IV Международный конгресс
исследователей русского языка

РУССКИЙ ЯЗЫК:
исторические судьбы
и современность

ТРУДЫ
и МАТЕРИАЛЫ

Москва, МГУ
20–23 марта 2010 г.

Русский язык: исторические судьбы и современность

IV Международный конгресс исследователей русского языка

Москва, МГУ имени М. В. Ломоносова, филологический факультет

20-23 марта 2010 года

Труды и материалы

Lomonosov Moscow State University (MSU)
Faculty of Philology

Russian Language: Its Historical Destiny and Present State

The Fourth International Congress of Russian Language Researchers

Moscow, Lomonosov Moscow State University,
FACULTY OF PHILOLOGY

March 20-23, 2010

Proceedings and materials

Collected by

Marina L. Remneva, Anatoliy A. Polikarpov

Moscow University Press

2010

Московский государственный университет имени М. В. Ломоносова
Филологический факультет

Русский язык: исторические судьбы и современность

IV Международный конгресс исследователей русского языка

Москва, МГУ имени М. В. Ломоносова,
ФИЛОЛОГИЧЕСКИЙ ФАКУЛЬТЕТ

20-23 марта 2010 года

Труды и материалы

Составители
М. Л. Ремнёва, А. А. Поликарпов

Издательство Московского университета
2010

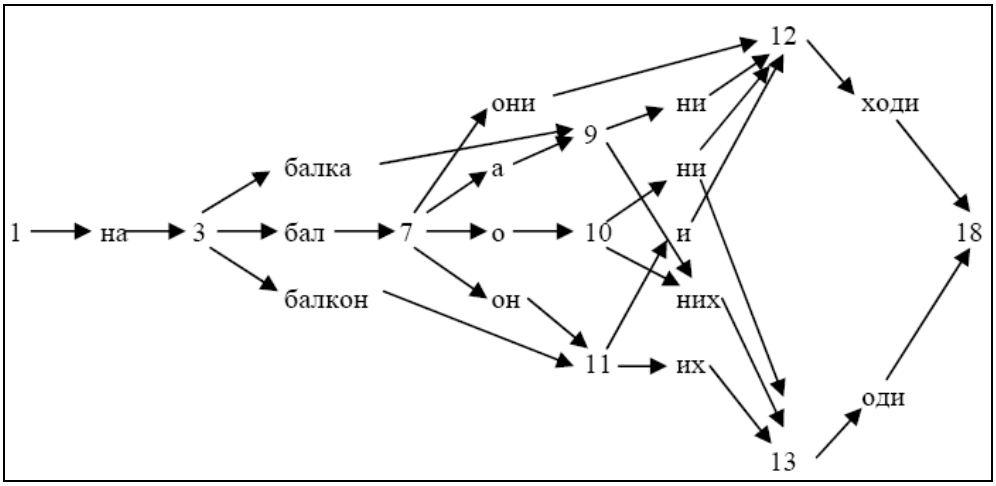

Рис. 1. Граф связей между выделенными фрагментами транскрипции предложения (ветви). Также показаны позиции начала и конца каждого фрагмента (вершины).

Следует отметить, что, например, фрагменту *-оди* соответствуют слова *йоде, коде, оде, поде*. Таким образом, количество вариантов даже для такого короткого предложения составляет O (m^5), где m — среднее число слов, имеющих одинаковую транскрипцию.

Учет ударений. Перепишем предложения примера, расставив ударения: *На ба́л ко́ни хо́дят* <–> *На балко́не хо́дят* и выполним транскрибирование: *наба́лко́нихо́ди* <–> *набалко́нихо́ди*. При этом число вариантов сокращается на 1–2 порядка. Дальнейшее сокращение числа вариантов требует учета синтагматики соседних слов и, далее, синтаксиса всего предложения. Статистика сочетаемости слов получена из синтаксически размеченного корпуса, мы же остановимся на применении синтаксиса для оценки наиболее вероятной последовательности слов, составляющих предложение (или его часть).

Статистический синтаксический анализ. Для анализа зависимостей в предложении разработан алгоритм [5], строящий покрывающее дерево всего предложения. В предлагаемой нами модели локальных связей структура зависимостей строится снизу вверх. Вначале устанавливаются связи между соседними словами (локальность), которые объединяются в юниты, затем устанавливаются связи между соседними юнитами, и так далее, пока не достигается последний, верхний уровень объединения, чем и завершается построение дерева зависимостей. Алгоритм локальных зависимостей проверялся на неразмеченном тексте и показал точность ~ 74.6% (число правильно установленных связей к общему числу связей в предложении). Применение алгоритма дает наиболее вероятный в лексическом и синтаксическом отношении вариант распознавания русского предложения (или его части) выделенного в слитной русской речи, как вариант с наибольшим весом установленных зависимостей.

Литература

1. *Зализняк А. А.* Грамматический словарь русского языка. М., 2008.
2. *Кривнова О. Ф., Захаров Л. М., Строкин Г. С.* Многофункциональный автоматический транскриптор русских текстов // Русский язык: исторические судьбы и современность. Международный конгресс исследователей русского языка. МГУ. М., 2001. С. 408–409.
3. *Потапова Р. К.* Речевое управление роботом. М., 1989; 2-е изд., доп. и пер. М., 2005.
4. *Потапова Р. К.* Перспективы прикладного речеведения // Речевые технологии. 2008. № 1. С. 5–17.
5. *Потемкин С. Б.* Неконтролируемый синтаксический анализ // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции «Диалог 2009». М., 2009. С. 409–414.
6. *Шелепов В. Ю., Ниценко А. В.* Структурная классификация слов русского языка. Новые алгоритмы сегментации речевого сигнала, распознавания фонем и их классов // Искусственный интеллект. 2005. № 4. Донецк. С. 679–690.
7. *Gao J., Suzuki H.* Unsupervised learning of dependency structure for language modeling // Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (7–12 July 2003). Sapporo, 2003. P. 521–528.

Развитие системы автоматического анализа текстов «СтилеАнализатор»

А. С. Кравцова, В. В. Поддубный, О. Г. Шевелев, А. А. Фатыхов

Томский государственный университет

askravtsova@gmail.com, pvv@inet.tsu.ru, oshevelyov@gmail.com, zyabloko@gmail.com

О. В. Кукушкина, А. А. Поликарпов

Московский государственный университет имени М. В. Ломоносова

kukush@orc.ru, anatpoli@mail.ru

Автоматический анализ текстов, извлечение частотных признаков, стиль текста, классификация текстов, кластеризация текстов

Summary. The thesis outlines a desktop text analysis tool developed by two universities. It specifies methods implemented, possibilities of the current version of the tool, and briefly describes its shortcomings. Finally, main features of a new web-generation of the tool that is being developed are presented.

В настоящее время лингвисты все чаще обращаются к автоматическим средствам анализа текстов. Простейший уровень анализа, как, например, подсчет количества слов в Word, прочно вошел в арсенал гуманитариев. К сложному — с использованием методов современной математической статистики и искусственного интеллекта — пока относятся с недоверием. Многие уже видят выгоды применения точных методов в лингвистике, но использовать эти методы пока проблематично даже в сотрудничестве с математиками и программистами. Успех исследований в квантитативной лингвистике во многом зависит от развитости и удобства программного инструментария.

В 2004 году на факультете информатики Томского государственного университета (ТГУ) началась работа над проектом «СтилеАнализатор». В 2005 году группа лингвистов филологического факультета Московского государственного университета имени М. В. Ломоносова (МГУ) подключилась к проекту. В 2006–2008 гг. совместный проект развивался на основе гранта РФФИ (06–07–89320). Суть проекта заключалась в создании многооконного (MDI) приложения для проведения разнообразных лингвистических исследований. Работа в программе делится на три этапа: 1) предобработка текстов, 2) преобразование текстов к количественному виду, 3) анализ количественных данных. Каждый

этап независим и предоставляет данные, доступные для использования в других системах.

В этап предобработки вошли такие операции, как унификация оформления, импорт грамматической разметки системы DicTUM-1 [1], замена по словарю (например, замена словоформ на их аффиксальные модели), специальные функции (например, удаление диалогов в тексте) и добавление заголовков.

Для этапа преобразования текстов к количественному виду был разработан специальный язык запросов, позволяющий подсчитывать частоты вложенных последовательностей элементов текста (букв, слов, предложений) с заданными параметрами (например, грамматические характеристики определенного слова). Полученные количественные данные сохраняют привязку к текстам, поэтому все исходные данные о произведениях и авторах можно использовать в анализе (например, классификация по авторам, жанрам, тематике) и отображать эту информацию на графиках и диаграммах. В 2007 году в «СтилеАнализатор» был добавлен специальный вид обработки — преобразование текстов к суффиксным структурам, позволяющим проводить анализ всех комбинаций элементов, присутствующих в наборе текстов.

Этап анализа в «СтилеАнализаторе» развит наиболее сильно. Реализованы три типа анализа: 1) структурный, 2) признаковый, 3) потоковый. В структурный анализ вошли функции работы со словарями текстов, фоносемантические функции, суффиксные деревья. Признаковый анализ, самый проработанный из трех, включил в себя иерархический кластерный анализ, проверку статистических гипотез, классификацию (деревья решений, нейронные сети, энтропийные методы), редукцию признакового пространства (через энтропию, анализ). Реализованные подходы содержат как оригинальные решения, так и модификации имеющихся. Для проверки результатов классификации реализованы современные методы тестирования (k-подмножеств, leave-one-out) и меры (точность, полнота, F-мера). Потоковые методы анализа работают на базе суффиксных деревьев. Пока они представлены в системе только кластеризацией по CS-, RS- or TS мерам.

«СтилеАнализатор» вот уже несколько лет активно тестируется и используется коллективом лингвистов МГУ с целью проведения множество исследований на больших корпусах текстов. Основная серия экспериментов была проведена в ходе работы по гранту РФФИ (06–07–89320). Разные корпуса текстов подверглись кластеризации и классификации с различными параметрами обработки. Главной целью экспериментов было выявление набора признаков, которые бы позволяли устойчиво различать тексты и авторов разных типов (функциональные стили, внутри них — жанры, авторы по полу, конкретные авторы и т. п.). Исследователями было отмечено, что хотя «СтилеАнализатор» и удобен для проведения большинства исследований и предоставляет большой спектр методов, в нем недостает средств обеспечения наглядности и прозрачности результатов. Основной интерес лингвистов состоит в раскрытии «черного ящика» математических процедур, в выявлении вопроса о том, как именно получен результат, какие языковые закономерности лежат в его основе. Работа лингвистов МГУ и математиков-программистов ТГУ, прежде всего, заключается в поиске оптимального сочетания определенных типов лингвистических признаков текстов (различающихся синтагматической протяженностью, например, буквы, морфемы, словоформы, словосочетания, предложения и т. п., а также различающихся степенью обобщения выбранных единиц по протяженности) с определенными статистическими средствами анализа, различными критериями значимости и т. п. при решении классификационных задач определенных типов.

Практическое использование «СтилеАнализатора», например, показало неудобство специального языка запросов (низкая скорость, излишняя вариативность). Изолированность системы (оконное приложение Windows) и работа с локальными файлами привели к путанице с многочисленными версиями текстовых и аналитических данных, затруднили предоставления системы третьим лицам без угрозы бесконтрольного распространения. Дополнительные проблемы возникают с дальнейшим увеличением объема исследуемых данных. Стало очевидным, что некоторые алгоритмы должны быть реализованы с учетом параллельных вычислений.

В итоге, в сентябре 2009 года было решено начать разработку нового поколения «СтилеАнализатора». Основная идея — на основе старой системы создать веб-приложение, работающее с текстами в базе данных. Такой подход существенно облегчает работу территориального распределенного коллектива, позволяет предоставлять отдельные функции системы заинтересованным людям. Разработка ведется на языке Java, используется СУБД MySQL и самые современные средства и технологии, такие как Spring, Google Web Toolkit. Распределение прав пользователей и параллельные вычисления закладываются в систему с самого начала.

В данный момент ведется работа над базовыми функциями работы с корпусом и реализацией словарно-аналитических методов, которые были слабо представлены в настольной версии программы. Предполагается, что первый год две системы будут использоваться совместно. Веб-версия в первую очередь воплотит в себе функциональность работы с корпусом текстов, обеспечит экспорт текстов в старую систему. Старая система пока будет использоваться для работы с количественными данными. В дальнейшем ее функции постепенно будут перенесены в новую систему.

Литература

1. *Kukushkina O. V., Polikarpov A. A.* DicTUM-1, a system for dictionary-text universal manipulations and analysis // http://www.philol.msu.ru/ ~lex/articles/dictum.htm.

Квантитативный анализ лексикографических материалов

С. В. Лесников

Государственное образовательное учреждение высшего профессионального образования «Сыктывкарский государственный университет»

serg@lsw.ru

Гипертекст, Интернет, компьютер, корпус, лексикография, лингвистика, текст, филология, языковедение, языкознание

Summary. Quantitative analysis of the text involves the calculation of a number of some quantitative characteristics of the body text. During the report expected to show the results of quantitative analysis of lexicographical material.

Во время доклада предполагается продемонстрировать результаты квантитативного анализа лексикографических материалов на примере корпуса художественных произведений на русском языке.

Для квантитативного (количественного, автоматического, автоматизированного, алгебраического, аналитического, вычислительного, инженерного, кибернетического, компьютерного, математического, механистического, статистического, численного…) анализа текстовой информации надо определиться с базовыми понятиями: что именно и по каким формулам будем считать.

Анализ текста осуществлялся по следующему алгоритму: 1) по заранее определенному списку разделительных символов (пунктуационных знаков, спец. знаков: конец строки, абзаца и др.) исследуемый текст разбивается на порции (том, книга, часть, раздел, глава, параграф, абзац, предложение, слово); 2) выделяются приставки, суффиксы, окончания (и др. аффиксы) для каждого слова; 3) определяется часть речи и уточняются атрибуты для каждого слова с помощью типовых алгоритмов; 4) определяются части предложения; 5) определяются субъекты и объекты в тексте и наличие связей между ними. Объекты и субъекты образуют в своих отношениях модель проблемы. Привнесение вопроса к модели замыкает ее.

Предложенный алгоритм прост, на первый взгляд, для исполнения человеком (с учетом уровня грамотности), однако для реализации на компьютере пока достаточно не формализован.