Декабрь, № 8. Т. 1 **Физико-математические науки** **2012**

УДК 519.765

АНДРЕЙ АЛЕКСАНДРОВИЧ КОТОВ
кандидат филологических наук, доцент кафедры русского языка филологического факультета, Петрозаводский государственный университет (Петрозаводск, Российская Федерация)
andrewcot@onego.ru

МИХАИЛ ЮРЬЕВИЧ НЕКРАСОВ
преподаватель кафедры теории вероятностей и анализа данных математического факультета, Петрозаводский государственный университет (Петрозаводск, Российская Федерация)
nekrassov.mikhail@gmail.com

АЛЕКСЕЙ ВЛАДИМИРОВИЧ СЕДОВ
преподаватель кафедры теории вероятностей и анализа данных математического факультета, Петрозаводский государственный университет (Петрозаводск, Российская Федерация)
sedov_a@mail.ru

АЛЕКСАНДР АЛЕКСАНДРОВИЧ РОГОВ
доктор технических наук, профессор, заведующий кафедрой теории вероятностей и анализа данных математического факультета, Петрозаводский государственный университет (Петрозаводск, Российская Федерация)
rogov@psu.karelia.ru

ИНФОРМАЦИОННАЯ СИСТЕМА ДЛЯ СОЗДАНИЯ РАЗМЕЧЕННЫХ КОРПУСОВ МАЛОЙ РАЗМЕРНОСТИ*

Описывается информационная система, позволяющая создавать грамматически и синтаксически размеченные корпуса. Приводится модель, на основе которой построена система. Содержится краткий анализ существующих русскоязычных размеченных корпусов текстов и их особенностей, сравнение этих корпусов с корпусами, созданными при помощи описываемой системы.

Ключевые слова: размеченные корпуса, система грамматической разметки, система синтаксической разметки, проблемы омонимии

ЛИНГВИСТИЧЕСКИЕ КОРПУСА

Корпусная лингвистика – молодое направление прикладной лингвистики. Являясь разделом языкознания, она занимается выявлением закономерностей функционирования языка через его анализ и изучение с помощью лингвистического корпуса. Корпусная лингвистика включает в себя два аспекта:

- создание и разметка (аннотирование) корпусов текстов и разработка средств поиска по ним,
- экспериментальные исследования на базе корпусов [7].

Различают корпуса с морфологической (грамматической), синтаксической, акцентной и прочими разметками. Зачастую при создании корпуса используется смешанная разметка.

Перечислим основные корпуса русских текстов, существующие на данный момент:

1. Национальный корпус русского языка (НКРЯ) (http://www.ruscorpora.ru/) – наиболее крупный и известный корпус русских текстов.

2. Хельсинкский аннотированный корпус русских текстов ХАНКО (http://www.slav.helsinki.fi/hanco/ index.html).

3. Корпус русского литературного языка (http://www.narusco.ru/).

4. Открытый корпус OpenCorpora (http://www.opencorpora.org/) – находящийся в разработке корпус с открытым кодом.

5. Корпус русских публицистических текстов второй половины XIX века СМАЛТ (http://smalt.karelia.ru/corpus/index.phtml), разработанный в ПетрГУ.

Здесь перечислены далеко не все существующие корпуса русского языка. НКРЯ, например, состоит из целого ряда более мелких специализированных корпусов.

Сравнение корпусов

Все рассматриваемые корпуса содержат морфологическую (или грамматическую) разметку, хотя сами наборы морфологических признаков отличны. Синтаксическая разметка не присутствует в корпусе русского литературного языка, в OpenCorpora она заявлена, но доступа к ней пока нет. В тех же корпусах, где есть синтаксическая разметка, она представлена различными подходами. Синтаксическая разметка ХАНКО опирается на общепринятые классификации традиционной описательной грамматики – в рамках не раз подвергавшейся обоснованной критике теории членов предложения. Синтаксическая разметка НКРЯ построена на классификации,

доступной узкому кругу специалистов и требующей детального предварительного знакомства. Она представлена в терминах деревьев зависимостей и синтаксических отношений, принятых в теории «Смысл-Текст».

В корпусе СМАЛТ в основу синтаксической разметки положена идея структурной схемы предложения, восходящая к трудам представителей Пражского лингвистического кружка и получившая развитие на почве русского языка благодаря работам Н. Ю. Шведовой и ее последователей [1], [5]. Синтаксическая разметка – это следующий этап обработки корпуса текстов (после морфологической), она позволяет решать потенциальному адресату (в первую очередь лингвисту) целый спектр различных научно-исследовательских задач как в сфере русского, так и общего синтаксиса. В частности, на ее основе можно судить о частотном соотношении двух- и однокомпонентных предикативных структур в русском языке, о формальных способах выражения предикативных отношений, о развитии синтаксического строя русского языка. Подобный анализ позволяет объективировать и упорядочить, насколько это возможно, систему разметки. Отметим, что все пути синтаксического аннотирования в целом решают разные задачи, имеют свои достоинства и недостатки и представляют интерес для различного рода лингвистических исследований.

Если сравнивать представительность корпусов, то НКРЯ отличается от остальных корпусов сбалансированностью и многообразием текстов. Однако это достоинство НКРЯ порождает и главный недостаток крупных корпусов – слабую точность разметки. Некоторые из проектов, в рамках которых создавался НКРЯ, предусматривали краткие сроки разработки и автоматическую разметку. В связи с этим процент брака и неточностей в этой разметке существенен. Наибольший процент неточностей в автоматически размеченных грамматических корпусах возникает из-за проблемы *омонимии*[1]. В изначальном варианте НКРЯ омонимия снималась вручную, и размер корпуса был невелик [6]. Однако потом стало происходить значительное расширение корпуса и во многих случаях переход на автоматическую разметку, оставляющую омонимию. При автоматической разметке, как правило, в случае омонимии слову ставятся в соответствие все возможные наборы параметров. Сейчас в НКРЯ подавляющая часть корпуса хранится без снятия омонимии. Для СМАЛТ было принято решение выбрать основным фактором создания корпуса не большой размер, а высокую точность и однозначность разметки. Высокая точность разметки достигается при использовании специалистов-филологов, часть работы должна проделываться ими вручную довольно тщательно.

Учитывая это, при разработке универсальной программы по созданию размеченных корпусов небольшого размера было решено разработать специальные программы для грамматической и синтаксической разметки текстов, с которыми должны работать специалисты. Программы должны удовлетворять следующим особенностям.

- Пользователь должен иметь возможность настроить программу для определения / изменения грамматических или синтаксических параметров слова или предложения.
- Разбор должен быть максимально удобен, прост и быстр.

В основе работы программы грамматической разметки корпуса СМАЛТ лежит универсальная модель построения корпуса с грамматической разметкой.

МОДЕЛЬ ПОСТРОЕНИЯ КОРПУСА С ГРАММАТИЧЕСКОЙ РАЗМЕТКОЙ

Корпус представляет собой набор текстов $Texts = \{T_1, T_2, ..., T_n\}$. В общем случае каждый текст может быть представлен в виде группированного набора слов $T_i = \{w_{ij}\}$ (j – номер слова в тексте). Для удобства навигации по корпусу лучше вводить дополнительные структурные единицы текста. Далее будем обозначать отдельное слово текста, как w. В общем случае *posAtr:* $w \rightarrow Pos \in Positions \subset N^k$. $Positions = (Pos_1, Pos_2, ..., Pos_k)$. Здесь для слова w задается k его координат в тексте. Примерами дополнительных структурных единиц текста могут быть: главы, параграфы, абзацы, предложения. Модель предусматривает объединение мельчайших структурных единиц (слов) в словосочетания. Каждое словосочетание представимо в виде некоторого подмножества из слов предложения. $SS_j = \{w_{i1}, w_{i2},..., w_{ik}\} \subset S$, причем $\bigcup_j SS_j = S$ (отдельные слова в модели можно тоже принимать за словосочетания). Каждое слово может входить только в одно словосочетание, поэтому пересечение любых двух словосочетаний всегда пусто: $SS_k \cap SS_l = \varnothing$, при $k \neq l$.

Кроме позиции для каждого слова необходимо определять набор его морфологических параметров. Для существительного это могут быть: падеж, число, род, одушевленность / неодушевленность. Сами морфологические параметры представляют собой еще один вектор: *intAtr:* $w \rightarrow intParams \in N^k$: $intParams = (Param_1, Parsm_2, ..., Param_k)$, где k – количество грамматических признаков слова. Морфологические признаки закодированы натуральными числами. Расшифровка этих признаков хранится в отдельном конфигурационном файле и при необходимости может быть изменена администратором программы или даже самим пользователем.

Было решено хранить морфологические признаки в виде дерева. Структуру дерева можно описать следующим образом: $MorfAtrs:G=<V,E>$, причем $V = MorfParams \cup MorfItems$, а $\forall(i, j) \in E$, либо $i \in MorfParams$, $j \in MorfItems$, либо $i \in MorfItems$, $j \in MorfParams$. Иными словами, различают множество морфологических

параметров *MorfParams* и значений этих параметров *MorfItems*. Ребро $(i, j) \in E$, $i \in MorfParams$, $j \in MorfItems$ характеризует принадлежность значения j параметру i.

Для каждого слова имеется возможность определять ряд дополнительных строковых параметров. Причем для каждого корпуса они могут быть индивидуальны: нет необходимости их встраивать для всех корпусов. При создании программы универсальной разметки необходим механизм инициализации произвольного количества таких параметров. Формально $StrAtr: w \rightarrow strParams = \{strParam_i\}$, $i = \overline{1,l}$, структура же *strParams* задается пользователем в отдельном конфигурационном файле.

В результате каждому слову ставится в соответствие кортеж: $w \rightarrow <Pos, intParams, sParams>$ (позиция слова, морфологический разбор и некоторые строковые атрибуты слова), который можно считать грамматической разметкой данного слова.

ПРОГРАММА ГРАММАТИЧЕСКОЙ РАЗМЕТКИ

Процесс грамматического разбора текста делится на 2 этапа: этап преформатирования и этап разметки.

Этап преформатирования

На вход программы подается текст в кодировке Unicode, который необходимо разметить. При загрузке текста программа автоматически разбивает текст на структурные компоненты: главы, абзацы и предложения. Это первый этап разбора текста – этап преформатирования. Он проходит в полуавтоматическом режиме: программа на основании ряда эмпирических правил разбивает текст на структурные единицы, после чего пользователь проверяет правильность этого разбиения и при необходимости редактирует его.

Этап разметки

Следующим этапом является грамматический разбор текста. Пользователю выводится текст, где возможна только покомпонентная навигация. Это значит, что при грамматическом разборе пользователь может перемещаться по структурным компонентам текста, в данном случае по словам. В нижней части экрана выводится информация о разборе текущего слова, пользователь может изменить ее. Разбор представляет собой ряд выпадающих списков для морфологических признаков и поля для ввода необходимой текстовой информации. Слова, не размеченные ранее пользователем, выделяются красным цветом, размеченные слова выделяются зеленым.

В программе имеется возможность объединения нескольких слов в словосочетания, образующие составную часть.

Автоматическая грамматическая разметка

В системе грамматического разбора реализован алгоритм автоматической разметки. Система ищет (по написанию) текущее разбираемое слово в словаре, в случае удачного поиска пользователю выдается разбор найденного слова в качестве разбора текущего. Далее пользователь может изменить разбор или сохранить предложенный. Чтобы избежать постоянного обращения к базе данных в процессе разбора, автоматическая разметка всех неразмеченных слов текста производится до того, как пользователь приступил к проверке, по окончании этапа преформатирования.

Отметим основные проблемы автоматической разметки.

1. Некоторые слова остаются неразмеченными, так как не все слова есть в словаре. Стоит отметить, что с расширением словаря неизвестных слов становится все меньше.

2. С расширением словаря остро проявляются проблемы *омонимии*. Она заключается в том, что для некоторого слова в словаре может содержаться несколько вариантов его атрибуции, поэтому встает вопрос, как выбрать из них верный. Один из вариантов решения этой проблемы представлен в виде следующей модели. Рассматриваются триады: тройки подряд идущих слов ν_1, ν_2 и ν_3 с соответствующими им разборами a_1, a_2 и a_3. Порядок слов считается важным. Введем понятие вероятности разбора $P(a_i) = \frac{n(a_i, \nu_i)}{N(\nu_i)}$, где $n(a_i, \nu_i)$ – количество встречаемых в корпусе разборов a_i слова ν_i, а $N(\nu_i)$ – общее количество разборов слова, встречаемых в корпусе. Запишем: $P(a_1a_2a_3) = \alpha P(a_3)P(a_2 \mid a_3)P(a_1|a_2a_3) + (1 - \alpha)P(a_1)P(a_2 \mid a_1)P(a_3 \mid a_1a_2)$.

Основу формулы составляют два слагаемых, каждое из которых представляет собой формулу условной вероятности. Разбор a_3, на котором достигается максимум функции $P(a_3)$, принимается как разбор по умолчанию в случае омонимии. Весовой коэффициент α задает степень влияния каждого слагаемого на итоговую сумму. Если триады слов ν_1 ν_2 ν_3 ранее не встречались в корпусе в данном порядке, рассматривается уже двойка слов, для которой приводятся аналогичные результаты. Если двойка не встречалась, в качестве разбора a_i слова ν_i принимается $\arg\max P(a_i) = \frac{n(a_i, \nu_i)}{N(\nu_i)}$, то есть самый часто встречаемый в корпусе разбор слова ν_i.

Приведем общий алгоритм автоматической грамматической разметки в виде псевдокода:

Цикл по всем словам в тексте
  Количество вхождений текущего слова в словаре равно?
    Если 0 => слово разбирается с нуля
    Если 1 => найденный разбор – разбор по умолчанию
    Если больше 1, цикл по каждому разбору i
      Количество слов перед текущим в предложении?

Если 0 => $\alpha_1 = 1$, $k_1 = P(a_3)$
Если 1 => $\alpha_1 = 2$, $k_1 = P(a_3)P(a_2 \mid a_3)$
Если больше 1 => $\alpha_1 = 3$, $k_1 = \alpha P(a_3)P(a_2 \mid a_3)P(a_1 \mid a_2 a_3)$

Количество слов после текущего в предложении?
Если 0 => $\alpha_2 = 1$, $k_2 = P(a_1)$
Если 1 => $\alpha_2 = 2$, $k_2 = P(a_1)P(a_2 \mid a_1)$
Если больше 1 => $\alpha_2 = 3$, $k_2 = P(a_1)P(a_2 \mid a_1)P(a_3 \mid a_1 a_2)$

$\alpha = \alpha_1 / (\alpha_1 + \alpha_2)$
$P_i = \alpha\, k_1 + (1-a)\, k_2$
$P = max\ P_i$

МОДЕЛЬ ПОСТРОЕНИЯ КОРПУСА С СИНТАКСИЧЕСКОЙ РАЗМЕТКОЙ

Опишем модель построения синтаксического корпуса СМАЛТ. Корпус представляет собой набор текстов $Texts = \{T_1, T_2, ..., T_n\}$. В отличие от грамматического разбора, минимальной структурной единицей текста здесь, в зависимости от выбранного подхода к синтаксической разметке, могут выделяться не только слова, но также предложения, части предложений либо клаузы[2]. Рассмотрим вариант с разбиением текстов на клаузы.

Определим формально понятие клаузы. Для этого возьмем за основу предложение S, позиционированное в тексте. Данное предложение разбивается на части и может быть представлено как множество: $S = \{Part_1, Part_2, ..., Part_n\}$. В этом виде клаузы предложения представимы в виде некоторого подмножества из частей предложения. $Cl_j = \{Part_{i1}, Part_{i2}, ..., Part_{ik}\} \subset S$, причем $\bigcup_j Cl_j = S$ и, вообще говоря, пересечение клауз может быть непустым. Резюмируя это, отметим два основных момента: каждая клауза может быть разбита на несколько частей внутри одного предложения; некоторые части предложения могут входить сразу в несколько клауз. Разметка клаузы очень проста: $Cl_j \rightarrow Scheme \in N$

В результате была разработана модель разбора, при котором предложения текста могут разделяться на части, а эти части группируются в клаузы, причем важен порядок этой группировки. Например, предложение поделено на 4 части: $Part_1$, $Part_2$, $Part_3$, $Part_4$. Если объединяется $Part_1$ и $Part_3$, а затем $Part_3$ и $Part_4$, то получается единая клауза $Part_1$-$Part_3$-$Part_4$. Если объединяется $Part_1$ и $Part_3$, а затем $Part_1$ и $Part_4$, то организуется ветвление и в итоге получаются две клаузы $Part_1$-$Part_3$ и $Part_1$-$Part_4$. Текст представляется в виде графа, в котором части предложения являются вершинами, а связные комбинации вершин образуют клаузы.

ПРОГРАММА СИНТАКСИЧЕСКОГО РАЗБОРА

Программа для синтаксической разметки имеет структуру, схожую со структурой программы для грамматической разметки. Однако из-за описанных особенностей модели синтаксического разбора в программе существует и ряд отличий. Разбор текста также делится на 2 этапа: этап преформатирования и этап синтаксической разметки.

Этап преформатирования аналогичен этапу преформатирования грамматического разбора. Отметим, что разбиение предложения на части и объединение частей в клаузы для удобства пользователей проводится на этапе разметки.

На *этапе синтаксической разметки* пользователь сопоставляет каждой клаузе текста одну из синтаксических схем путем выбора одной из них из списка, предложенного на нижней панели. Пользователь имеет возможность редактировать разметку текста и разбивать предложения на части. Отдельно выделяется текущая клауза, которая размечается в данный момент, и выводится контекст этой клаузы. Покомпонентная навигация программы позволяет пользователю передвигаться по частям предложения. Основная сложность для разметчика кроется в правильном разбиении предложения на части и сцеплении этих частей в клаузы.

ПРИМЕНЕНИЕ ИНФОРМАЦИОННОЙ СИСТЕМЫ

На базе описанной выше информационной системы были построены несколько корпусов: корпус СМАЛТ и корпус финноязычных текстов.

Подробнее про создание корпуса СМАЛТ написано в [4] и [2]. Основу корпуса составляют публицистические тексты разной тематики из петербургских журналов XIX века в дореволюционной графике, при этом все слова ретранслируются и в современную графику. В корпусе присутствует метаразметка текстов по следующим параметрам: автор текста, журнал, дата написания. Корпус на данный момент состоит из 101 текста, из них 74 содержат грамматическую разметку и 95 – синтаксическую. В грамматически атрибутированных текстах корпуса содержится 169 136 слов. Количество лексем в грамматической части корпуса равно 21 630, словоформ – 48 320. В синтаксической части корпуса размечено 47 336 клауз. Строковые параметры корпуса СМАЛТ – начальная форма, современное написание и современное написание начальной формы.

Подробнее про корпус финноязычных текстов написано в [3]. Корпус состоит из 107 статей газеты «Karjalan Sanomat», содержащих грамматическую разметку. Строковые грамматические параметры корпуса финноязычных текстов – начальная форма и перевод слова.

ПРЕДОСТАВЛЕНИЕ ДОСТУПА К КОРПУСАМ

Для доступа к корпусам было решено использовать веб-ресурс, предоставляющий возможности просмотра информации о корпусе, списке публикаций, авторах, а также поиска интересующей информации в корпусе с возможностью выбора только необходимых текстов (формиро-

вание собственного подкорпуса). Доступ к корпусу СМАЛТ расположен по адресу: http://smalt.karelia.ru/corpus/index.phtml. Корпус финноязычных текстов находится в тестовом режиме, доступ к нему пока не открыт.

На данном ресурсе пользователь может ознакомиться с текстами корпусов, осуществить поиск по заданным параметрам.

ЗАКЛЮЧЕНИЕ

Проанализировав различные существующие на данный момент корпусы и сравнив их с корпусом СМАЛТ, разрабатываемым в ПетрГУ, мы пришли к следующим выводам.

- Корпус СМАЛТ не претендует на представительность и полноту НКРЯ, однако он изначально планировался как корпус со специфическим содержанием и точной разметкой.
- Синтаксический подкорпус СМАЛТ построен на принципе разметки, не применяемом в других крупных русскоязычных корпусах, и поэтому может быть интересен определенному кругу исследователей.
- Универсальность моделей и программ, используемых для разметки корпуса СМАЛТ и корпуса финноязычных текстов, позволяет использовать их для создания различных корпусов, в том числе и для специфических языков.
- Наличие как оригинальной графики, так и современного написания слова и начальной формы в корпусе СМАЛТ делает его интересным для исследователей дореволюционного русского языка и его связи с современным языком.

Размеченный корпус СМАЛТ может быть использован при научных изысканиях в области истории языка, грамматики, лексикографии, а также при изучении соответствующих курсов студентами филологических специальностей. Кроме того, он может быть востребован специалистами по литературе XIX века.

* Работа выполнена при поддержке Программы стратегического развития (ПСР) ПетрГУ в рамках реализации комплекса мероприятий по развитию научно-исследовательской деятельности на 2012–2016 гг.

ПРИМЕЧАНИЯ

[1] Омонимия – в нашем случае это ситуация, при которой одинаково написанные слова имеют различные грамматические разборы.

[2] Клауза – элементарное предложение в составе сложного, вершиной которого является глагол.

СПИСОК ЛИТЕРАТУРЫ

1. Грамматика современного русского литературного языка / Под ред. Н. Ю. Шведовой. М.: Наука, 1970.
2. Котов А. А., Гурин Г. Б., Седов А. В., Некрасов М. Ю., Сидоров Ю. В., Рогов А. А. Особенности создания электронного ресурса «материалы к синтаксическому словарю» // Российский научный электронный журнал «Электронные библиотеки», Том 13 – Выпуск 2, 2010 г. [Электронный ресурс]. Режим доступа: http://elbib.ru/index.phtml?page=elbib/rus/journal/2010/part2/KGSNSR
3. Петрова А. А., Ленина А. А. Результаты совместного проекта РГНФ – АФ «перевод с финского на русский и с русского на финский в поликультурном мире»: создание лингвистического корпуса финноязычных текстов газеты «Karjalan sanomat» и его перспективы [Электронный ресурс]. Режим доступа: http://www.petrsu.ru/Faculties/Balfin/AAPetrova_2011.html
4. Рогов А. А., Гурин Г. Б., Котов А. А. Некоторые особенности грамматически размеченного корпуса по русской публицистике второй половины XIX века // Труды международной конференции «Корпусная лингвистика-2008». СПб., 2008. С. 326–333.
5. Русская грамматика / Под ред. Н. Ю. Шведовой. М.: Наука, 1980. Т. 1, 2.
6. Сичинава Д. В. Национальный корпус русского языка: очерк предыстории // Национальный корпус русского языка: 2003–2005. М.: Индрик, 2005. С. 21–30.
7. Энциклопедическая статья «Корпусная лингвистика» // Электронный портал «Фонд знаний Ломоносов» [Электронный ресурс]. Режим доступа: http://www.lomonosov-fund.ru/enc/ru/encyclopedia:01210:article