Программный комплекс «СМАЛТ»♣

© А.А. Рогов, Г.Б. Гурин, А.А. Котов, Ю.В. Сидоров, Т.Г. Суровцова

Петрозаводский государственный университет
rogov@psu.karelia.ru

Аннотация

В современном языкознании 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. Корпусы активно используются в практике составления словарей, в проведении разнообразных исследований языка.

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 из петербургских журналов XIX века в оригинальной орфографии.

1 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 [2, 4]. Корпусы активно используются в практике составления словарей, в проведении разнообразных исследований языка. Наиболее известны общепризнанные образцы корпусов: Британский национальный корпус, Мангеймский корпус немецкого языка, Чешский национальный корпус и др. Отечественная лингвистика несколько отстает в этом отношении. Достаточно вспомнить, что до недавнего времени существовал единственный русскоязычный корпусов, Упсальский корпус русских текстов, который был создан в 60-е гг. прошлого века вне России и, по оценке многих, во многом устарел. Однако в последнее время появилось немало интересных проектов такого рода, самый масштабный из них – Национальный корпус русского языка (www.ruscorpora.ru).

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 из петербургских журналов «Время», «Эпоха», «Современник», «Гражданин» «Светоч», «Молва», «Библиотека для чтения», «Заря» XIX века [1–3] в оригинальной орфографии. Проект был поддержан грантами РГНФ № 02-04-12015в, 05-04-12418в, 08-04-12105в (руководитель А. А. Рогов). Адрес в Интернете: http://smalt.karelia.ru.

2 Программный комплекс «СМАЛТ»

Программный комплекс «СМАЛТ» предоставляет несколько систем доступа к единой базе данных, хранящей синтаксические и морфологические разборы литературных произведений. Общая схема ПК «СМАЛТ» представлена на рис. 1. Он состоит из базы данных, системы подготовки данных, системы контроля знаний учащихся, системы доступа к БД и информационно-аналитической системы для анализа текстов. 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.

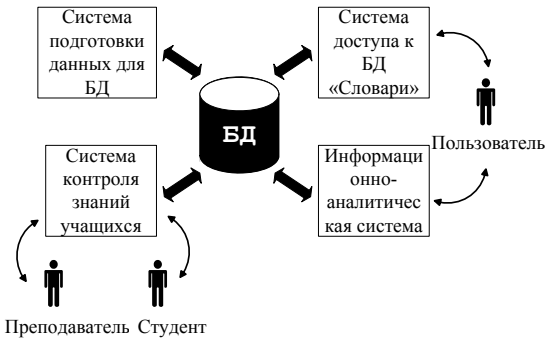

Рис. 1. Схема программного комплекса

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008, Дубна, Россия, 2008.

С помощью системы подготовки данных ПК «СМАЛТ» была проведена лемматизация и морфологическая разметка текстов, запланирована синтаксическая аннотация предикативных клауз. Обработка каждого текста в БД предполагает три стадии: преформатирование, грамматический анализ, синтаксический анализ (рис. 2). 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, который может рассматриваться как самостоятельный продукт.

Рис. 2. Система подготовки размеченных текстов

3 Состав и особенности корпуса, его адресаты

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 по понятным причинам: 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, представляющие богатейший языковой материал, 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, орфография той эпохи не была устойчивой, широко была распространена орфографическая и фономорфологическая вариативность, например: *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 может быть полезно исследователю, описывающему динамику норм правописания [см. 4, 6]. При этом корпус удобен для пользователя, незнакомого с особенностями дореволюционной графики и орфографии: в корпусе реализован поиск слов по современной орфографии, позволяющий отыскивать, например, по лемме *мужчина* все орфографические варианты (*мужчина, мущина, мужщина*).

Потенциально корпус адресован самому широкому кругу пользователей – всем, кто интересуется русским языком и русской литературой 19 столетия: профессиональным лингвистам, в том числе историкам языка, филологам-литературоведам, студентам, преподавателям средней школы и школьникам. Это обусловило некоторые особенности морфологической разметки, подачи материала и системы поиска.

4 Морфологическая разметка корпуса

Общеизвестно, что «представление в корпусе информации о морфологических формах и значениях (часть речи, род, падеж, вид…) является самостоятельной научной проблемой» [6], которая может решаться по-разному [4]. Данный корпус опирается в основном на морфологическую модель, представленную в «Грамматическом словаре русского языка» А. А. Зализняка (М., 1977; 4-е изд. М., 2003).

Однако специфика корпуса и тем более корпуса языка XIX века, ориентированного на широкого пользователя, такова, что в некоторых случаях требовались особые решения. Для сохранения упорядоченности и единообразия разметки, особенно при принятии решений в области частеречной разметки, последовательно применялись рекомендации «Словаря русского языка» (2-е изд. М., 1981–1984) и «Толкового словаря русского языка» С. И. Ожегова и Н. Ю. Шведовой. Несомненные минусы этого решения очевидны, однако принципиальной установкой разработчиков было обеспечение доступности и простоты в использовании корпуса. Все перечисленные критерии учитывались при формировании системы грамматических параметров, с помощью которой описывались словоформы.

В корпусе использована система морфологической разметки, использующая систему традиционных морфологических понятий, причем в двух вариантах – в виде формально-грамматической и морфолого-семантической разметки

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 – и предоставляет возможность поиска по значениям базовых морфологических категорий соответствующих частей речи.

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

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

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

В настоящий момент в базе данных словаря с морфолого-семантический разбором находится более 40 000 лемм из текстов общим объемом более 140 000 словоформ. В словаре с формально-грамматическим разбором находится более 26 000 лемм из текстов общим объемом около 100 000 словоформ.

5 Подача материала и системы поиска

Реализация модулей доступа к БД системы производится с использованием языка PHP 4. Общая схема доступа представлена на рис. 3. Пунктирными линиями обозначены модули, разработка которых незакончена. Для обеспечения поддержки символов дореволюционного русского алфавита все тексты произведений, словоформы хранятся в кодировке Unicode. 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 частоты встречаемости, отсортированнные по убыванию).

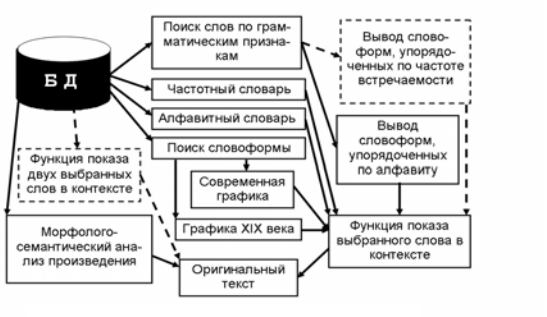

Рис. 3. Функции доступа к базе данных размеченных текстов «СМАЛТ»

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.

6 Дополнительные возможности ПК «СМАЛТ»

Предусмотрена возможность расширения корпуса электронных документов за счет данных, подготовленных в других информационных системах. Проведена работа по синхронизации данных ПК «СМАЛТ», и конкордансов В. И. Даля (http://elibrary.karelia.ru/dahl/user_new/index.php). Одна база данных реализована на платформе СУБД Interbase 6.0, другая – на платформе СУБД Oracle. Создано универсальное приложение, которое проводит синхронизацию данных.

Для пользователей, которые активно работают с ПК «СМАЛТ», в целях экономии денежных средств на Интернет-трафик, а также для пользователей, испытывающих проблемы с доступом к сети Интернет, была разработана локальная версия ПК «СМАЛТ», которую можно установить на локальные (персональные) рабочие станции пользователей. В локальную версию включена возможность работы с корпусом публицистических произведений. Для этого достаточно около 500 мегабайт свободного дискового пространства на компьютере и устройство для чтения компакт-дисков. Сам процесс установки достаточно прост даже для неспециалистов - инструкция по установке загружается автоматически и достаточно просто следовать пошаговым инструкциям, указанным в ней. Во время инсталляции на рабочую станцию устанавливается web-сервер Apache, серверный язык обработки скриптов PHP и система управления базами данных InterBase. Все вышеперечисленное программное обеспечение является свободно распространяемым и не требует покупки дополнительных лицензий.

Таким образом, на рабочие станции пользователей устанавливается программное обеспечение ПК «СМАЛТ», которое позволяет работать с корпусом произведений. В дальнейшем, для увеличения количества разобранных произведений пользователю достаточно будет обновить базу данных. Кроме этого в локальной версии ПК «СМАЛТ» предусмотрена функция выбора произведений, по которым строится словарь, что позволяет сделать его более сбалансированным. Адрес в Интернете: http://smalt.karelia.ru.

7 Описание информационно-аналитической системы для анализа текстов

Модуль статистической обработки, входящий в ПК «СМАЛТ», представляет собой информационно-аналитическую систему, которая включает в себя ряд общеизвестных методов, позволяющих проводить классификацию и кластеризацию текстов и групп текстов на основании определяемых лингвостатистических параметров, вычисляемых по тексту, например, распределению длины предложения или употреблению частей речи на заданных позициях предложения. Могут быть рассчитаны и количественные характеристики для произведения или групп произведений, например, количество употребления части речи, средняя длина предложения и т.п. Лингвостатистические параметры определяются на основании морфологических и синтаксических разборов литературных произведений получаемых с использованием корпуса произведений, поддерживаемого ПК «СМАЛТ». Модульная структура и использование архитектуры «клиент-сервер», позволяет обеспечить расширяемость, гибкое изменение конфигурации, отсутствие необходимости в установке.

Для проведения кластеризации и классификации текстов в информационно-аналитической системе предлагаются следующие группы методов:

- разбиение анализируемых текстов на группы с использованием методов кластерного анализа;
- разбиение анализируемых текстов на группы с использованием методов дискриминантного анализа;
- метод «сильного графа» для оценки парной связи грамматических и синтаксических классов.

Кроме этого, есть группа методов, осуществляющая проверку статистических гипотез об однородности распределения частотных характеристик текстов.

Для облегчения работы пользователя выполняемое им исследование хранится в виде проекта, включающего тексты, которые он хочет проанализировать, и использованные им методы. Для анализа доступны разобранные тексты, входящие в корпус произведений ПК «СМАЛТ», в основном это публицистические статьи XIX века. Могут быть использованы и другие тексты, но для этого они должны быть предварительно подготовлены с использованием ПК «СМАЛТ», а именно получены их морфологический и синтаксический разборы. База знаний информационно-аналитической системы подскажет пользователю методы анализа данных, которыми он может воспользоваться. Предложенные методы классифицируются по категориям, как наиболее предпочтительные для выбранных текстов, информативные или часто используемые. Если необходимо, то для выбранных методов указываются начальные условия.

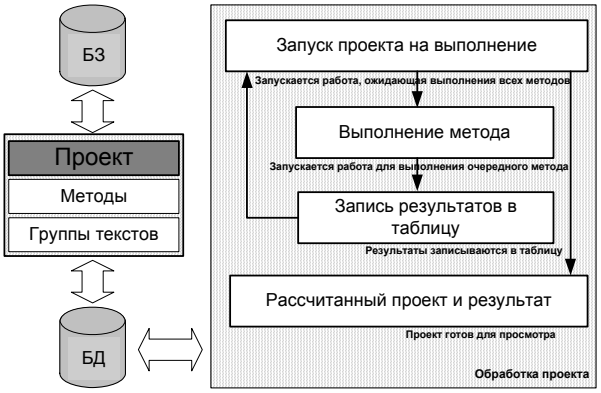

Рис. 4. Интерфейс пользователя

Число рассматриваемых методов можно расширять за счет модульной структуры информационно-аналитической системы. В процессе работы накапливается информация, позволяющая определить методы, которые обеспечивают более эффективную работу, то есть провести некоторое ранжирование методов. На основании этого проводится комплексное рассмотрение результатов тестов, с помощью введения оценки, учитывающей каждый из полученных результатов. Так каждый из тестов может давать положительное или отрицательное заключение о близости групп произведений, а комплексная оценка может учитывать все результаты и выдавать единый результат, причем вклад каждого теста берется в соответствии с введенным ранее ранжированием методов.

После выбора методов пользователь запускает проект на выполнение. Так как часть методов предполагает достаточно большой объем вычислений, то запускается работа, которая обеспечивает полную обработку проекта и не зависит от действий, выполняемых пользователем. Пользователь может закрыть браузер или наблюдать за тем, как происходят вычисления, см. рисунок 4.

После окончания расчетов можно проводить анализ полученных данных. Если обнаружена близость групп текстов по рассматриваемым признакам, то выводится соответствующее сообщение. На основании результатов всех методов, с учетом их ранжирования и используя «метода комитетов», может быть сделан вывод о наличии или отсутствии близости групп текстов.

В процессе работы пользователь может изменить проект, а именно добавить новые произведения и методы, в этом случае потребуется повторное проведение вычислений. Если получен интересный результат, например, обнаружено, что употребление глаголов определенным автором сильно отличается от остальных, то проект может быть продублирован исследователем, чтобы сохранить текущее состояние и начать работу с того же места, но в новом проекте, добавив для рассмотрения новые лингвостатистические параметры. Доступ к проекту может быть разделен с другими зарегистрированными участниками информационной системы.

8 Результаты апробации

С использованием ПК «СМАЛТ» был проведен ряд исследований по выявлению групп текстов, принадлежащих перу одного автора, а также их количественных характеристик [3, 5, 6, 8].

Проверялась методика Гейра Хетсо [9] для установления авторства. Было показано, что результаты исследования не зависят от того, какой текст рассматривается, в авторской орфографии и пунктуации или отредактированный для полного собрания сочинений. Проверена гипотеза о нормальности исходных выборок, тем самым правомерность использования статистических критериев в исследовании Г. Хетсо. Выявлена неустойчивость параметров «Средняя длина слова в буквах» и «Индекс разнообразия лексики» на разных объемах выборок.

Рассмотрен параметр «Распределение частей речи на первых трех и последних трех позициях в предложении». Показано, что увеличение числа формально-грамматических признаков не приводит к существенному изменению результатов классификации.

Метод оценки парной связи для грамматических классов (метод «сильного графа [1]) не дал возможности провести атрибуцию авторства обработанных статей, но была выявлена возможность анализа жанров.

Был проведен анализ публицистических произведений, с использованием синтаксического разбора, на основе методов многомерного статистического анализа, а также метода «сильного графа». Группы объектов, которые получились в результате, не дали четкого разделения по авторам произведений, причем группировка менялась в зависимости от выбранной меры и метода построения кластеров. При этом происходило разделение текстов маленького и большого размера, предположительно из-за того, что маленькие тексты имеют более бедный набор рассматриваемых параметров.

9 Заключение

С помощью программного комплекса создан и поддерживается размеченный корпус, основанный на публицистических статьях XIX века. Для текстов, включаемых в корпус, проводятся в автоматизированном режиме морфологический и синтаксические разборы. Доступ к корпусу поддерживается с помощью словарей, содержащих контекты. Использованные технологии могут быть применены для создания, хранения и доступа к мультиязычному корпусу, состоящего из произвольных текстов.

Входящая в ПК «СМАЛТ» информационно-аналитическая система для анализа текстов включает методы, позволяющие выделять группы текстов близкие по рассматриваемому набору лингвостатистических параметров. С одной стороны корпус используется как единый тестовый материал, дающий возможность исследовать сами методы на надежность и устойчивость, сравнивать их, создавать рекомендаций по применению.

А с другой стороны он сам является объектом для проведения исследований. Так с помощью ПК «СМАЛТ», совместно со специалистами-филологами, проводился поиск лингвостастистических параметров, определяющих стиль автора.

Литература

[1] Бородкин Л. И., Милов Л. В., Морозова Л. Е. К вопросу о формальном анализе авторских особенностей стиля в произведениях // Математические методы в историко-экономических и историко-культурных исследованиях. – М., 1977, – С. 298–326.

[2] Волков С. Св. Корпус текстов и исторический словарь / С. Св. Волков, В. П. Захаров // Русский язык конца XIX века: Проблемы изучения и лексикографического описания. – СПб., 2004. – С. 38–43.

[3] Захаров В. Н. Программная система поддержки атрибуции текстов статей Ф. М. Достоевского / В. Н. Захаров, А. А. Леонтьев, А. А. Рогов, Ю. В. Сидоров // Труды Петрозаводского государственного университета: Сер. Прикладная математика и информатика. Вып. 9. – Петрозаводск, 2000.

[4] Корпус текстов как особый тип лингвистической электронной библиотеки / С. Св. Волков, А. С. Герд, О. Н. Гринбаум и др. // Словарь русского языка XIX века. Проблемы. Исследования. Перспективы. – СПб., 2003. – С. 92–108.

[5] Рогов А.А., Сидоров Ю.В., Солопова А.И., Суровцова Т.Г. Информационно-аналитическая система «СМАЛТ» // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции "Диалог 2007" (Бекасово, 30 мая – 3 июня 2007 г.) / под ред. Л.Л. Иомдина, Н.И. Лауфер, А.С. Нариньяни, В.П. Селегея. – М. : Издательский центр РГГУ, 2007. – С. 470–474. – на рус. яз.

[6] Рогов А.А., Сидоров Ю.В., Суровцова Т.Г. Математические методы атрибуции литературных текстов небольшого объема // Материалы XIII Всероссийской конференции «Математические методы в распознавании образов». – М. : МАКС Пресс, 2007. – С. 525–528. – на рус. яз.

[7] Стенограмма обсуждения «Проекта Словаря русского языка XIX века» // Словарь русского языка XIX века. Проблемы. Исследования. Перспективы. – СПб., 2003. – С. 109–154.
[8] Суровцова Т. Г. Экспертная система для выявления скрытых количественных характеристик литературных произведений // Информационные технологии моделирования и управления. – 2007, №6(40). – С. 650–655. – на рус. яз.
[9] Хетсо Г. Принадлежность Достоевскому: К вопросу об атрибуции Ф.М. Достоевскому анонимных статей в журналах Время и Эпоха / Г. Хетсо – Oslo: Solum Forlag A.S., 1986. – 82 с.

Program Complex “SMALT”

A.A. Rogov, G.B. Gurin, A.A. Kotov,
Yu.V. Sidorov, T.G. Surovcova

In modern linguistics traditional methods of language data collection such as introspection, text data collection, experiment, and questioning are being replaced now by the corpus method. The creation of linguistic corpora of texts is one of the most important tasks of modern linguistics. Corpora are actively used in lexicography and in different linguistic research projects.

Researches of Petrozavodsk state university have been working with electronic text processing since 1995. The result of this work is the program complex “Statistical Methods of Literary Texts Analysis” (PC “SMALT”). It is based on a text database, which is composed of publicistic articles on different topics from St.Petersburg magazines of the XIX century in original spelling.

♣ Проект был поддержан грантами РГНФ № 02-04-12015в, 05-04-12418в, 08-04-12105в (руководитель А. А. Рогов). Адрес в Интернете: http://smalt.karelia.ru