наборы окончаний при одинаковых грамматических признаках.

Анализ проводится следующим образом. Для входной словоформы ищется в словаре статья, заголовок которой максимально вкладывается слева в словоформу. Если она найдена и если в ней словоизменительный класс равен 1, то считываем словоизменительный тип из словарной статьи. Окончание, полученное отсечением от входной цепочки букв словарной основы, ищется в строке таблицы окончаний существительных, номер строки определяется словоизменительным типом. При окончании хранится набор грамматических признаков, который и является результатом анализа.

Слова 2-го словоизменительного типа имеют парадигму, состоящую из 29 позиций — 24 для полных форм, 4 для кратких и 1 — сравнительная степень, превосходная степень рассматривается как отдельное прилагательное, не имеющее степеней сравнения. В словаре хранятся словоизменительный тип и тип чередования, вырабатываемые в результате анализа признаки — род, число, падеж, краткость, степень сравнения и возвратность.

Анализ проходит следующим образом. После нахождения словарной статьи и отсечения основы проверяется вложение возвратной частицы *ся* справа в словоформу: если остаток заканчивается на *ся*, то частица отбрасывается и запоминается факт возвратности. Затем проверяется тип чередования, если он равен 1 и остаток начинается с *н*, то *н* отбрасывается. Оставшаяся часть ищется в таблице окончаний прилагательных. При найденном окончании хранится набор характеристик, которые и являются результатом анализа. При необходимости результат уточняется с помощью характеристик, указанных в словарной статье.

В 3-й класс входят глаголы и глагольные формы. Парадигма глаголов велика — она насчитывает около 185 форм в общем случае. В словарной статье глагола хранятся следующие признаки: тип чередования, типы окончаний, типы суффиксов и 2 специальных признака. Вырабатываемые грамматические характеристики включают основные характеристики — времени, вида, залога и возвратности, а также дополнительные — лицо и число (для настоящего и будущего времени), род и число (для прошедшего времени обоих видов), число, падеж и краткость (для причастий).

Любая глагольная форм представима в виде конкатенации буквенных цепочек:

$$\rho = \chi_1 \oplus \chi_2 \oplus \chi_3 \oplus \chi_4,$$

где $\chi_1$ и $\chi_2$ — 1-й и 2-й элементы чередования; $\chi_3$ допускает две формы — или окончание глагола, или конкатенацию суффикса и окончания прилагательного; $\chi_4$ — возвратная частица.

На месте первого элемента чередования может стоять одна из буквенных цепочек $\chi_{1,1}$ или $\chi_{1,2}$, причем только после $\chi_{1,2}$ может стоять $\chi_2$, который тоже допускает одну из двух форм — $\chi_{2,1}$ и $\chi_{2,2}$. $\chi_{1,1}$ и $\chi_{2,1}$ не могут быть пустыми цепочками. Каждый элемент разложения остатка имеет так называемую категорию — указание на то, в каких глагольных формах может участвовать данный элемент.

Анализ глагольной формы состоит в построении указанного разбиения. От словоформы отсекается словарная основа, получается рабочий остаток. Если есть чередования (рассматривается тип чередования из словарной статьи), то по его типу устанавливается, что может стоять на месте первого и второго элементов чередования. Если буквенная цепочка $\chi_{1,1}$ вкладывается слева в рабочий остаток, то отсекаем ее, новый рабочий остаток и запоминаем грамматические признаки, соответствующие элементу остатка. Если $\chi_{1,1}$ не вложился, проверяем вложение $\chi_{1,2}$. Он должен вложиться, иначе необходимо искать другую словарную основу. Итак, если $\chi_{1,2}$ вложился в рабочий остаток, то отделяем его, сохраняем грамматические признаки и повторяем процесс с $\chi_2$, после чего получаем новый рабочий остаток. Далее проверяем вложение частицы *ся* в полученный остаток и при необходимости отсекаем его и фиксируем характеристику возвратности. Теперь рабочий остаток представляет собой или окончание глагола, или суффикс глагола и окончание прилагательного. Рассматриваем остаток как окончание глагола, т. е. пытаемся отыскать его в соответствующих таблицах (номер строки таблицы, в которой ведется поиск, определяется соответствующим типом окончания). После того, как окончание глагола найдено, проверяется наличие специальных признаков, с помощью которых корректируется полученный результат. Если окончание глагола не было найдено в таблицах окончаний, то рассматриваем рабочий остаток как суффикс глагола и окончание прилагательного. Если один из суффиксов глагола вкладывается слева в рабочий остаток, то он отсекается. Оставшийся остаток является окончанием прилагательного, его необходимо отыскать среди окончаний прилагательных и определить число, падеж и род причастия.

### Определение позиции ударения в словоформе

При создании словаря О.С. Кулагина оптимизировала для автоматизированного анализа словоформ результаты, опубликованные А.А. Зализняком в "Грамматическом словаре русского языка" [4]. Грамматический словарь создан для синтеза словоформ с участием человека, что обусловило менее строгое разбиение слов на грамматические разряды [4] по признаку сходства словоизменительных парадигм.

Основным делением слов в словаре [4] является деление на грамматические разряды. Грамматический разряд — это совокупность слов, у которых набор клеток, образующих парадигму, одинаков, т.е. одинаково число клеток и их названия. Деление на разряды тесно связано с делением на части речи, но не совпадает с ним. Все разряды снабжены достаточно большим количеством поправок и помет, корректирующих или дополняющих построенную парадигму. Так как существительные, прилагательные, числительные и местоимения имеют сходную парадигму (изменение по падежу, числу и роду — кроме существительных), то в "Грамматическом словаре" для них используется единая структура обозначений. Парадигма глаголов сильно отличается от парадигмы имен, для них используется своя система обозначений.

Считается, что существительные имеют субстантивное склонение, прилагательные — адъективное, местоимения — местоименное, а числительные — склонение числительных. При наличии у слова отклонения от этого правила оно отдельно указывается в словарной статье [4]. Для собирательных и количественных числительных в словарной статье непосредственно указана их парадигма, так как их число невелико, парадигма невелика и их склонение не подчиняется каким-либо общим правилам.

Глаголы представляют собой отдельную категорию. Основное деление глаголов внутри их разрядов — по типу спряжения. Каждый тип спряжения характеризуется определенным спо-

собом построения трех основных форм глагольной парадигмы. В словарной статье глагола хранятся признаки, характеризующие его вид, переходность, безличность, многократность, а также тип спряжения, пометы, указывающие на наличие беглой гласной или различные чередования. Если у глагола есть какие-либо дополнительные особенности спряжения, то они также указываются с помощью специальных признаков. Как и у слов других частей речи, словарная статья глагола может содержать несколько вариантов значений одной и той же пометы, пояснения, ограничения лексической сочетаемости, особые формы во фразеологизмах, а также схему ударения глагола.

Грамматический словарь позволяет определять позицию ударения в слове. Для этого автор для каждого склонения указывает основные схемы ударений и некоторые отклонения от них. Кроме того, словарь содержит сведения о переносе ударения на предлог в словосочетаниях. Для форм с ударением на окончании, которое является неслоговым, указаны правила переноса ударения на основу. Схема ударения позволяет определить, на какой слог падает ударение в конкретной форме слова, при условии, что известно ударение начальной формы слова. Так, для всех слов, кроме глаголов, указаны 7 (a–f) схем ударения: a — неподвижное ударение на основе, b — неподвижное ударение на окончании, c–f — разные виды подвижного ударения. Для слов адъективного склонения указаны две схемы ударения — для полных (a–b — только неподвижное ударение) и для кратких (a–c) форм. Чтобы указать основные типы колебания ударения в кратких формах, используются обозначения (a'–c', c''). Для глаголов также указаны две схемы ударения — для настоящего (a–c, c') и для прошедшего (a–c, c') времени, кроме глаголов *дать* и *взять* — у них представлена особая схема ударения. Для причастий действительного и страдательного залога настоящего и прошедшего времени и деепричастий настоящего и прошедшего времени отдельно указаны схемы ударения. Все глаголы в каждой из перечисленных форм имеют одинаковую схему ударения.

## Построение морфологического словаря

Хотя О.С. Кулагиной был предложен очень эффективные алгоритм и структура словаря для анализа словоформ, отсутствие в свободном доступе данных для этого словаря ограничивало использование этого алгоритма. В то же время грамматический словарь русского языка доступен в текстовом виде в сети Internet.

После реализации алгоритма синтаксического анализа необходимо было наполнить словарь, предназначенный для анализа, словарными основами, снабженными необходимыми признаками. Для этого существовало несколько способов. Первый заключался в заполнении словаря вручную, просклоняв каждое слово и определив его признаки в соответствии с таблицами. Этот способ требовал значительных усилий и филологической квалификации для построения всех возможных форм слов, вносимых в словарь.

В то же время грамматический словарь [4] позволяет построить все возможные словоформы для слов на основе признаков, которые были приписаны каждому слову. Автором была предпринята попытка найти соответствие между наборами признаков словаря для анализа и грамматического словаря. Оказалось, что не существует однозначного соответствия между наборами признаков двух словарей. Было принято решение реализовать алгоритм морфологического синтеза на основе грамматического словаря, автоматически построить все возможные формы каждого слова и найти, какие наборы флексий могут входить в словоформы каждого слова, т.е. определить значения признаков словарной статьи.

После того, как был реализован алгоритм синтеза словоформ, оказалось, что доступный в текстовом виде "Грамматический словарь" мало пригоден для автоматизированного синтеза, так как он ориентирован на восприятие человеком и при формировании словоформ на конечном этапе алгоритм опирается на языковой опыт человека. Кроме того, словарные статьи грамматического словаря в некоторых случаях сформированы достаточно свободно, что затрудняет его использование для автоматического анализа. Поэтому словарь был преобразован в удобный для автоматической обработки формат. После этого стало возможным формирование словарных статей словаря для морфологического анализа.

В настоящий момент завершено преобразование статей "Грамматического словаря". Наибольшие трудности при этом были вызваны глаголами, так как их парадигма очень велика

и для них надо определить 9 признаков, тогда как для остальных словоизменительных типов необходимо определить 1–2 признака. Кроме того, в грамматическом словаре члены видовой пары считаются разными глаголами, а в словаре для анализа они считаются одним глаголом, поэтому их необходимо выявить и объединить.

Таким образом, в результате преобразования статей грамматического словаря сформированы словарные статьи словаря для эффективного морфологического анализа. Кроме того, в качестве сопутствующей задачи был реализован алгоритм синтеза русских словоформ, а "Грамматический словарь" был переведен в более формализованную форму, что позволяет эффективно строить формы русских слов. Оба словаря имеют объем около 100 000 слов и охватывают практически все слова русского языка.

При текущей реализации скорость анализа составляет в среднем около 64 000 слов в секунду. Скорость анализа глаголов и глагольных форм составляет около 56 000 слов в секунду, остальных слов — от 93 000 (для прилагательных) до 105 000 (для существительных) слов в секунду. Различие в скорости анализа объясняется различной сложностью алгоритмов разбора. Скорость расстановки ударений несколько меньше скорости анализа, что объясняется наличием нескольких дополнительных шагов. Текущая реализация алгоритма выполнена на интерпретируемом языке программирования java, тесты выполнялись на машине с частотой процессора 3,2 ГГц (Pentium IV) и объемом оперативной памяти 1 Гб.

С помощью построенного аппарата можно эффективно решать проблему расстановки ударений в русских словоформах. Кроме того, реализованные словари и алгоритмы можно использовать в задачах автоматического перевода, извлечения информации из текста, автоматического индексирования баз данных в информационно-поисковых системах, сжатия текстовых баз данных, проверки грамматической правильности текста, создания электронных словарей и обучающих систем, проведения лингвистических исследований.

## AUTOMATED RUSSIAN TEXT PROCESSING

P.A. VEINIK

**Abstract**

The allolog analysis-based solution of the problem of accenting in Russian texts is considered. The morphological analysis and synthesis algorithms are described. Some problems of creating vocabularies for analysis and synthesis are discussed. The effectiveness of analysis algorithm is measured and presented.

**Литература**

1. *Вейник П.А.*// Изв. Белорус. инж. акад. 2002. № 1/2. С. 39–41.
2. *Кулагина О.С.*// Морфологический анализ русских именных словоформ. М., 1986.
3. *Кулагина О.С.*// Морфологический анализ русских глаголов. М., 1986.
4. *Зализняк А.А.*// Грамматический словарь русского языка — словоизменение. М., 1980.