А.А. Рогов, Г.Б. Гурин, А.А. Котов

НЕКОТОРЫЕ ОСОБЕННОСТИ ГРАММАТИЧЕСКИ РАЗМЕЧЕННОГО КОРПУСА ПО РУССКОЙ ПУБЛИЦИСТИКЕ ВТОРОЙ ПОЛОВИНЫ XIX ВЕКА

Введение

В настоящее время на смену таким традиционным методам получения языковых данных, как интроспекция, сбор текстового материала, эксперимент, опрос, приходит ***корпусный метод***; создание лингвистических корпусов текстов осознается как одна из актуальных задач современного языкознания[1]. Корпусы активно используются в практике составления словарей, в проведении разнообразных исследований языка. Отечественная лингвистика несколько отстает в этом отношении. Однако в последнее время в появилось немало интересных проектов такого рода, самый масштабный из них – Национальный корпус русского языка.

Программный комплекс «СМАЛТ»

В Петрозаводском государственном университете работы по компьютерной обработке текстов ведутся с 1995 года. Их результатом явилась разработка программного комплекса «Статистические методы анализа литературных текстов» (ПК «СМАЛТ»), имеющего в своей основе базу данных текстов, состоящую из публицистических статей разной тематической направленности

[1] См., например: *Волков С.Св., Захаров В.П.* Корпус текстов и исторический словарь // Русский язык конца XIX века: Проблемы изучения и лексикографического описания. СПб., 2004. С. 38–43; *Волков С.Св., Герд А.С., Гринбаум О.Н.* и др. Корпус текстов как особый тип лингвистической электронной библиотеки // Словарь русского языка XIX века. Проблемы. Исследования. Перспективы. СПб., 2003. С. 92–108.

из петербургских журналов XIX века «Время», «Эпоха», «Современник», «Гражданин» «Светоч», «Молва», «Библиотека для чтения», «Заря» в оригинальной орфографии. Проект был поддержан грантами РГНФ № 02-04-12015в, 05-04-12418в, 08-04-12105в (рук. А.А. Рогов). Адрес в Интернете: **http://smalt.karelia.ru**.

ПК «СМАЛТ» предоставляет несколько систем доступа к единой базе данных, хранящей синтаксические и морфологические разборы литературных произведений. Он состоит из базы данных, системы подготовки данных, системы контроля знаний учащихся, системы доступа к БД и экспертной системы по выявлению скрытых количественных характеристик. Для хранения базы данных используется СУБД Interbase 6.0. В качестве исходного источника данных для клиентского приложения используется текстовый файл в кодировке Unicode, что позволяет избежать проблем, связанных с использованием в отдельных текстах символов, специфичных как для отдельных языков, так и для орфографии разных периодов одного языка.

С помощью системы подготовки данных ПК «СМАЛТ» была проведена лемматизация и морфологическая разметка текстов, в настоящее время ведется работа по синтаксической аннотации предикативных клауз.

Обработка каждого текста в БД предполагает три стадии: преформатирование, грамматический анализ, синтаксический анализ. На этапе преформатирования выполняется автоматизированное разбиение исходного текста на единицы, среди которых выделяются часть (или раздел), абзац, предложение, слово. Полученное разбиение может быть откорректировано вручную. Важнейшим модулем ПК «СМАЛТ» является морфологически размеченный корпус текстов русской публицистики второй половины XIX века как самостоятельный продукт.

Состав и особенности морфологического корпуса

Большинство современных русскоязычных корпусов ориентированы на язык XX–XXI веков, тексты предшествующих периодов, в силу трудности их автоматической обработки, включаются в корпусы реже. Данный корпус является историческим, поскольку сформирован из оригинальных текстов русской публицистики 60–70-х годов XIX века.

Публицистические тексты обязательно включаются в состав современных лингвистических корпусов: именно в публицистике в силу ее определенной жанровой свободы и тесной связи с социально-культурной, политической и экономической жизнью социума полнее и ярче отражаются разнообразные языковые изменения, прослеживаются формирующиеся тенденции развития языка. Насколько нам известно, публицистические тексты эпохи второй половины XIX века до сих пор специально не привлекались в качестве особого объекта корпусной презентации.

В корпусе принципиально сохранены исконная графика текстов, а также все особенности дореформенной орфографии, как известно, неустойчивой, отличавшейся орфографическая и фономорфологическая вариативностью. В частности в корпусе сохраняются такие написания, как *очень-многіе, само-по-себе, до-сихъ-поръ, на-дняхъ, какъ-будто, ничемъ другъ-къ-другу необязанныхъ, студентскій миръ, взмахнутый, въ самомъ-достойнѢйшемъ, самоновѢйшій, низачто, само-малѢйшей, истинно-умные расположонъ, предстоитъ современем, состарѢлась, мущина, выростетъ, комунистъ, колосальный* и проч.

Сохранение этих особенностей в диахроническом корпусе представляется крайне полезным и даже необходимым не только при описании динамики норм правописания, но и при выявлении некоторых тенденций развития грамматической системы языка XIX века. При этом корпус принципиально адресован самому широкому кругу пользователей (лингвистам, в том числе истори-

кам языка, литературоведам, студентам, преподавателям и учащимся средней школы), в том числе и тем, кто незнаком с особенностями дореволюционной графики и орфографии. Поэтому реализован поиск слов по современной орфографии, позволяющий отыскивать, например, по лемме *мужчина* все орфографические варианты (*мужчина, мущина, мужщина*).

Словарь, наполнение которого происходит в процессе разбора, существенно ускоряет проведение морфологического анализа, а также позволяет рассматривать разные виды омонимии, возникающие в тексте. Формирование собственного словаря позволяет в перспективе работать с текстами на разных языках

Широкий круг потенциальных пользователей корпуса обусловил также особенности морфологической разметки, подачи материала и системы поиска.

Принципы морфологической разметки

Общеизвестно, что «представление в корпусе информации о морфологических формах и значениях (часть речи, род, падеж, вид…) является самостоятельной научной проблемой»[1]. Корпус опирается в основном на морфологическую модель, представленную в «Грамматическом словаре русского языка» А. А. Зализняка. Однако специфика корпуса языка XIX века, ориентированного на широкого пользователя, такова, что в некоторых случаях требовались особые решения. Для сохранения упорядоченности и единообразия разметки, в первую очередь частеречной, последовательно применялись рекомендации Малого академического словаря и Словаря С.И. Ожегова и Н.Ю. Шведовой[2].

---

[1] **www.ruscorpora.ru**

[2] *Словарь русского языка*: В 4 т. 2-е изд., испр. и доп. М., 1981–1984; *Ожегов С.И., Шведова Н.Ю.* Толковый словарь русского языка: около 80 000 слов и фразеологических выражений. 4-е изд., М., 2006.

Минусы этого решения вполне очевидны для разработчиков, однако принципиальной установкой было обеспечение доступности и простоты в использовании корпуса, что учитывалось при формировании системы грамматических параметров. В корпусе использованы 2 варианта морфологической разметки, основанной на системе традиционных морфологических понятий.

Разметка 1 опирается на следующий инвентарь частей речи: существительное, прилагательное, числительное, местоимение, глагол, причастие, деепричастие, наречие, предикатив, союз, предлог, модально-дискурсивное слово или частица, междометие, компонент идиомы, антропоним.

Предоставляется возможность поиска по значениям базовых морфологических категорий соответствующих частей речи.

Разметка 2 ориентирована на школьную традицию, включает дополнительные грамматические параметры: лексико-грамматические разряды существительных. прилагательных, числительных, местоимений, типы склонения и спряжения. Она предназначена для использования в образовательных целях, может рассматриваться как параллельный обучающий корпус, подобный тому, что реализован в рамках Национального корпуса русского языка.

Отличие «лексико-грамматического» и «формально-грамматического разбора» можно пояснить на примере. В словосочетании «*первый ученик*», при лексико-грамматическом анализе слово «*первый*» будет кодировано как прилагательное (в значении «лучший»), а при формально-грамматическом разборе как числительное. Формально-грамматический разбор обладает меньшей вариативностью, но и меньшей степенью субъективности. Заметим, что взаимнооднозначное соответствие между разборами удалось установить только в 90% случаев. Формализовать остальные 10% не удалось.

В настоящий момент в базе данных словаря с морфолого-семантический разбором находится более 40 000 лемм из текстов

общим объемом более 140 000 словоформ, с формально-грамматическим разбором – более 26 000 лемм из текстов общим объемом около 100 000 словоформ.

Подача материала и системы поиска

Реализация модулей доступа к БД системы производится с использованием языка PHP 4. Для обеспечения поддержки символов дореформенного алфавита все тексты произведений, словоформы хранятся в кодировке Юникод. Для отображения используется шрифт Palatino Linotype.

Для удобства работы и полноты информации корпус реализован в виде словаря с алфавитной системой построения. Реализовано несколько систем поиска по различным критериям: 1) по словам в старой орфографии; 2) по словам в современной орфографии; 3) по грамматическим признакам (с возможностью сохранения заданных параметров).

Кроме того, возможен поиск через Сводный список текстоформ: 1) алфавитный, 2) алфавитно-частотный (с указанием частотности по убывающей).

При использовании любого поиска пользователь получает информацию в следующей последовательности: 1) морфологический разбор (или множество морфологических разборов); 2) сведения об авторе и произведении, сведения о контексте с точностью до номера главы, параграфа и предложения; 3) контекст в пределах предложения; 4) расширенный контекст – полный оригинальный текст.

Проект синтаксически аннотированного корпуса

В настоящее время ведется работа по созданию синтаксической разметки текстов, уже включенных в состав морфологического корпуса. Синтаксическое аннотирование представляет не меньшую сложность, если учитывать разнообразие подходов и

синтаксических теорий. На наш взгляд, прежде всего необходимо сформулировать те базовые принципы, которыми следует руководствоваться в дальнейшем. Они могут быть различны.

Например, в одном из немногих корпусов русского языка с интегрированной синтаксической разметкой «ХАНКО» аннотация строится с опорой на базовый принцип доступности: «С потребностью в доступности корпуса связано то обстоятельство, что при его создании мы опираемся на устоявшиеся теоретические концепции, которые используются в известных лингвистических трудах и/или учебной литературе по русской грамматике»[1]. Таким образом, используется общепринятая в школьной (и отчасти вузовской) практике система описания традиционного синтаксиса, в рамках которого одно из центральных мест занимает учение о членах предложение, при этом различные аспекты его организации не разграничиваются.

Минусы этого решения, и в первую очередь **нечеткость и множественность характеристик** при разметке, признают сами авторы: «В силу нечеткости критериев выделения членов предложения ряд единиц получил множественную синтаксическую разметку. Такие единицы получают также двойную графическую разметку»[2].

В создаваемом корпусе в основу синтаксической разметки для положена идея структурной схемы в понимании Н. Ю. Шведовой и ее последователей, впервые отчетливо заявленная в «Грамматике современного русского литературного языка» (1970), позднее наиболее полно отраженная и развитая в «Русской грамматике» (1980)[3]. С одной стороны, это несколько су-

[1] **www.slav.helsinki.fi/hanco/**

[2] Там же.

[3] *Грамматика* современного русского литературного языка. М., 1970. С. 544–574; *Русская* грамматика. М., 1980. Т. 2. Синтаксис. С. 92–98, 237–386.

жает круг потенциальных пользователей, с другой – позволяет объективировать и упорядочить, насколько это возможно, систему разметки.

Создание полного списка структурных схем простого предложения (в корпусе размечаются предикативные клаузы) – отдельная научная проблема, не имеющая пока своего решения. На данный момент мы можем говорить о том, что в научном обороте существуют как минимум три списка структурных схем – различные как количественно, так и качественно: 1) список схем «Русской грамматики»; 2) список «минимальных схем» В.А. Белошапковой; 3) список схем Е.Н. Ширяева[1].

Е.Н. Ширяев на основе достаточно убедительного теоретического обоснования значительно переработал и дополнил исходный список свободных структурных схем «Русской грамматики». Именно он является на сегодняшний день наиболее полным и точным и взят за основу для дальнейшей переработки. Ее необходимость объясняется двумя причинами: во-первых, использование структурных схем для синтаксической разметки в корпусе имеет свою специфику, во-вторых, объективная ситуация изучения вопроса такова, что ни один из существующих списков структурных схем нельзя признать окончательно полным.

Таким образом, на выходе мы получим наиболее полный и сбалансированный список структурных схем простого предложения, который будет использован для синтаксической разметки текстов.

---

[1] См. подробнее в порядке перечисления в тексте: *Русская* грамматика. М., 1980. Т. 2. Синтаксис. С. 97; *Современный русский язык* / Под ред. В.А. Белошапковой. 2-е изд., испр. и доп. М., 1989. С. 637–644; *Современный русский язык* / Под общ. ред. Л.А. Новикова. 3-е изд. СПб., 2001. С. 632–640.