Секция XIV.
Компьютерная и квантитативная лингвистика

Периодизация творчества Ф. Тютчева при помощи количественного анализа его стиля

С. Н. Андреев

Смоленский государственный университет

smol.an@mail.ru

Стилеметрия, дискриминантный анализ, периоды творчества

Summary. The paper is devoted to the stylometric study of the changes in the style of a famous Russian poet of the 19th century, F. Tutchev. The analysis is based on the count of parts of speech and syllabic length of the words in his lyrics which were written at different periods of his life. Discriminant analysis allowed to reveal characteristics, which differentiate various periods and to suggest where to draw the demarcation line between the periods.

Одним из направлений стилеметрического анализа является поиск и анализ возможных изменений в индивидуальном стиле авторов, что позволяет найти объективные критерии для разграничения творчества на периоды. В последнее время этот аспект в стилеметрии привлекает все большее внимание лингвистов [1].

В данном сообщении ставится задача при помощи количественных методов анализа стихотворных текстов определить изменения стиля одного из наиболее известных российских поэтов второй половины XIX в. Федора Тютчева на протяжении его жизни и установить наиболее существенные границы его основных творческих этапов.

В качестве материала взяты все лирические стихотворения Тютчева, написанные четырехстопным ямбом. Признаковая схема включает частеречную отнесенность лексики (существительное, глагол, прилагательное, наречие, различные виды местоимений, функциональные слова, междометия), а также слоговой состав.

Для выявления дифференциальной силы указанных признаков использовался многомерный дискриминантный анализ.

В результате ряда тестовых процедур с различными наборами признаков и разным определением границ выделяемых этапов, было выявлено оптимальное разбиение лирики Тютчева на три группы. Эти три группы произведений достаточно хорошо соответствуют биографическим данным Тютчева. Первая группа включает произведения, написанные поэтом в то время, когда он в основном жил вне России (с 1822 до 1844). Ко второй группе относятся стихотворения 1844–1857 годов, то есть произведения, созданные после переезда в Россию и составляющие своего рода «переходный период». Третий период — это всеобщее признание поэта в России, отмеченное избранием его в Российскую Академию в 1857 году.

Выделение указанных трех периодов основано на установленной в ходе исследования дискриминантной модели, включающей такие признаки, как количество существительных, прилагательных, наречий, притяжательных местоимений (2-го лица), указательных местоимений, функциональных слов, личных местоимений первого лица и некоторых других в произведениях Тютчева. Из числа фонетических признаков релевантным оказалось количество трехсложных и шестисложных слов.

В целом, проведенное исследование показало наличие языковых изменений в стиле Тютчева, которые совпадают с важными биографическими данными поэта.

Литература

1. *Juola P.* Authorship attribution // Foundations and Trends in Information Retrieval. Volume 1, Issue 3 (December 2006). Hanover; MA, USA, 2006, P. 233–334.

Информационные и метаинформационные компоненты речи при ее автоматической обработке

С. В. Андреева

Саратовский государственный университет имени Н. Г. Чернышевского

svandreeva@rambler.ru

Виды информации, основные и вспомогательные единицы, информационный мусор

Summary. In speech information of different kinds (factual, metacommunicative, discursive, signalling) is rendered simultaneously, and one can find elements of these structures in one phrase. Informatively heterogeneous discourse units can be represented as a three-level hierarchy: basic — supplementary — informational “garbage”.

1. Единицы дискурса передают информацию четырех разных видов [2]. Под **фактуальной** информацией понимается все то, что пополняет интеллектуальный запас знаний человека или содержит сведения бытового характера, необходимые в данный момент.

Со стороны адресата коммуникация предполагает не только декодирование семантики языковых знаков, но и распознавание разнообразной метакоммуникативной и дискурсивной информации. **Метакоммуникативная** информация отражает взаимодействие между автором речи и ее адресатом и, следовательно, направлена на организацию общения: информативность в контакто-регулирующем плане (поддержание контакта, оформление этапов интеракции); информативность в оценочно-интерпретационном плане речи / ситуации; информативность в плане интерперсональных отношений (интимизация общения, смягчение категоричности суждений и т. п.). **Дискурсивная** информация направлена на организацию дискурса и ориентацию адресата в нем для оптимального восприятия как фактуальной, так и коммуникативной информации: информативность в плане структурирования дискурса, обозначения роли фрагмента в тексте, отношения к нему автора и т. д. Можно выделить и **сигнальную** «информацию» — информативность непроизвольных речевых проявлений (в том числе табуированных восклицаний и их заменителей), т. е. выражение психофизиологического и эмоционально-чувственного состояния говорящего.

В результате исследования «многоликой» речевой реальности нами выделены два типа единиц: **основные**, составляющие «костяк» синтаксиса речи, и **вспомогательные**, используемые говорящим факультативно, т. к. их функционирование ограничено конкретными речеорганизующими задачами [1]. Основные единицы речевой коммуникации представлены как предложенческими структурами на основе свободного конструирования, так и непредложенческими клишированными единицами релятивного типа, не обла-

дающими признаками грамматической моделируемости и предикативности, но отличающимися семантической и интонационной завершенностью и вместе с предложенческими единицами образующие «костяк» («тело») дискурса, особенно в устной диалогической речи.

2. Появление в речи вспомогательных единиц «оправдано не с языковой, а с речевой точки зрения» [4]. Такие единицы не несут фактуальной (предметно-фактической) информации, но выполняют в речи определенные функции. Они направлены **на оптимизацию** передачи фактуальной информации. **Дискурсивные** элементы структурируют речевой поток, показывая последовательность информационных блоков (*во-первых, прежде всего*), обозначая роль фрагмента в тексте (*теперь о главном*) и отношение к нему автора (*на мой взгляд, к сожалению*), вводя дополнительную информацию в качестве «мостиков-переходов» (*впрочем, кстати*), подытоживая сказанное (*таким образом, итак, ладно*) и т. п.

Метакоммуникативные элементы способствуют выполнению функции поддержания контакта: этикетные формулы (*добрый день!*), актуализаторы (*да, ага, правда*), интимизаторы общения (*слушай, знаешь, смотри, представь*), заполнители пауз и т. п. В целях гармонизации коммуникации, психологической синхронизации собеседники используют «опустошенные», лишенные номинативного содержания единицы с формальной предикативностью (***Знаешь*** / *где я видела хорошие? В «Архипелаге»!; Наташа* / ***слушай*** / *у тебя есть все адреса?*). Посредством целых предикативных структур регулятивной направленности говорящий заявляет о своем коммуникативном намерении, привлекая и активизируя внимание слушающего (***Хотела задать тебе вопрос*** / *у тебя у телефона нет такой функции?*; ***Я знаешь что придумала*** / ***хочу тебе рассказать*** // *На прошлой неделе…*). К нашим выводам близки теоретические положения О. Б. Йокоямы о семи видах знания, в том числе метаинформационном, совершенно необходимом для успешного осуществления информационного дискурса [3]. Для нас важно утверждение исследовательницы о том, что средства передачи метаинформационного знания (знания кода и дискурсивной ситуации) вырабатываются в каждом языке, следовательно, само это знание универсально для разных языковых систем и без него невозможен прием фактуальной информации.

3. К понятию «информационного мусора» можно отнести прежде всего слова-паразиты, обусловленные в основном низким уровнем речевой культуры говорящего (например, частотное и неуместное *значит, там* (*Ну ты когда распечатываешь* ***там*** / *черно-белым* ***там*** / *нельзя никак подставить?*), а также смысловую тавтологию (*Есть ли* ***свободные вакансии*** / *нет на сегодня?*). Сюда же относятся слова, не осуществляющие определенной функции в речи, а возникающие под воздействием временных экстралингвистических факторов. Например, чрезмерно частое употребление говорящими таких слов, как *ну, вот, прям, просто, вообще, типа, как бы*, что объясняется данью моде (*И вообще* / *я* ***как бы*** *хочу сказать…*). Использование явно ненужных слов и структур связано с повышенным эмоциональным состоянием говорящего (*Это к тебе еще просьба…* ***вот*** // *Можешь…* ***вот*** / *записи отдать?* или ***А ты знаешь что? Я не знаю*** / *а у тебя никакого покрывала* / *ничего такого нету?*). Появление «информационного мусора» может быть обусловлено также обстановкой, в которой протекает коммуникативный акт, например, шумовым фоном, рассеянным вниманием собеседника или перебиванием говорящего слушающим (*Вот* ***я говорю*** / *ты отправь по почте* // *На меня* ***говорю*** *не надейся* //).

4. В речи информация разных видов передается единим потоком и в одной фразе можно обнаружить элементы этих структур. Единицы дискурса с разной информативностью можно представить в виде шкалы: основные — вспомогательные — информационный мусор. При автоматической обработке речи неизбежны потери части содержания («информационный сброс»). Учет закономерностей формирования и языкового выражения информационных и метаинформационных компонентов позволит избежать избыточности «глобальной структуры текста» за счет устранения единиц, получивших в ходе анализа малый информационный вес.

Литература

1. *Андреева С. В.* Типология конструктивно-синтаксических единиц в русской речи // Вопросы языкознания. 2004. № 5. С. 32–45.
2. *Андреева С. В.* Речевые единицы русской речи: Система, зоны употребления, функции. Изд. 2-е, испр. М, 2006.
3. *Йокояма О. Б.* Когнитивная модель дискурса и русский порядок слов. М., 2005.
4. *Кокошкина И. В.* Функционирование «лишних» слов в русской речи // Проблемы речевой коммуникации. Саратов, 2008. С. 182–191.

Звуковой корпус как материал для многоуровневого анализа русской речи и база для различных инновационных проектов*

А. С. Асиновский, Н. В. Богданова, А. И. Рыко, С. Б. Степанова, Т. Ю. Шерстинова

Санкт-Петербургский государственный университет

a.s.asinovsky@gmail.com, nvbogdanova_2005@mail.ru, aryko@mail.ru, stsvet_2002@mail.ru, sherstinova@gmail.com

Звуковой корпус русского языка, грамматика речи, многоуровневая лингвистическая разметка, спонтанная речь, фонетика, повседневная коммуникация, информационные и речевые технологии

Summary. The paper presents the Russian speech corpus which consists of two main parts: 1) the ORD corpus of Russian everyday communication, which contains recordings of all spoken episodes recorded by 40 subjects during twenty-four hours and 2) recording of linguistically balanced speech material from different social groups. The main directions of innovational scientific research projects based on this corpus are discussed.

1. Звуковой корпус русского языка и его основные модули

Одним из наиболее актуальных направлений современной лингвистики является сбор и систематизация живого речевого материала. Наилучшим способом решения этих задач представляется **корпусный подход**, который позволяет осуществлять мониторинг и многоуровневое описание современной русской речи, а также является удобной базой для различных инновационных проектов.

Именно такой подход был реализован при создании **Звукового корпуса русского языка** (ЗКРЯ), который активно разрабатывается в настоящее время на факультете филологии и искусств СПбГУ. Перспективной задачей данного корпуса является фиксация разных форм естественного языка в целях создания **грамматики русской речи**, описания фонетики спонтанной речи и решения актуальных задач современных речевых технологий.

ЗКРЯ включает в себя два блока, устроенных принципиально по-разному. Первый блок реализует уникальный проект «Один речевой день» (ОРД) и строится по **принципу невода**: разные люди живут один день «с диктофоном на шее» и записывают всю свою речевую продукцию. В результате в зоне внимания исследователей оказываются не только записи речевого материала, но и все характеристики информантов и их коммуникантов [3]. Модуль ОРД содержит 320 часов звучания реальной речевой коммуникации, полученные от 40 информантов (20 мужчин и 20 женщин) и более чем 500 их коммуникантов. Речевой материал разбит на крупные коммуникативные эпизоды (994 файла). В результате расшифровки 34 часов звукозаписей получены тексты общим объемом в 244075 слов на уровне *Frase*, которые относятся к 125 коммуникативным эпизодам. Сплошное аннотирование звукозаписей осуществляется по следующим восьми уровням: *Frase* (реплики говорящих), *Speaker* (код

*Исследование выполнено при поддержке гранта РГНФ «Разработка информационной среды для мониторинга устной русской речи» (09–04–12115в).

говорящего), *Events* (невербальные аудиособытия), *Voice* (качество голоса говорящего), *FonetCom* (фонетический комментарий), *FraseComment* (фразовый комментарий), *Notes* (общий комментарий), *Episode* (мини-эпизод речевой коммуникации). Проводится выборочная сегментация, аннотирование и транскрибирование на лексическом и морфемном уровнях.

Второй блок корпуса строго сбалансирован по разным параметрам — собственно лингвистически, социологически и психологически (реализован **принцип ковчега**). В результате объектом внимания исследователей стали устные монологи различной степени лингвистической мотивированности и спонтанности — **чтение** и **пересказ** исходного текста, **описание изображения** и **свободный рассказ на заданную тему**, — произнесенные информантами с различными социальными и психологическими характеристиками [1]. Этот модуль содержит речевой материал, записанный более чем от 120 дикторов разных социальных групп и профессий (медицинские работники, юристы, программисты, преподаватели-филологи, студенты разных специальностей и некот. др.), всего более 30 часов звучания.

Для автоматической обработки материалов корпуса используется информационная исследовательская среда для мониторинга русской устной речи (см. [1]). Она позволяет описывать и аннотировать разные сегментные типы речевого потока (от «эпизода разговорного дня» до отдельных звуков), осуществлять автоматический анализ полученной лингвистической и паралингвистической разметки, исследовать корреляцию данных на разных уровнях, заниматься моделированием, оценкой и прогнозированием речевого поведения в зависимости от разных условий.

2. Основные направления исследования

Материал ЗКРЯ позволяет описывать специфику устной спонтанной речи на всех уровнях, анализировать внутриязыковую интерференцию (бытовая речь различных профессионально ориентированных групп); описывать дистрибуцию тех или иных грамматических классов слов или их форм в устной монологической и диалогической речи разных социальных групп, дать лексикографическое описание бытовой спонтанной звучащей речи; решать разнообразные задачи обработки естественного языка / речи, а также интегрального моделирования звуковой формы естественного языка.

Наиболее интересными **инновационными проектами**, которые уже осуществляются на этом материале, являются следующие:

- публикация расшифровок спонтанных монологов разного типа (см., например: [2]);
- создание конкордансов разного типа: 1) по текстам, произнесенным одними и теми же группами информантов, но в рамках различных коммуникативных сценариев; 2) по текстам информантов, объединенных признаками пола, возраста, профессии и т. п.; 3) по всему корпусу ОРД;
- создание словаря русской бытовой разговорной речи;
- создание словаря контекстных экспрессем русской разговорной речи;
- создание словаря редуцированных форм русской речи;
- описание дистрибуции и реализации русских аффиксов в реальной речи;
- описание дистрибуции частей речи;
- описание метакоммуникации в спонтанной речи;
- описание коммуникативных стратегий, используемых говорящими для реализации различных речевых сценариев;
- уточнение инвентаря метаединиц для описания устной спонтанной речи;
- разработка учебных материалов нового поколения для студентов-русистов и для обучения русскому языку нерусских.

Литература

1. *Богданова Н. В., Асиновский А. С, Русакова М. В., Рыко А. И., Степанова С. Б., Шерстинова Т. Ю.* Звуковой корпус как способ мониторинга и фиксации разных форм естественного языка» // Международная конференция «Диалог'2009» (Бекасово, 27–31 мая 2009 г.): Труды и материалы. Вып. 8 (15). М., 2009. С. 38–44.
2. Русская спонтанная речь. Свободные монологи-рассказы на заданную тему. Тексты. Лексические материалы / Сост. В. В. Куканова; отв. ред. и автор предисловия Н. В. Богданова. СПб., 2008.
3. *Asinovsky A., Bogdanova N., Rusakova M., Stepanova S., Ryko A., Sherstinova S.* The ORD Speech Corpus of Russian Everyday Communication «One Speaker's Day»: Creation Principles and Annotation // LNCS / LNAI series. «Text, Speech and Dialogue» TSD-2009. Berlin; Heidelberg. 2009. P. 250–257.

Сайт «Казанская лингвистическая школа»: лингвоинформационное обеспечение

К. Р. Галиуллин, Е. А. Горобец, Р. Н. Каримуллина

ГОУ ВПО «Казанский государственный университет им В. И. Ульянова-Ленина»

Kamil.Galiullin@ksu.ru, Elena.Gorobets@ksu.ru, elena_gorobets@mail.ru, re_ka@mail.ru

Сайт, Казанская лингвистическая школа, история языкознания, интернет-ресурс

Summary. Kazan linguistic school is one of scientific schools well-known in Russia and all over the world. Linguoinformational providing of Kazan linguistic school site is the project of great importance. The report describes materials of our site: its structure, technical details, linguistic components — and touches upon problems of its development in different trends.

Казанская лингвистическая школа (КЛШ) — одна из научных школ Казанского университета, имеющих мировую известность. Идеи КЛШ послужили основой для развития многих направлений современного языкознания.

К сожалению, до начала XXI века получение подробной информации о КЛШ для многих ученых было довольно проблематичным. В особенности это касается полнотекстовых версий трудов. Часть работ И. А. Бодуэна де Куртенэ вошла в известный двухтомник 1963 года, но это лишь малая часть. То же касается и трудов остальных представителей КЛШ. Для зарубежных и многих отечественных коллег эти материалы практически недоступны.

В целях оптимизации обмена информацией с научным миром в России и за рубежом, в целях сохранения и продолжения традиций всемирно известной Казанской лингвистической школы, а также в целях информационного обеспечения процесса преподавания ряда основных лингвистических дисциплин, связанных с идеями представителей КЛШ, на кафедре теоретической и прикладной лингвистики филологического факультета Казанского государственного университета в 2004 году был создан сайт, посвященный КЛШ (http://www.kls.ksu.ru). Для размещения материалов сотрудниками кафедры была разработана система администрирования сайта на языке PHP с использованием базы данных MySQL, которая в настоящее время активно используется и совершенствуется. Разделы сайта постоянно пополняются, проводятся работы по разысканию и оцифровке материалов, которые вносятся в базу и подвергаются дальнейшей выборке по различным параметрам. Целесообразность максимально активного развития сайта состоит в том, что в сети Интернет информация о КЛШ представлена очень скудно. Сайт «Казанская лингвистическая школа» является на данный момент единственным источником развернутой информации в Интернете о деятельности и научном наследии КЛШ.

На электронный почтовый адрес сайта (kazanlinguist@mail.ru) приходит множество благодарственных писем от российских и зарубежных коллег, которым материалы, размещенные на сайте, помогли в их научных разысканиях. Публикация малодоступных текстов представителей известных научных школ вообще является делом необходимым и перспективным; таким образом, одной из ближайших наших задач мы видим подготовку к публикации и размещение на сайте трудов представителей КЛШ, которые содержатся в

Научной библиотеке им. Н. И. Лобачевского Казанского государственного университета и в Российской государственной библиотеке.

В настоящее время сайт состоит из шести блоков: «Персоналии», «История», «Новости», «Материалы конференций», «О сайте» и «Контакты». В блоке «Персоналии» (http://www.kls.ksu.ru/persons.php) представлена информация об основателе Казанской лингвистической школы И. А. Бодуэне де Куртенэ и о представителях школы (В. А. Богородицкий, Н. В. Крушевский, А. И. Александров, А. И. Анастасиев, А. С. Архангельский, С. К. Булич, П. В. Владимиров, Н. С. Кукуранов, В. В. Радлов). В каждом разделе представлена информация биобиблиографического характера, публикуются полнотекстовые версии трудов каждого из ученых и труды, посвященные им. Выставлены результаты аналитических интернет-обзоров, посвященных жизни и деятельности представителей КЛШ. К публикации готовятся материалы, связанные с жизнью и творчеством таких учеников Бодуэновской школы, как В. В. Плотников и А. А. Царевский (студенты Казанской духовной академии). Сведения о них практически не проникали в работы о Казанской лингвистической школе в двадцатом веке; необходимым является, таким образом, обращение к архивным материалам. На сайте также опубликованы выходные данные ряда публицистических трудов И. А. Бодуэна де Куртенэ, которые еще нигде не были описаны. Ведется подготовка к публикации их полнотекстовых версий.

В 2006 году в рамках Бодуэновских чтений издан перечень просмотренных de visu публикаций И. А. Бодуэна де Куртенэ на русском языке, в печати находится полный библиографический перечень трудов В. А. Богородицкого и В. В. Радлова. На сайте опубликована полнотекстовая версия двухтомника И. А. Бодуэна де Куртенэ 1963 года, а также ряд статей ученого, написанных им для «Русской энциклопедии» (1913 год, 4 том) (http://www.kls.ksu.ru/boduen/bibbod.php?id=15&num=1000000), идет работа по оцифровке русскоязычных публикаций основателя КЛШ. Труды В. А. Богородицкого (http://www.kls.ksu.ru/bogor/work_bogor.php) представлены полнотекстовой версией его статьи о Казанской лингвистической школе, в процессе размещения находятся «Общий курс русской грамматики», «Введение в изучение современных романских и германских языков». Опубликованы труды Н. В. Крушевского «К вопросу о гуне. Исследования в области старославянского вокализма», «Очерк науки о языке», «Заговоры как вид русской народной поэзии», «Лингвистические заметки. I. Новейшие открытия в области арио-европейского вокализма. II. Изменения согласных групп вида ЕЕ. III. О морфологической абсорбции», «Об аналогии и народной этимологии», «Предмет, деление и метод науки о языке»; размещены документы о Н. В. Крушевском из Национального архива Республики Татарстан и «Программа чтений приват-доцента Н. В. Крушевского в 1879–1880 гг.», опубликованные в ходе II Международных Бодуэновских чтений в сборнике, посвященном Н. В. Крушевскому (http://www.kls.ksu.ru/krush/workrush.php); готовятся к публикации на сайте «Очерки по языковедению. I. Французская грамматика. II. Антропофоника», «Восемь гимнов Ригведы».

Размещены полнотекстовые версии материалов традиционных международных Бодуэновских чтений, проходивших в Казанском университете в 2001, 2003, 2006 годах; размещаются тезисы чтений 2009 года, материалы конференции, посвященной В. А. Богородицкому (http://www.kls.ksu.ru/boduen/chtenia_list.php). В разделе «История» приводится перечень работ о КЛШ; некоторые из этих работ представлены в полнотекстовой версии (http://www.kls.ksu.ru/events.php).

В процессе разработки находятся система рассылки новостей и гостевая книга, а также раздел «Контакты», где планируется размещение перекрестных ссылок на лингвистические сайты в Интернете.

В создании сайта принимают участие преподаватели, сотрудники и студенты кафедры теоретической и прикладной лингвистики (http://www.kls.ksu.ru/tpl.php) филологического факультета Казанского государственного университета.

Индикаторный подход к многоаспектному поиску информации

В. Д. Гусев, Н. В. Саломатина

Институт математики им. С. Л. Соболева СО РАН (Новосибирск)

gusev@math.nsc.ru, nataly@math.nsc.ru

Извлечение информации, аспект содержания, прецедентный текст, индикаторный словарь

Summary. The problems of forming, enriching and use of cue-words dictionaries for the automatic detection of the various aspects of scientific text content, and also for the identification of structures such as «text in the text», are considered. The results of real text processing are presented.

Введение. Одним из методов извлечения данных и знаний из текста является использование индикаторных словарей, несущих существенную **сопутствующую** информацию (в виде слов, словосочетаний или шаблонов) об отдельных аспектах содержания текста или наличии специфических объектов в нем. Например, в научных текстах сочетания «в статье рассматриваются», «в настоящей работе», «ставится задача» и т. п. часто сигнализируют о **цели работы**, а клише «народная мудрость гласит», «справедлива пословица», «как говорил N.» предшествуют вкраплению прецедентного текста в авторский.

Основы индикаторного подхода были заложены в 70-е годы прошлого столетия (см. обзоры [2]; [3]). Его достоинствами являются простота и возможность применения к широкому кругу объектов и явлений. Однако каждому объекту или аспекту содержания сопутствуют свои индикаторы, а формирование словарей индикаторов требует больших затрат ручного труда. **Целью доклада** является описание разработанной нами **технологии частичной автоматизации** этого процесса и возможностей **пополнения словарей** без привлечения дополнительного обучающего материала.

Автоматизация формирования и обогащения индикаторных словарей. Для создания индикаторных словарей необходима представительная подборка текстов $T = \{T_1, \ldots, T_i, \ldots, T_m\}$. Автоматизация отбора индикаторов основана на предположениях об их **повторяемости в разных текстах**, **невысокой частоте встречаемости в отдельно взятом тексте** и выполнении ряда других условий. Удобным инструментом для проверки всех требований является программа вычисления **совместного L-граммного спектра** $\{T_i\} \in T$ [1]. Спектр $\Phi(T)$ аккумулирует частотно-позиционную информацию о цепочках из L подряд следующих слов (L-грамм), представленных хотя бы в паре текстов из T ($L = 1, 2, \ldots, L_{max}(T)$, $L_{max}(T)$ — длина максимального межтекстового повтора). Процедура отбора индикаторов сводится к вычислению $\Phi(T)$ для нормализованных T_i с последующей фильтрацией L-грамм, не удовлетворяющих введенным ограничениям. На завершающем этапе эксперт компонует словарь из найденных потенциальных индикаторов.

Полученный словарь может быть обогащен без увеличения исходной подборки путем варьирования отобранных L-грамм с использованием «допустимых» редакционных операций. Например, ввиду близости понятий «задача» и «проблема» и при наличии в словаре индикатора «важнейшая проблема» можно пополнить словарь и индикатором «важнейшая задача», если он там отсутствует. Такие условно синонимичные подстановки извлекаются из $\Phi(T)$ путем анализа частично перекрывающихся L-грамм. Например, наличие цепочек вида *в (данной, настоящей, предлагаемой, …) работе* позволяет считать слова в скобках условными синонимами. Так возникают индикаторы в виде шаблонов с одной или несколькими переменными: в\$x\$y\$z, $x \in$ {*этот, данный, предлагаемый, …*}, $y \in$ {*статья, доклад, работа, …*}, $z \in$ {*рассматриваться, анализироваться, обсуждаться, …*}. Они могут включать в себя и индикаторы, отсутствовавшие в исходной подборке. Пополнение словаря возможно и при использовании других типов варьирования: **словообразования** (*важный — важнейший, говорил — говаривал*), **пере-**

становок слов (*актуальной задачей является — является актуальной задачей*), ограниченных по длине **вставок** (*как говаривал N. — как говаривал известный писатель N.*) и др.

Выявление аспектов содержания в научных текстах. На материале докладов конференции по компьютерной лингвистике (Диалог'2002, 146 докладов, ~ 442 тыс. словоупотреблений) построены индикаторные словари для выявления 12 аспектов содержания (**рассматриваемая проблема**, **цель**, **новизна**, **возможности дальнейшего развития**, …). Они содержат свыше 1000 слов, словосочетаний и шаблонов. Ведется их пополнение путем рассмотрения других предметных областей (индикаторы «решается», «доказывается» и др. присутствуют в текстах по математике, но практически отсутствуют в текстах по компьютерной лингвистике). Эксперименты по выявлению аспектов содержания научных текстов показали, что полнота и точность поиска зависят от конкретного аспекта и в среднем составляют, соответственно, ~ 70% и 60%. Фразы, содержащие аспектные индикаторы, как правило, присутствуют в авторских аннотациях.

Возможности автоматизации процесса выявления прецедентных текстов. Прецедентные тексты в виде крылатых фраз, цитат, стихотворных строк и т. п., их состав, способы варьирования и встраивания в основной текст являются важными характеристиками авторского стиля и интересуют лексикологов с разных точек зрения. Автоматическая идентификация этих объектов в тексте часто затрудняется отсутствием индикаторов, многообразием, малой длиной объектов (одно-два слова), вариативностью и пр.

В отличие от предыдущего случая (индикаторы аспектов) большое значение в плане информационного поиска приобретает каталогизация и инвентаризация прецедентных текстов. Поиск следует осуществлять не только по индикаторам, но и по объектам каталога. В первом случае (поиск по индикаторам) пополняется новыми объектами каталог. Во втором (поиск по объектам каталога) — словарь индикаторов (при известном объекте может появиться новый индикатор). Последовательные итерации обоих типов позволяют улучшить полноту и точность поиска. На данный момент по такой схеме составлен словарь из 250 индикаторов разной длины (**афоризм**, **поговорка**, **народная мудрость**, ***недаром говорят***, ***на ум приходит знаменитое***, …). В сочетании с пунктуационными индикаторами этот словарь обеспечил на текстах «Политкома» довольно высокую полноту поиска (~ 80%) при относительно невысокой точности (~ 24%). Примеры найденных объектов: ***Как говорил*** *маршал Маклюэн, «если хочешь что-то узнать про воду, не спрашивай у рыбы»*; ***Принято говорить****: не расстраивайся — это кино*. Полученные результаты мы оцениваем оптимистично, т. к. имеются различные резервы для повышения точности. Кроме того, следует учитывать, что именно индикаторы обеспечивают возможность выявления «новых» (не содержащихся в каталоге) объектов.

В заключение следует отметить, что индикаторный подход весьма перспективен в плане многоаспектного извлечения информации. Однако его целесообразно использовать в сочетании с другими подходами, поскольку искомые объекты (или аспекты) не всегда сопровождаются индикаторами.

Литература

1. *Гусев В. Д., Саломатина Н. В.* *L*-граммное представление текстов на естественном языке и его возможности // Материалы Всеросс. конф. «Квантитативная лингвистика: исследования и модели» (КЛИМ-2005), 6–10 июня 2005 г. Новосибирск, 2005. С. 256–270.
2. *Пащенко Н. А., Кнорина Л. В., Молчанова Т. В. и др.* Проблемы автоматизации индексирования и реферирования // Итоги науки и техники. Информатика. 1983. Т. 7. С. 7–164.
3. Advance in Automatic Text Summarization // Ed. by: I. Mani and M. T. Maybury. Cambridge, Massachusetts, 1999. P. 433.

Электронный словарь лингвистической терминологии тезаурусного типа

Н. П. Дарчук, Л. А. Алексеенко

Киевский национальный университет имени Тараса Шевченко (Украина)

lingvo@voliacable.com

Термин, информационно-поисковая система, тезаурус, тезаурусные отношения

Summary. The paper is devoted to the description of the lexicographical and encyclopedic electronic database of linguistic terms which is arranged and sorted in 4 ways, in particular, by alphabetic order, by the meaning, and as thesaurus, containing totally 3400 entries from the following fields: morphology, syntax, lexis, semantics, computer linguistics. It was developed a technology of information retrieval system i. e. thesaurus which may function as reference system or as an integral part of any other intellectual system.

Одной из актуальных междисциплинарных задач нашего времени является логико-понятийное моделирование терминосистем различных областей знаний, поскольку эти модели необходимы при составлении терминологических словарей, тезаурусов, баз данных и баз знаний, систем искусственного интеллекта. Частным случаем моделирования знаний можно считать построение электронного тезауруса, который, с одной стороны, является способом формализованного представления терминологии, а с другой — считается важным источником совершенствования систем знаний конкретных наук.

Этой проблеме посвящен проект «Электронный словарь лингвистической терминологии с информационно-поисковой системой (тезаурус)», выполненный в лаборатории компьютерной лингвистики Киевского национального университета.

Цель проекта — 1) составление электронного Словаря лингвистических терминов с использованием новой формализованной методики конструирования тезауруса, отвечающей современным стандартам терминографии, и представление его в сети Интернет; 2) благодаря разработанным компьютерным технологиям верификация теоретической тезаурусной модели путем применения ее для анализа корпуса текстов на русском / английском / украинском языках по различным разделам лингвистики.

Работа над проектом осуществлялась в **два этапа**. На **первом этапе** создавался электронный тезаурус в виде лексикографической и энциклопедической электронной базы лингвистических терминов, который состоял из трех словарей: алфавитного, толкового и тезаурусного. В алфавитном для каждого терминологического слова или словосочетания (3400 терминов) поданы русский и английский эквиваленты и толкования из авторитетных источников (около тридцати): терминологических словарей, грамматик, монографий. В словарь включены общелингвистические термины (преимущественно существительные или именные словосочетания) из всех разделов грамматики, лексикологии, прикладной и компьютерной лингвистики.

Тезаурусный словарь представляет собой перечисление логико-семантических функций между лингвистическими терминами (список функций заимствован из работы [1], но дополнен и модифицирован нами). Построение тезауруса предусматривает раскрытие всех типов отношений между понятиями, основными из которых являются гипонимия (род / вид), соподчинение на одном уровне — парциация (часть / целое), синонимия, корреляция, ассоциация, функция, способы выражения функции и др. Разработанный электронный словарь включает не только множество отдельных терминов, представленных в виде алфавитного списка с их толкованиями, но и сами модели представления отношений между терминами в виде семантической сети — иерархизированной структуры данных, в которой выделяются узлы (термины) и дуги, выражающие разные типы отношений между узлами.

Сетевое представление данных имеет не только чисто прикладное значение, но и позволяет глубже проникнуть в систему логики данной науки, точнее смоделировать терминосистему по лингвистике. Тезаурус состоит из 3394 терминов, которые охвачены семантической сетью в 9265 семантических отношений. Созданная на этом этапе модель является **статическим** представлением логико-понятийных отношений между терминами данной лингвистической терминосистемы.

Можно рассмотреть эту модель в плане отображения **динамических** аспектов структуры научного знания, а именно как верификацию теоретической тезаурусной модели путем применения ее к анализу корпуса текстов из разных разделов лингвистики. Важность такого исследования объясняется тем, что любое знание имеет текстовое выражение и познается через текст. Энциклопедическая модель научного знания является производной от множества реальных текстов и репрезентацией этих текстов на уровне семантической модели. По отношению к множеству терминов определенной науки логико-понятийная система области знания — это модель плана содержания области знания. Логико-понятийная структура текста отражает основные элементы семантической парадигматики текстов.

На **втором этапе** осуществлялось построение динамической логико-понятийной модели путем наложения тезаурусной модели лингвистических терминов в виде иерархической классификационной схемы — сети на словник научного текста. В результате получаем также иерархическую классификационную сеть конкретного анализированного текста с абсолютной частотой употребления в конкретном тексте. Этим обеспечивается подход — от терминологического словаря — к тексту, причем частота употребления и контексты дают возможность концентрировать разрозненную терминологическую информацию для разрешения различных терминологических задач (напр., целесообразность включения предтермина в создаваемый словник терминов).

Методика состоит из таких основных этапов автоматической обработки корпуса текстов: а) лемматизация и упорядочение по частеречной принадлежности; б) определение для каждой леммы (существительного или прилагательного) абсолютной частоты употребления; в) построение тезаурусного графа терминов конкретного текста с абсолютными частотами употребления в тексте путем наложения тезаурусной сети терминосистемы; г) снятие омонимии значений терминов; д) построение дополнительного словника слов с абсолютными частотами, которые не вошли в тезаурус; е) поиск слов-предтерминов с иллюстративными контекстами. (Проверка работы тезауруса осуществлялась на корпусе научных статей журнала «Вопросы языкознания»; длина текста около 80 тыс словоупотреблений).

Важность Проекта в том, что, во-первых, электронный тезаурус в мультимедийном пространстве обеспечивает лингвистов современным словарем лингвистических терминов; во-вторых, достижением проекта является методика конструирования, а также компьютерный инструментарий для реализации этой модели; в-третьих, тезаурус совместим с интеллектуальными системами обработки текстовой информации, в которых он может быть использован как база знаний и инструмент распознавания смысла.

Литература

1. *Никитина С. Е.* Тезаурус по теоретической и прикладной лингвистике. М., 1979.

Наиболее продуктивные и наиболее частотные аффиксальные модели лексики Словаря языка Пушкина*

Д. Жакетова, Е. Ф. Пирятинская

Московский государственный университет имени М. В. Ломоносова

Язык Пушкина, словообразование, аффиксальная модель, частота, продуктивность

Summary. This paper is devoted to the most productive and most frequent affixal models of Pushkin's Language Vocabulary. Analysis done contributes to a total research on author's language and, which is more important, to the whole system of Russian word-formation regularities.

В данной работе в качестве объекта анализа используется та часть слов из Словаря языка Пушкина, которая представлена простыми словами (словами с одним корнем — корневыми и аффиксальными производными). Их количество насчитывает 19349 лексем. (За пределами представляемого здесь анализа остались сложные — двух-, трех- и четырехкорневые слова. Всего — 1578 лексем). На базе материала корневых слов и аффиксальных производных, его структуризации и категоризации была создана база данных (БД) в оболочке Excel-2007, что и позволило нам обнаружить наиболее продуктивные и наиболее частые модели словообразования этого класса пушкинских слов.

Заметим, что для описания всех анализируемых слов в строке описания в БД было использовано по три позиции перед корнем и по восемь позиций после корня, включая постфикс. Строка описания также содержит информацию о номере лексемы, ее частеречной принадлежности и индексе ее частоты употребления в текстах А. С. Пушкина.

После ряда сортировок, проведенных в БД, были построены словари частоты употребления и словари продуктивности аффиксальных моделей пушкинских слов. Ниже представлены наиболее продуктивные аффиксальные модели простых слов «Словаря языка Пушкина» в наиболее абстрактном их отображении (т. е. с указанием только типа морфем):

№№ п/п	Абстрактная аффиксальная. модель	Продуктивность модели (кол-во лексем с данной афф. моделью)
1	R-F	3290
2	R-S-F	3125
3	Pr-R-S-S	2851
4	R-S-S-F	1435
5	R-S-S	1112
6	Pr-R-S-S-PF	1006
7	Pr-R-S-F	973
8	Pr-R-S-S-F	859
9	Pr-R-S	550
10	Pr-R-F	507
11	R-S-S-PF	353
12	Pr-R-S-S-S	348
13	R-S	327
14	R	281
15	R-S-S-S-F	276

* Настоящее исследование — составная часть НИР, выполняемой в лаборатории общей и компьютерной лексикологии и лексикографии МГУ по гранту РФФИ № 08–07–00435–а «Создание и первичный анализ лингвистически-ориентированной базы данных „Электронная энциклопедия языка Пушкина“ (1 очередь)».

Ниже приводятся результаты того же анализа, но сгруппированные по категориальной принадлежности лексем и с указанием конкретных аффиксов, т. е в виде конкретных аффиксальных моделей. Среди наиболее продуктивных словообразовательных моделей этого класса для имен существительных выделяются следующие:

Конкретная аффиксальная модель существительных	Продуктивность модели	Примеры слов из СЯПа (с частотой их употребления)
R-0	1925	*Рай* (34), *октябрь* (64)
R-*а*	666	*Белка* (15), *рана* (58)
R-*к-а*	242	*Подушка* (20), *беседка*(15)
R	234	*Так* (1475), *но*(3862)
R-*я*	140	*Неделя* (125), *баня* (31)
R-*о*	127	*Озеро* (37), *небо* (288)
R-*ок*-0	81	*Кусок* (23), *пирожок* (2)
R-*ец*-0	76	*Резец* (11), *телец* (3)
R-*и-я*	70	*Идиллия* (13), *репутация* (2)
R-*ик*-0	63	*Старик* (72), *шарик* (1)

Как видно, среди первых десяти самых продуктивных конкретных моделей существительных представлены только суффиксальные производные и голые корни. Среди глаголов выделяются следующие наиболее продуктивные модели:

Конкретная аффиксальная модель глаголов	Продуктивность модели	Примеры слов из СЯПа (с частотой их употребления)
R-*и-ть*	287	*Разорить* (24), *простить* (201)
R-*а-ть*	267	*Мечтать* (15), *играть* (178)
R-*и-ть-ся*	171	*Мучиться* (12), *ложиться* (30)
R-*ну-ть*	106	*Махнуть* (16), *лопнуть* (4)
R-*а-ть-ся*	105	*Кусаться* (2), *питаться* (12).
за-R-*а-ть*	86	*Закричать* (89), *замирать* (12)
про-R-*а-ть*	64	*Проведать* (9), *прогнать* (40)
на-R-*а-ть*	58	*Нападать* (22), *налетать* (4)
про-R-*и-ть*	34	*Проказить* (8), *проложить* (5)

Префиксальные модели словообразования оказываются более распространенными среди глаголов. Это объясняется тем фактом, что наличие или отсутствие приставки определяет вид — одну из важнейших категориальных характеристик глагола. Посткорневая часть обычно ограничивается двумя-тремя позициями для суффиксов и позицией для окончания, иногда добавляется постфикс. Модели с постфиксом в принципе дублируют такие же модели без него, например, R-*а-ть* и R-*а-ть-ся*. Среди имен прилагательных выявлены следующие наиболее продуктивные аффиксальные модели:

Конкретная аффиксальная модель прилагательных	Продуктивность модели	Примеры слов из СЯПа (с частотой их употребления)
R-*н-ый*	368	*Подобный* (103), *обычный* (11)
R-*ск-ий*	310	*Невский* (21), *адъюнкткий* (25)
R-*ый*	117	*Милый* (698), *подлый* (12)
R-*ов*-0	113	*Расинов* (3), *Миллеров* (3)
R-*ич-еск-ий*	69	*Анакреонтический*(2), *балтический* (2)
R-*ин*-0	44	*Ольгин* (4), *Мариин* (21)
без-R-*н-ый*	33	*Безобразный*(11), *беззаботный* (10)
R-*к-ий*	28	*Крепкий* (34), *дерзкий* (34)
R-*и-тель-н-ый*	26	*Разорительный* (3), *пленительный* (26)

Продуктивные словообразовательные модели имен прилагательных представляют собой, в основном, вариации типа R-s-f, где количество суффиксов варьируется от одного до трех (*крепкий, милостивейший*). Префиксальный способ словообразования достаточно распространен среди имен прилагательных лексики СЯПа. Такие модели, практически, повторяют бесприставочные аналоги (например, R-*н-ый* и *без*-R-*н-ый*).

Данные частотных словарей, полученные в результате сортировок основной БД, в обобщенном виде дают интересную картину. Как видно из нижеследующей таблицы, общий список наиболее продуктивных аффиксальных моделей похож на список наиболее частых моделей, но в отдельных моментах существенно с ним расходится. Практически совпадают в обоих таблицах по своей порядковой позиции такие модели, как: R-F (1–1), R-S-F (2–3), Pr-R-S-S (3–4), R-S-S (5–6), R-S-S-F (7–4), R-S-S-PF (11–12), R-S-S-S-F (15–15) и нек. др. Одновременно, обращает на себя внимание существенное расхождение в порядковой позиции продуктивности и частоты употребления такой абстрактной модели, как R. Если по частоте употребления эта модель на втором месте, то по продуктивности — во втором (!!!) десятке (14).

Вопрос о точном характере зависимости между частотой и продуктивностью аффиксальных моделей слов требует специального рассмотрения.

№№ п/п	Абстрактная аффиксальная модель	Общая частота употребления слов с данной аффиксальной моделью
1	R-F	126169
2	R	92329
3	R-S-F	48768
4	Pr-R-S-S	31889
5	R-S	27959
6	R-S-S	26700
7	R-S-S-F	17765
8	Pr-R-S	13927
9	Pr-R-S-F	10981
10	Pr-R-S-S-F	9559
11	Pr-R-F	8833
12	Pr-R-S-S-PF	7952
13	R-S-S-PF	7153
14	Pr-R-S-S-S	3770
15	R-S-S-S-F	2700

Если сравнить список наиболее продуктивных аффиксальных моделей словообразования лексики СЯПа со списком моделей, представленным в РГ-80 (раздел «Словообразование»), то можно сказать, что перечень продуктивных моделей отличается, но не настолько, чтобы нужно было бы говорить о «языке» Пушкина как о другом языке, нежели общелитературный русский. Различия в наборе моделей объясняются тем, что СЯП — это авторский словарь, где в словообразовательных моделях отображено личностное отношение к языку. Кроме того, пушкинские тексты, естественно, не покрывают столь же обширное смысловое пространство, как и тексты всего русского литературного языка, на материале которого, в основном, и базируются грамматические описания типа РГ-80. А поэтому пушкинская лексика не может быть столь же лексически и структурно разнообразна, как та лексика, которая кладется в основу упомянутых общеязыковых описаний. Различия, естественно, объясняются и тем фактом, что лексика СЯПа относится к другому историческому периоду, хотя и принято считать язык Пушкина основой русского литературного языка. Однако требуется систематическое сопоставление этих разных языковых стихий, что, естественно, требует одинаковости процедур отбора и описания их материала. Т. е. как и пушкинский, общеязыковой материал тоже должен извлекаться из некоторого представительного для всего русского языка корпуса текстов.

Лингво-методические возможности русско-белорусского параллельного корпуса текстов

А. В. Зубов

Минский государственный лингвистический университет (Беларусь)

proscien@mslu.by

Белорусский, корпус, параллельный, русский, тэггированный

Summary. In the report the author presents two groups of language parallel texts, that are created in Minsk state Linguistic University. The first group includes Russian-Belorussian texts of different styles. Russian-Belorussian educational texts belong to the second group. The author describes the common method that can be applied to tag multilingual parallel text and explains how the tagged texts can be used for scientific and educational purposes.

Особую специфическую разновидность корпусов текстов представляют корпусы параллельных текстов. Корпусом параллельных текстов называется множество текстов на каком-то одном языке и их переводов на один или несколько других языков.

С 2006 года в рамках государственной программы «Белорусский язык и литература» Минским государственным лингвистическим университетом совместно с Институтом языка и литературы им. Янки Купалы и Якуба Коласа Национальной Академии наук Республики Беларусь выполняется тема «Создание большого корпуса текстов белорусского языка и его использование для изучения белорусского языка и его связи с другими языками Европы». В процессе выполнения этой темы создается тэггированный корпус текстов белорусского языка размером в 1 млн. словоупотреблений и параллельный русско-белорусский корпус текстов, содержащий 300000 словоупотреблений.

В последний включены, в основном, тексты художественной литературы (Якуб Колас, Владимир Короткевич, Змитрок Бядуля, Иван Шамякин и др.), некоторые научные тексты и тексты деловой прозы.

Основной особенностью создаваемых параллельных корпусов текстов является то, что все их единицы являются тэггированными (размеченными). Иными словами, каждое словоупотребление текстов имеет специальные индексы (меты, тэги), свидетельствующие об их лингвистической специфике. Для этих целей использован стандарт тэггирования CES (Corpus Encoding Standard), который широко использовался при разработке европейских проектов MULTEX 135 и EAGLES (Expert Advisory Group on Language Engineering Standard) в сотрудничестве с американским партнером Vassar Cillege и французским партнером CNRS (CENTRE National de la Recherché Scientifique). При его использовании в оформлении информации к словоупотреблениям текста были внесены некоторые незначительные изменения.

Этот стандарт удобен также тем, что он специально создан для автоматического решения задач прикладной лингвистики, машинного перевода, лексикографии и т. п.

В соответствии с этим стандартом каждое словоупотребление белорусских и русских текстов получало набор определенных единых лексико-морфологических признаков. Так, для существительного указывались коды класса слова, одушевленность, число, падеж, личность, сокращение ли это или имя собственное. Для глагола — коды класса слова, вид, время, залог, переходность, спряжение, лицо, число, род. Аналогично имели свои коды и слова других классов слов.

В отличии от тэгов CES, в создаваемом параллельном корпусе каждое словоупотребление имело определенные структурные признаки. Для слов всех классов указывалось число слогов в словоупотреблении и место ударного слога в нем.

Тэггирование всех словоупотреблений белорусских и русских текстов проводилось в полуавтоматическом режиме.

В последние годы во многих вузах Республики Беларусь все большее число специальных и естественных дисциплин читаются на белорусском языке. В школах многие дисциплины также преподаются как на русском, так и на белорусском языках. Как в процессе преподавания таких дисциплин и создания соответствующих учебников и учебных пособий, так и при проведении научных исследований по сопоставительному изучению русского и белорусского языков неоценимую помощь может оказать параллельный тэггированный русско-белорусский корпус учебных текстов.

Созданием параллельного русско-белорусского корпуса учебных текстов с 2009 года занимается кафедра информатики и прикладной лингвистики Минского государственного лингвистического университета в рамках научной темы «Структурно-семантический анализ текстов с использованием методов корпусной лингвистики». Все словоупотребления таких учебных текстов также кодируется по описанному выше методу с использованием стандарта тэггирования CES.

Предварительный анализ нескольких пар школьных учебников по различным дисциплинам дал возможность выделить в них 12 видов информации: «теоретические темы», «главные выводы», «материал для повторения», «упражнения», «контрольные задания», «исторические сведения», «основные события и даты» и др.

Используя параллельные тэггированные корпусы текстов можно автоматически извлечь из них большое число информации, необходимой как для проведения сравнительно-сопоставительного изучения русского и белорусского языков, так и для обучения белорусскому языку. Так, с опорой на эти тексты можно:

1. Автоматически выделять в двух языках группы слов определенного словоизменения или словообразования.
2. Искать и выделять слова с определенными грамматическими характеристиками.
3. Выделять грамматические модели некоторых понятий и их эквивалентов в переводном языке.
4. Выделять структурные модели словосочетаний исходного и переводного языков.
5. Проводить сопоставительный анализ двух языков на синтаксическом уровне.
6. Автоматически строить переводные конкордансы.
7. Верифицировать значения лексических единиц, уже зафиксированных в белорусско-русских и русско-белорусских словарях, особенно в том, что касается идиом, метафор и терминологических выражений.
8. Выделять новые устойчивые словосочетания и идиоматические выражения, которые целесообразно вводить в существующие словари и др.

Если говорить о применении параллельных корпусов текстов в учебном процессе, то они позволяют оперативно решать следующие задачи:

1. Отбирать текстовые примеры для создания учебников и учебных пособий.
2. Создавать русско-белорусские и белорусско-русские словари по различным видам информации учебников и по всем материалам учебников.
3. Создавать двуязычные терминологические словари по различным дисциплинам.
4. Отбирать предложения и словосочетания определенных синтаксических структур.
5. Изучать стилистические особенности разных авторов.
6. Строить словари рифм.
7. Изучать ритмическую структуру строк, стихотворных текстов и др.

Марковские модели синтаксиса русского текста и анализ структуры множества их состояний

Е. А. Ильюшина, Д. А. Третьяков

Московский государственный университет имени М. В. Ломоносова

ilyushina@newmail.ru

Синтаксические связи, марковская модель, структура множества частей речи

Summary. The paper offers a model of Russian text syntax as a sequence of parts of speech and syntactic relations. The structure of variety of Markovian chain states is analyzed and stationarity of the final distribution is established.

Выполнен синтаксический разбор большой выборки русских текстов с помощью программного пакета Cognitive Dwarf, при этом в качестве единиц, составляющих текст, выбирались как части речи (ЧР), так и синтаксические отношения или связи (СО). Построены частотные словари ЧР и СО, а также матрицы условных вероятностей для линейной последовательности частей речи. Для синтаксических отношений линейной, а не древесной структурой является результат нестандартного алгоритма разбора, заложенный в пакете, что позволяет определить условные вероятности последовательности СО.

На первом этапе рассматривалась модель текста как последовательности частей речи, образующих простую однородную цепь Маркова. С помощью программы, реализующей алгоритм Уоршола, проведен анализ структуры множества состояний цепи, параметры модели (значения вероятностей перехода) оценивались для каждого конкретного текста по частотам синтаксического разбора. Установлено, что множество ЧР каждого текста образует один класс эквивалентности, в который входят практически все части речи, что свидетельствует об одинаковой синтаксической организации любого текста в данной модели. Тем самым доказывается, что для всех рассмотренных текстов предельное распределение является стационарным (это предположение уже заложено в алгоритме программы синтаксического анализатора).

Далее строится марковская модель текста, в которой состояниями являются синтаксические отношения. СО в тексте устроены не линейно, т. е. слова, связанные СО, могут стоять далеко друг от друга, но результат синтаксического разбора в программном пакете Dwarf записывается в виде линейной последовательности синтаксических отношений. Матрицы вероятностей перехода размера 324×324 являются существенно разреженными, что затрудняет анализ структуры множества состояний цепи. Процедуры сглаживания не всегда дают правильную оценку вероятности, что может изменить свойства исходной системы синтаксических связей. Здесь более уместным представляется использование сингулярного разложения (латентно-семантического анализа) для устранения нулей в матрице вероятностей перехода. Эта задача пока не решена.

В работе получены также частотные словари частей речи и синтаксических связей, проведен их статистический анализ и сравнение для текстов различных авторов, найдены энтропии распределений ЧР (Тургенев, Толстой, Достоевский, Лесков, Шолохов).

Литература

1. *Соколов Г. А., Чистякова Н. А.* Управляемые цепи Маркова в экономике. М., 2005.
2. *Dagan I., Markus S., Markovitch S.* Contextual word similarity and estimation from sparse data // Proceedings of the 31st Annual Meeting of the ***Association for Computational Linguistics (ACL)***. Colombus, Ohio, ***1993.*** P. 164–171.
3. http://cs.isa.ru:10000/dwarf/doc.htm.
4. www.statsoft.ru.

Особенности тематической организации интернет-эгоисторий в ранговом распределении Ципфа

А. С. Инфантьева

Кемеровский государственный университет

nomad5@mail.ru

Интернет-эгоистория, закон Ципфа, тематическая организация, тематическое поле, покрываемость текста

Summary. The report is devoted to research of the thematic organisation of Russian self-histories in Internet. Reserch is based on words` rank distributions of the Zypf`s law. The material is normal and pathological self-historic texts of Russian Inernet-diaries.

1. В настоящее время для лингвистики характерен возрастающий интерес к изучению новых дискурсивных практик, одной из которых является интернет-эгоистория. В отечественные филологические исследования термин «эгоистория» был введен Ю. Л. Троицким в значении сюжетного повествования о собственной жизни, имеющего свою интригу и не совпадающего ни с автобиографией, ни с автопсихологией [4: 58]. Одной из разновидностей рассказов о себе является текст личного интернет-дневника — так называемого блога.

2. Данное исследование направлено на выявление через ранговое распределение Ципфа закономерностей тематической организации интернет-эгоисторий, созданных носителями русского языка. Тематическая организация — это распределение тем в целостном тексте. Под темой (тематическим полем) понимается «совокупность данных, актуальных для определенной темы и являющихся фоном для ее раскрытия» [6: 10]. Под индексами темы отдельного текста могут пониматься наиболее частотные существительные, интенсивность использования которых выражается в долях покрываемости всего текста. Для ее вычисления используется метод рангового распределения Ципфа.

3. Согласно закону Ципфа, относительная частота слова (F) в тексте обратно пропорциональна рангу слова (n), причем «выполнение данного рангового распределения — это признак „правильности“ данного текста» [1: 9]. Принципиальные отклонения от нормы в ранговом распределении слов текста свидетельствуют об иных закономерностях организации текста. Интенсивность использования наиболее частотных существительных равна покрываемости текста участком словаря (10 первых существительных в частотном списке) и обозначается $Z_{n.}$ Таким образом, Z_n вычисляется как $F(x_{n1}) + F(x_{n2}) \ldots + F(x_n)$, где x — слово с рангом n. Час-

тота слова в тексте, обозначаемая $F(x_n)$, является отношением количества словоупотреблений 1 слова к объему текста в словоупотреблениях. Материалом исследования является пять полных нормальных (непатологичных) интернет-эгоисторий. В качестве текстов контрольной группы было выбрано пять интернет-эгоисторий наркозависимых личностей.

3.1. Z_n для беллетристических, эпистолярных и разговорных текстов удерживается в интервале 0.93 ÷ 1.45% [3: 81], а в интернет-эгоисториях Z_n варьируется 4.5 ÷ 27.4%. Такой разброс обусловлен интенсивностью записей в интернет-эгоистории: чем чаще добавляются записи, тем вероятней доминирование в тексте одной и той же темы. В интернет-эгоистории со средней интенсивностью записей 1\20.6 дней Z_n = 4.5%; при интенсивности 1\1.07 дней Z_n = 27.4%. Прямая зависимость Z_n от интенсивности следует из особенностей интернет-эгоистории как типа дискурса: текст посвящен субъекту, тематически ограничен повествованием о его жизни и дискретен, т. е. состоит из добавляемых записей малых по объему. При высокой интенсивности записей, внимание субъекта сосредоточивается на определенной теме, волнующей его, поэтому Z_n в таком тексте превышает 15% — максимально возможный Z_n в нормальных традиционных текстах [3: 81].

3.2. Тексты интернет-эгоисторий наследуют отдельные свойства малых нарративных жанров. Как и в ранее исследованных малых текстах с объемом слов порядка $N=10^3$ и $N = 10^4$ [1: 15], у частотных слов интернет-эгоисторий с объемом $N=10^1$ и $N=10^2$ достаточно высокий индекс покрываемости, что обусловлено именно малым объемом текста.

3.3. Ранговое распределение в контрольной группе текстов характеризуется еще более высоким Z_n, что связано с особой ролью существительных при психопатологии речи: «По мере утяжеления психического заболевания доля именных единиц в патологическом тексте имеет тенденцию к увеличению» [3: 82]. Максимальный Z_n = 31.1% в данном случае не показателен, так как близок к максимальному Z_n в нормальных интернет-эгоисториях (27.4%). Минимальный Z_n = 14.8% является очень высоким, что говорит об узкой тематической направленности интернет-эгоисторий наркозависимых и подтверждает тенденцию к сведению «истории жизни к истории наркотизации» [5: 59] и нескольких ведущих мотивов к одному — влечению к психоактивным веществам [2: 327].

4. Ранговое распределение слов свидетельствует, что тематическая организация интернет-эгоисторий характеризуется более высокой степенью покрываемости текста частотными существительными: их максимальный показатель покрываемости в два раза превышает аналогичный, описывающий ранговое распределение слов в традиционных нарративных малых жанрах. Такое повышение индекса покрываемости сопровождается интенсивностью добавляемых записей и выдвижением на первый план какой-либо определенной темы. На этом фоне интернет-эгистории наркозависимых характеризуются максимально высоким Z_n, отличным от показателей нормальных текстов, что демонстрирует зацикленность субъекта на доминирующей теме — влечении к психоактивным веществам.

Литература

1. *Арапов М. В., Ефимова Е. Н., Шрейдер Ю. А.* О ***смысле ранговых распределений***. Научная и техническая информация. Серия 2. № 1. С. 9–20.
2. *Елшанский С. П.* Семантика внутреннего восприятия при зависимостях от психоактивных веществ (на модели опийной наркомании). М., 2004.
3. *Пашковский В. Э., Пиотровская В. Р., Пиотровский Р. Г.* Психиатрическая лингвистика. М., 2009.
4. *Троицкий Ю. Л.* Эгоистория // Дискурс. 1996. № 1.
5. *Фоломеева Н. М., Шурышина И. И., Новикова Н. А., Чешнева Т. В.* Наркомания как форма девиантного поведения. М., 1997.
6. *Rosenthal G.* The Narrated Life Story: On the Interrelation Between Experience, Memory and Narration // Narrative, Memory & Knowledge: Representations, Aesthetics, Contexts. Huddersfield, 2006 // http://www2.hud.ac.uk/hhs/nme/books/2006/Chapter_1_-_Gabriele_Rosenthal.pdf.

Средства семантико-синтаксической обработки в системе русского синтеза RussLan

М. И. Канович, З. М. Шаляпина

Институт востоковедения РАН (Москва)

zmshal@yandex.ru

Русский синтез, семантико-синтаксическая обработка, сущностный подход

Summary. *Means of semantico-syntactic processing in the RussLan system of Russian generation.* The object of consideration is the formalism of R-attributes developed within the RussLan system of Russian generation for representing structural relations and processing linguistic rules associated with them, as attributes of the corresponding linguistic entities. It can be used for specifying the markers of syntactic arguments by the predicate valency frames, computing such frames for occasional lexemes, verifying word-form definitions, modifying them in case of conflicts, performing local syntactic transformations, etc.

Система русского синтеза RussLan [2]; [3] создается в рамках экспериментального комплекса ЯРАП для лингвистических исследований по японско-русскому автоматическому переводу и обеспечивает все этапы синтеза для русского языка, начиная с семантико-синтаксического и кончая морфологическим, так что на ее выходе выдается цепочка реальных русских словоформ. В своем функционировании она независима от входного языка перевода и от других компонентов комплекса: связь между ними осуществляется только через входные данные, представляемые в виде текстовых файлов. Реализована она на языке программирования Turbo Pascal.

Система носит экспериментальный характер, и объем ее лингвистического обеспечения ограничен задачами отладки формальных средств и механизмов синтеза. Ее теоретико-лингвистической основой является сущностный подход к языку, который ставит в центр описания языка и текста элементарные лингвистические сущности (конкретные лексические и грамматические морфемы и обобщающие их категории), а все отношения и правила трактует как атрибуты таких сущностей.

Семантико-синтаксический (**СемСинт**-) синтез понимается в системе RussLan как этап, промежуточный между семантическим синтезом (при переводе — межъязыковым переходом) и морфологическим синтезом в узком смысле — построением синтетических словоформ по их лексико-морфологическим (**ЛМ**-) определениям. Предварительные варианты ЛМ-определений имеются уже в составе входного представления текста, сформированного семантическими или переводческими процедурами. Но такие процедуры в общем случае не предполагают учета специфики выходного языка, и формируемые ими ЛМ-определения словоформ могут не отвечать каким-то из его требований. СемСинт-синтез должен устранить лакуны и некорректности входа и свести его к цепочке полных, непротиворечивых и синтетически реализуемых ЛМ-определений словоформ. В этом смысле он предстает не как переход между глубинным и поверхностным уровнями Синт-представления текста, но как верификация и коррекция представления текста в пределах одного и того же уровня.

Это позволяет определить и реализовать СемСинт-синтез как принципиально рекурсивный процесс и на всех его этапах применять для представления текста принципиально одни и те же формальные средства, так что его вход и выход различаются не форматом, но лишь конкретным составом входящих в них определений словоформ и степенью эксплицированности разных компонентов этих определений. Входные определения в общем случае содержат струк-

турно-синтаксические атрибуты определяемых словоформ, задающие их лексико-синтаксический (**ЛС**-) контекст, морфологические же категории могут отсутствовать либо, как и лексические единицы, не отвечать (или неполностью отвечать) заданному ЛС-контексту. На выходе ЛС-атрибуты отсутствуют, зато лексические и морфологические компоненты задают словоформу корректно и однозначно.

Для осуществления СемСин-синтеза в этом его понимании в системе RussLan разработан аппарат так наз. R-отсылок [1]. Они определяются как реляционные атрибуты лингвистических сущностей, задающие языковые или текстовые функции этих сущностей относительно некоторых других сущностей. Для текстовых сущностей в виде таких атрибутов определяются, в частности, элементы их ЛС-контекста, связанные с ними отношениями синтаксических зависимостей или кореферентности, для языковых — правила учета таких элементов ЛС-контекста при синтезе.

Средствами этого аппарата в системе RussLan решаются следующие задачи:

– проверка корректности морфологических и синтаксических атрибутов обрабатываемых единиц и модификация некорректных сочетаний таких атрибутов с учетом типа обнаруженных конфликтов, в том числе:
  - коррекция номера зависимости по ее семантической интерпретации;
  - снятие конфликта между семантическими требованиями к актанту и его собственной семантикой;
  - морфолого-синтаксическое перефразирование;
  - коррекция морфологических характеристик;
– оформление актантов с учетом модели управления управляющего предиката, включая учет таких явлений, как:
  - влияние синтаксических «хозяев» на оформление «слуг»;
  - влияние «слуг» на оформление их «хозяев»;
  - влияние позиции «слуги» на его оформление;
  - влияние дистантного ЛС-контекста;
  - влияние числительных на их соседей в именной группе;
  - ограничения, связанные с лексической обусловленностью, включая «двустороннюю» лексическую обусловленность в предложных конструкциях;
– вычисление синтаксической информации к окказионализмам и прочим лексемам, отсутствующим в словаре, по их морфологическому составу;
– развертывание фразеологии;
– аналитическое формообразование;
– синтаксическая обработка по умолчанию

и т. п. Все перечисленные виды СемСинт-обработки могут быть проиллюстрированы на конкретных русских примерах.

В целом аппарат R-отсылок может рассматриваться как многофункциональное формальное средство, позволяющее решать широкий спектр задач семантико-синтаксического синтеза. Существенно также то, что этот аппарат не имеет жесткой привязки только к одному языку, но применим и для других языков, сходных с русским по тем своим типологическим характеристикам, которые учтены в реализуемых им механизмах умолчаний. Это прежде всего соотношения, существующие в языке между структурой синтаксических зависимостей и связей кореферентности, с одной стороны, и линейной структурой текста, с другой (например, препозитивное размещение определений, использование в качестве служебных показателей предлогов, а не послелогов и т. п.). Для языков соответствующей типологии разработанный аппарат может использоваться как универсальное средство семантико-синтаксического синтеза.

Литература

1. *Канович М. И., Шаляпина З. М.* Аппарат R-отсылок как универсальное средство синтаксического синтеза (на опыте разработки системы русского синтеза RussLan) // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции «Диалог 2006». М., 2006. С. 207–213.
2. *Шаляпина З. М., Борисова Е. Г., Канович М. И. и др.* Проблемы русского лексико-синтаксического синтеза при сущностном подходе к языку // Русский язык: исторические судьбы и современность. Международный конгресс исследователей русского языка. М., 2001. С. 427–428.
3. *Shalyapina Z. M., Kanovich M. I., et al.* RUSSLAN: A System of Russian Language Generation // Investigations into Formal Slavic Linguistics. Contributions of FDSL IV. Part I. Frankfurt am Main et al., 2003. P. 385–403.

Разделение омофонов с использованием лексико-синтаксической информации

Г. Е. Кедрова, С. Б. Потемкин

Московский государственный университет имени М. В. Ломоносова

kedr@philol.msu.ru, potemkin@philol.msu.ru

Транскрипция, распознавание речи, синтагма, синтаксический анализ, критический путь

Summary. Homophone disambiguation using lexical and syntactic information is presented. An isolated part of the Russian sentence is coded as a sequence of transcription symbols. Analysis of the sequence with the use of a special lexicon provides an oriented net of possible words of the sentence. Word instances are disambiguated using concordance and syntactic information.

Одним из основных методов современных систем распознавания речи является использование как можно большего объема неакустической информации, особенно, информации более высоких уровней, т. е. семантической и прагматической [3]. В основном используются два базовых подхода: а) разработка более совершенной фонетической системы, состоящей из контекстных вариантов фонем; б) адаптация моделей обучения, учитывающих высшие уровни языка. Применение каждого из подходов в отдельности повышает эффективность распознавания по сравнению с существующими методами, а комбинация обоих характеризуется максимальной эффективностью [4]. Задача разделения омофонов важна, прежде всего, при распознавании слитной звучащей речи. При ее решении активно используются разные системы **автоматического транскрибирования**, в частности может использоваться автоматический транскриптор, предназначенный для синтеза речи, разработанный в МГУ [2]. В отличие от синтеза речи, когда транскрипция должна быть максимально подробной, при распознавании допустима и даже желательна упрощенная транскрипция. Мы используем простой транскриптор для русского языка [6], в котором произношение русского слова кодируется в основном соответствующими русскими буквами (плюс *j, l, t, h*), что упрощает восприятие и программирование алгоритма. Транскриптор производит подстановки символов в соответствии с правилами, вида *ъе → je, ъё → jё, ъю → jю*, … Учитывается начало и конец слова (знак #), а также ударная гласная (знак \).

Фонетическое кодирование и распознавание слитной фразы. На основе Грамматического словаря русского языка А. А. Зализняка [1] был составлен **словарь словоформ и их транскрипций**. Для отрезка звукового сигнала, ограниченного паузами или иными маркерами строится его фонетическая запись. Транскрибирование предложений *На бал кони ходят* <-> *На балконе ходят* дает: *набалканихади.* Далее выполняется **алгоритм выделения фонетических фрагментов**: а) отделяем начальную последовательность из i букв и ищем эту последовательность в словаре транскрипций; b) если поиск успешен, запоминаем длину последовательности. Таким образом, выделяется несколько слов разной длины — кандидатов на первое слово предложения. с) для каждого слова — кандидата выполняем пп. а) — b) для поиска следующего за ним слова — кандидата, начиная с буквы i + 1. Когда все кандидаты выписаны, просматриваем список с конца (с последнего слова) и отбираем только те слова, у которых конец слова совпадает с концом предложения. Для каждого из этих слов находим предыдущее; и так до начала предложения.

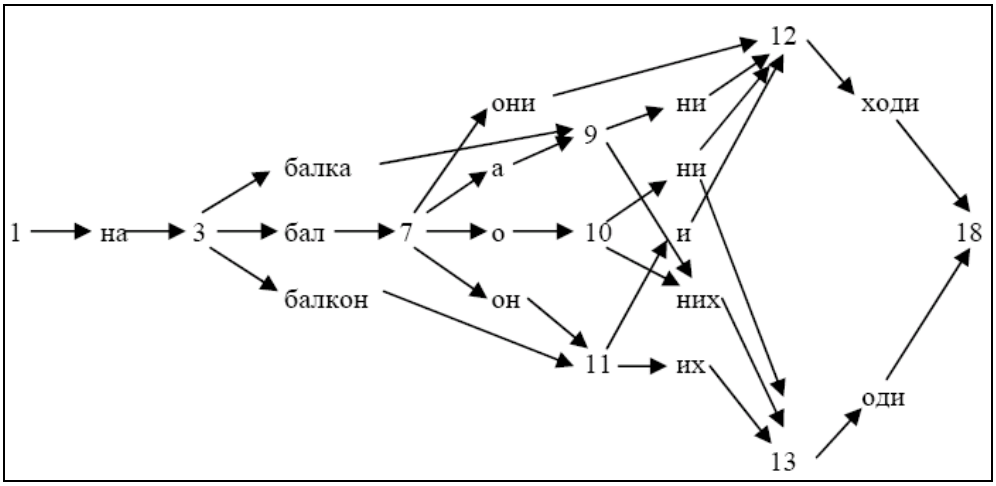


Рис. 1. Граф связей между выделенными фрагментами транскрипции предложения (ветви). Также показаны позиции начала и конца каждого фрагмента (вершины).

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.

Собственно говоря, количественный анализ текста предполагает расчет ряда некоторых квантитативных характеристик корпуса текстов [1]:

N — объем текста — число лексических единиц (ЛЕ, ЛЕ = словоупотребление, словоформа или лексема) в тексте.

L — число ЛЕ в тексте, которые встретились в тексте хотя бы один раз.

Lf_1 — ЛЕ, которые встретились в тексте только один раз.

Lf_k — число ЛЕ, которые встретились в тексте с частотой больше одного раза.

r — ранг ЛЕ. Ранг ЛЕ может измеряться следущим образом: а) по частоте встречаемости в тексте (или фрагменте) — самая частотная ЛЕ имеет ранг равный 1 и далее ранг r увеличивается по мере уменьшения частоты встречаемости ЛЕ в тексте (ЛЕ имеющие одинаковую частоту имеют и равные ранги); б) по длине слова (напр., число букв в ЛЕ); в) число значений ЛЕ (по толковым словарям).

Lr_1 — максимальная частотность ЛЕ.

F_i — абсолютная частота ЛЕ.

F^*_i — накопленная абсолютная частота ЛЕ = сумме частоты данной ЛЕ и всем предшествующих абсолютных частот ЛЕ.

f_i — относительная частота ЛЕ = отношению абсолютной частоты ЛЕ к объему текста N.

f^*_i — накопленная относительная частота ЛЕ = отношению накопленной абсолютной частоты ЛЕ к объему текста N.

H_i — удельная энтропия ЛЕ = $-f_i \log f_i$.

H^*_k — накопленная энтропия текста, равная сумме удельных энтропий.

C — индекс дистрибуции (чем, эта величина больше, тем богаче словарь текста) $C = (f_{r1}^2 + L^2)^{\wedge 1/2}$.

I_i — индекс итерации (индекс повторения ЛЕ в замкнутом тексте) $I_i = N / L$.

I_e — индекс исключительности (специфичности) лексики $I_e = 20*Lf_1 / N$

P — индекс предсказуемости (чем P меньше, тем привлекательнее текст) $P = 100 – (Lf_1*100) / N$.

I_q — индекс плотности текста. Пропорционален числу повторяющихся ЛЕ в тексте и объема текста N. (чем богаче тематика, тем выше I_q, чем однообразнее тема, тем I_q ниже)

n — число абзацев в тексте.

m — число абзацев текста, в которых встретилась ЛЕ.

$K_i = F_i * m / (N * n)$ — коэффициент важности ЛЕ.

I_{ext} — объем экстенсивности словаря текста. Пропорционален широте лексики, разнообразия выражения.

I_f — индекс стереотипности текста. Длина интервала средней части повторяющихся ЛЕ. Если I_f больше, то главное не форма, а содержание (для беглового чтения, нестилизовано, спонтанная речь). I_f меньше у художественных текстов, беллетристики.

При положительной установке длина и глубина предложений, количество сложных предложений больше, чем при отрицательной установке.

Закон Ципфа: $F_i = (1 / 10)* N *(1 / r)$.

Закон Ципфа в общем виде: $F_i =p* N *(r**(-b))$.

Закон Ципфа-Мандельброта: $F_i =p* N *(r+v)**(-b)$.

** — возведение в степень; p, v, b — параметры распределения: p — коэффициент отдельной частоты наиболее частой ЛЕ, v — поправочный коэффициент частых ЛЕ, b — коэффициент лексического богатства текста, N — число ЛЕ в тексте, r — ранг ЛЕ.

Ципф показал, что распределение ЛЕ в тексте подчиняется закону: если к какому-либо достаточно большому тексту составить список всех встретившихся в нем ЛЕ, затем расположить их в порядке убывания их частоты (встречаемости в данном тексте) и пронумеровать в порядке от 1 (порядковый номер наиболее часто встречаемой ЛЕ) до N, то для любой ЛЕ произведение ее порядкового номера (ранга) в таком списке и частоты ее встречаемости в тексте будет величиной постоянной, имеющей примерно одинаковое значение для любой ЛЕ из списка.

Литература

1. *Бектаев К. Б.* Статистико-информационная типология тюркского текста. Алма-Ата, 1978.
2. Записки Тартуского университета. Квантитативная лингвистика и автоматический анализ текстов. Тарту, 1985; 1986. Вып. 745; 1989. Вып. 872.
3. *Засорина Л. Н.* Введение в структурную лингвистику. М., 1974.
4. *Зубов А. В., Зубова И. И.* Информационные технологии в лингвистике. М., 2004.
5. *Зубова И. И.* Информационные технологии в лингвистике. Минск, 2001.
6. *Пиотровский Р. Г.* О некоторых стилистических категориях // Вопросы языкознания. 1954. № 1. С. 55–68.
7. *Пиотровский Р. Г., Бектаев К. Б., Пиотровская А. А.* Математическая лингвистика. М., 1977.

О неоднородности количественных характеристик авторского стиля в романе «Тихий Дон»

А. Г. Макаров, А. А. Поликарпов

Московский государственный университет имени М. В. Ломоносова

anatpoli@mail.ru

Дискуссия вокруг проблемы авторства романа «Тихий Дон» (ТД) продолжается вот уже более трех десятилетий [3]; [4]. Текстологический анализ романа показывает крайнюю неоднородность художественного текста (провалы в исторической достоверности, анахронизмы, резкие изменения стиля, разрывы в повествовании и т. д.), что может свидетельствовать о сложной предыстории текста, участии в его создании нескольких авторов [2], в том числе самого Шолохова — лишь как соавтора. Однако окончательное решение вопроса о возможном авторстве романа с помощью лишь традиционного текстологического и литературоведческого анализа оказалось весьма трудным в силу недостатка или неоднозначности получаемой при этом информации. Требуется разработка новых подходов и методов в исследовании и разрешении проблемы авторства литературного произведения.

Попытки количественного лингвистического анализа текста ТД известны давно, однако уже первые работы [7]; [1]; [3: 183–194]; [5] показали существование на этом пути значительных трудностей, связанных с отсутствием отработанных общепризнанных методик и однозначной интерпретации получаемого результата. Наиболее успешной можно признать работу [6], в ходе которой в тексте романа подсчитывалась частота употребления служебных слов. В тексте ТД авторы обнаружили значительные вариации этих частотных характеристик в различных его частях, что могло быть связано с наличием у текста нескольких авторов и сосуществованием в нем авторского и соавторского слоев.

В настоящей работе была предпринята попытка проверить однородность распределения количественных характеристик стиля в разных частях романа на примере объемов прямой речи персонажей в различных частях текста «Тихого Дона». Для подсчета выбранной характеристики мы использовали авторское разбиение текста на части (с 1-й по 8-ю: ТД1–ТД8), при этом текст 6-й части мы разбили на две половины (ТД6–1: 1–29 гл.; ТД6–2: 30–65 гл.), что соответствовало двум этапам публикации 6-й части (в 1929 г. и в 1932 г.). Такое разделение текста вполне оправдано, поскольку при текстологическом анализе 6-й части именно с момента возобновления публикации романа в 1932 г. обнаруживаются многочисленные нестыковки и противоречия. Одновременно нами также определялась доля прямой речи в текстах других произведений М. А. Шолохова: в «Поднятой целине» (1-я кн. — ПЦ1; 2-я кн. — ПЦ2) и «Они сражались за Родину» (ОСР). Результаты приведены на рис. 1.

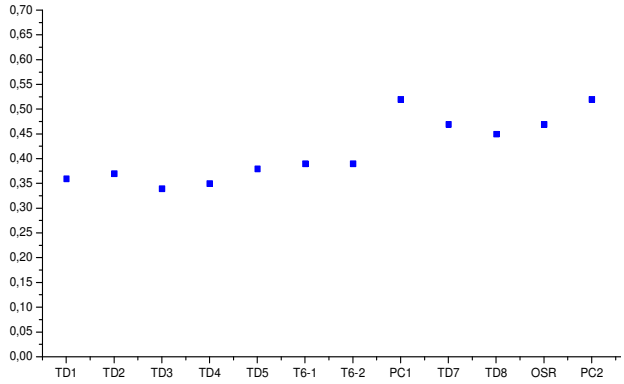


Рис. 1. Доля прямой речи в текстах Михаила Шолохова.

Хорошо видны существенные отличия объемов прямой речи в текстах первых шести частей ТД и в последних двух частях.

Для первого интервала мы видим близкие значения доли прямой речи в тексте (35–38%). Заметим, что все эти тексты были опубликованы Шолоховым в короткий промежуток времени с 1928 по 1932 гг. Для сравнения мы определили долю прямой речи в некоторых произведениях Федора Крюкова, одного из предполагаемых претендентов на авторство ТД: в рассказе «Казачка» — 44%; «В родных местах» — 32%; Зыбь — 27%; «На речке лазоревой» — 41%; «Картинки школьной жизни» — 32%; «Неопалимая купина» — 38%. Можно заключить, что исследованный параметр (доля прямой речи в художественном тексте) достаточно устойчива на протяжении первого текстового интервала ТД и не сильно отличается от характеристики текстов Ф. Д. Крюкова, у которого мы наблюдаем определенное постоянство доли прямой речи в его произведениях — примерно 35% от общего объема текста. Для очерков и коротких рассказов это значение возрастало до величины примерно 40%. Причем в более значительных по объему произведениях, в которых художественный материал, очевидно, прошел более глубокую и основательную авторскую обработку, мы отмечаем соответствующее снижение доли прямой речи.

Если теперь рассмотреть тексты седьмой и восьмой частей ТД, то видно, что ситуация меняется, объем прямой речи заметно возрастает (до 45%). Такие же и даже большие значения объемов прямой речи (50–55%) мы встречаем в 1-й книге «Поднятой целины» (1932 г.) и в послевоенных текстах Шолохова — «Поднятая целина», кн. 2-я (1960 г.) и «Они сражались за Родину». Одним из первых такое изменение стиля шолоховских произведений обнаружил еще в 2000 г. ростовский исследователь А. Венков [2: 30–34], который отметил появление в тексте, начиная с конца 6-й части, длинных монологов отдельных персонажей.

Исследование относительного вклада в прямую речь шолоховских текстов «диалогов» разного объема показало, что увеличение общих объемов прямой речи в последних частях ТД, в ПЦ и в ОСР происходит за счет резкого возрастания доли «„длинных“ диалогов»-монологов персонажей: с нескольких процентов до 10–15% и более. Их относительный вклад в общий объем прямой речи представлен на рис. 2.

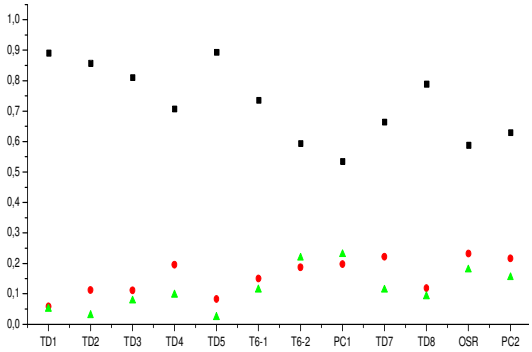


Рис. 2. Доля коротких (число букв в непрерывном фрагменте «прямой речи» менее 200 — ■), среднего объема (от 200 до 350 букв — ●) и длинных «диалогов» (▲) в текстах Крюкова, Шолохова и в «Тихом Доне»/

Мы видим, что в тексте «Тихого Дона» имеет место четко выраженная тенденция: к шестой части романа доля коротких диалогов заметно снижается при одновременном увеличении диалогов среднего и большого объемов. Начиная с середины 6-й части ТД мы можем сказать, что в текстах и «Тихого Дона», и обеих книг «Поднятой целины», и в главах романа «Они сражались за Родину» не менее трети объема прямой речи приходится на «протяженные» диалоги, а сама общая доля прямой речи в тексте возрастает с 32–35% до 45–50% и более (что соответствует появлению в тексте пространных рассуждений и баек в духе деда Щукаря).

Таким образом, проведенное изучение шолоховских текстов подтверждает полученные другими авторами наблюдения о заметном изменении количественных характеристик авторского стиля (в данном случае — доли прямой речи) в последних частях романа «Тихий Дон», а также в последующих произведениях М. А. Шолохова — в романе «Поднятая целина» и в романе «Они сражались за Родину».

Литература

1. *Аксенова Л. З. (Сова), Вертель Е. В.* О скандинавской версии авторства «Тихого Дона» // «Вопросы литературы». 1991. Февраль. С. 68–81.
2. *Венков А. В.* «Тихий Дон»: источниковая база и проблема авторства. Ростов-на-Дону, 2000.
3. Загадки и тайны «Тихого Дона». Итоги независимых исследований текста романа. 1974–1994. Самара, 1996.
4. *Макаров А. Г., Макарова С. Э.* Цветок-Татарник. В поисках автора «Тихого Дона»: от Михаила Шолохова к Федору Крюкову. М., 2001.
5. *Марусенко М. А., Бессонов Б. А., Богданов Л. М., Аникин М. А., Мясоедова Н. Е.* Темные воды «Тихого Дона» // В поисках потерянного автора. СПб., 2001.
6. *Фоменко В. П., Фоменко Т. Г.* Авторский инвариант русских литературных текстов. Приложение. Кто был автором «Тихого Дона»? // Новая Хронология Руси и Рима. Т. 2. М., 1995.
7. *Хьетсо Г., Густавссон С., Бекман Б., Гил С.* Кто написал «Тихий Дон»? М., 1989.

Анализ количественных характеристик авторского стиля романа «Тихий Дон» и его соотношение с другими текстами М. А. Шолохова на основе иерархической кластеризации

А. Г. Макаров, А. А. Поликарпов

Московский государственный университет имени М. В. Ломоносова

anatpoli@mail.ru

В. В. Поддубный, О. Г. Шевелев

Томский государственный университет

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

Как показали ранее проведенные эксперименты на текстах второй половины XX века, метод иерархической кластеризации текстов, основанный на использовании гипергеометрического критерия и интегральной меры рассогласования, дает осмысленную группировку текстов по различным стилям, в том числе по авторским [1]; [3]; [5]. Одними из наиболее успешно проявивших себя признаков являются частоты появления 55 служебных слов из списка А. Т. Фоменко.

Для анализа авторства Тихого Дона нами для начала был проведен ряд экспериментов по кластеризации с помощью данного метода и данных признаков текстов самого «Тихого Дона», других произведений М. А. Шолохова, а также произведений авторов, предположительно связанных с романом «Тихий Дон». В качестве первого шага исследования мы провели анализ текста «Тихого Дона» и других произведений М. А. Шолохова на «макроуровне», то есть выбрали

разбиение текстов на крупные блоки: для «Тихого Дона», как уже упоминалось выше, за основу были взяты тексты восьми частей романа, причем 6-й часть разбивалась на две половины — гл. 1–29 и гл. 30–65 (ТД-1 — ТД-5, ТД-6–1,ТД-6–2, ТД-7,ТД-8). К ним были добавлены тексты 1–й и 2–й книг «Поднятой целины» (ПЦ1, ПЦ2), «Они сражались за Родину» (ОСР) и послевоенный рассказ М. Шолохова «Судьба человека» (СЧ). «Донские рассказы», в силу жанровых отличий, а также незначительности объемов текста отдельных рассказов для надежной статистической обработки стилевых характеристик, на данном этапе исследования были из анализа исключены.

В качестве исследуемых признаков нами были выбраны служебные слова (предлоги, союзы, частицы) по списку В. П. Фоменко и Т. Г. Фоменко [4]. На рис. 1. представлено полученное дерево кластеризации для вышеупомянутых признаков для текста «Тихого Дона», разделенного на девять частей. На рис. 2 представлены результаты анализа текстов для случая, когда к исследуемому корпусу текстов были добавлены тексты ПЦ1, ПЦ2 и ОСР.

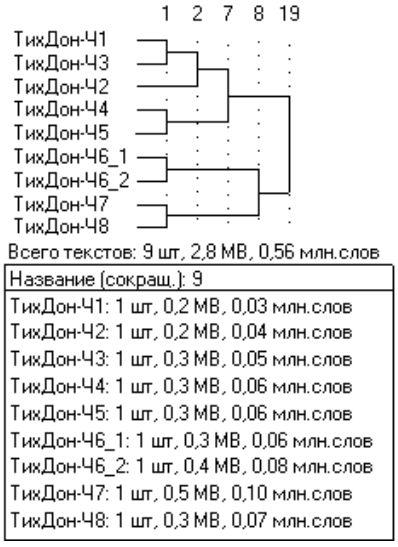


Рис. 1.

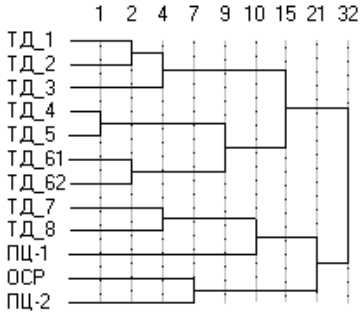


Рис. 2.

Полученные результаты количественной кластеризации выявляют характерную особенность: тексты частей ТД объединяются, группируются в **четыре** относительно самостоятельные кластера: первый кластер включает в себя первые три части ТД, второй — 4-ю и 5-ю части, третий — первую и вторую половины 6-й части. Наконец, четвертый кластер объединяет 7-ю и 8-ю части ТД. При этом наблюдается сближение образовавшихся кластеров в две группы, объединяющие *первую* и *вторую* пару кластеров.

Анализ результатов кластеризации по всем указанным текстам М. А. Шолохова (ТД, ПЦ и ОСР) показал добавление к наблюдаемым четырем кластерам нового, в который вошли «послевоенные» произведения М. А. Шолохова (ПЦ2 и ОСР). Текст ПЦ1 подсоединился к четвертому кластеру текстов ТД (ТД-7 и ТД-8). Заметим, что три эти текста создавались и были публикованы примерно в одно время начиная с 1932 по 1940 гг. В конечном счете, по результатам анализа кластеризации можно констатировать группирование исследованных текстов в три ярко выраженные стилевые группы. **Первая группа** включает тексты ТД первых шести частей романа. При этом имеет место выраженное структурирование и внутри этой группы: в отдельные подгруппы-кластеры устойчиво объединяются тексты первых трех частей ТД, четвертой пятой частей ТД и, наконец, первая и вторая половины шестой части ТД. Отметим здесь, что все эти тексты были опубликованы в течение короткого промежутка времени в 1928–1932 гг. **Вторая группа** включает в себя тексты Шолохова 30-х годов: кластер из последних двух частей романа — ТД-7 (1937 г.) и ТД-8 (1940 г.) и примыкающей к ней, но заметно отличной по характеристикам, 1-й кн. ПЦ (1932 г.). **Третья группа** образуется послевоенными произведениями М. Шолохова (2-я кн. ПЦ, ОСР) и резко выделяется по частотным характеристикам из остальных текстов.

Полученные результаты оказываются весьма интересными. Характер частотных характеристик подтверждает существенную неоднородность текста романа, которая была в свое время выявлена при текстологическом анализе ТД, когда именно с середины 6-й части ТД отмечалось нарастающее соавторское вмешательство в текст, при котором нарушалась и хронология описываемых событий, и историческая точность и достоверность, менялись язык и мировоззренческая направленность авторского текста [2]. Что касается текстов ПЦ и ОСР, то они по своим стилевым количественным характеристикам заметно дистанцированы от характеристик текстов ТД. При этом стилевые характеристики послевоенных произведений Михаила Шолохова (2-я кн. ПЦ и ОСР) резко отличаются не только от стиля ТД, но и от стиля 1-й кн. ПЦ. И, наконец, следует отметить еще раз, что характеристики 1–й кн. ПЦ сближаются, но не объединялся, со стилевыми характеристиками последних частей ТД (7-й и 8-й).

Литература

1. *Кукушкина О. В., Поддубный В. В., Поликарпов А. А., Шевелев О. Г.* Автоматическая классификация текстов корпуса русских текстов конца XX века по жанровым типам и источникам // Русский язык: исторические судьбы и современность. Международная конференция. Труды и материалы. Москва, МГУ, 20–23 марта 2007 г. М., 2007. С. 391–392.
2. *Макаров А. Г. и Макарова С. Э.* Цветок-Татарник. В поисках автора «Тихого Дона»: от Михаила Шолохова к Федору Крюкову. М., 2001.
3. *Поддубный В. В., Шевелев О. Г.* Сравнение стилей текстовых произведений по частотному признаку на основе гипергеометрического критерия // Теоретическая и прикладная информатика. Томск, 2004. Вып. 1. С. 101–110.
4. *Фоменко В. П., Фоменко Т. Г.* Авторский инвариант русских литературных текстов. Приложение. Кто был автором «Тихого Дона»? // Новая Хронология Руси и Рима. Т. 2. М. 1995.
5. *Polikarpov A., Kukuškina O., Poddubnyj V., Shevelyov O.* Optimization of feature space in tasks of classification and clustering of texts of various types / Abstracts from Qualico-2009. Graz, 2009.

Числовой профиль сюжета

Г. Я. Мартыненко, Т. Ю. Шерстинова

Санкт-Петербургский государственный университет

g.martynenko@gmail.com, sherstinova@gmail.com

Русский рассказ, Серебряный век, сюжет, динамика, размер предложения, временной ряд, динамический контур

Summary. The paper concerns dependence of quantitative text parameters on linear text structure. Russian short stories of the beginning of the XXth century are investigated. For each short story a dynamic contour of sentence length is built. Common types of dynamic contours are determined for the given text corpus. The most typical contours for Russian short stories of the Silver Age are presented.

1. В последние годы возрос интерес к ритмической организации непоэтического текста. Причем ритм понимается достаточно широко — как проявление симметрийных свойств на различных уровнях структуры теста. Чем вызван этот ин-

терес? Прежде всего, распространением в языкознании системных, а в последние годы и синергетических представлений, в которых ритм и гармония выступают в качестве системного параметра. Далее, и по сей день не теряет привлекательности идея переноса или приспособления методик, накопленных в стиховедении к изучению прозаического материала. Естественным представляется также использование идей и методов экспериментальной фонетики к изучению закономерностей реализации просодических комплексов. Значительную роль в этой пограничной области может сыграть теория временных рядов. И наконец, содержательным фоном и принципиальным арбитром в такого рода исследованиях могут явиться формализованные методики изучения динамики текста (сюжета, фабулы и композиции).

2. При статистическом описании структуры текста исследователи часто пользуются объемными (экстенсивными) переменными (признаками). Среди них наибольшей популярностью пользуется размер предложения, история изучения которого имеет уже вековую традицию. Примечательно, что этой переменной интересуются не только филологи, но и классики статистики — такие как М. Кендалл, Э. Юл, Г. Хердан и др. Привлекательность данного параметра коренится в том, что он весьма «доступен», легко измерим и открыт для интерпретации.

Другим параметрам повезло меньше. Иногда обсуждается размер словоупотребления, причем не только лингвистами, но и специалистами в области кодирования и автоматической обработки текста. Время от времени эта переменная появляется в теоретической статистике в качестве классического примера бимодального распределения. Что касается размера абзаца, то основательных исследований здесь нет.

Важно также отметить, что экстенсивные параметры рассматриваются как правило без учета линейной развертки текста, так как для исследователей важны прежде всего обобщающие показатели, характеризующие структуру текста в целом и позволяющие тем самым производить типолого-классификационные операции: атрибуционные, таксономические, диагностические и т. п.

3. Предположим, что экстенсивные параметры, например, размер предложения, по мере развертки сюжета произведения ведут себя не случайно, а подчиняются некоторым скрытым закономерностям, регулируемым сложным переплетением содержательных факторов.

Эксперимент был проведен на материале текстов выдающихся русских прозаиков первой трети XX века — А. Чехова, Л. Андреева, А. Куприна, И. Бунина, М. Горького, З. Гиппиус, Ф. Сологуба, М. Зощенко, М. Булгакова и др. Каждый автор представлен в корпусе тремя рассказами; было исследовано 50 рассказов. Эксперимент включает несколько шагов:

1. В каждом рассказе фиксируются размеры каждой переменной, например, размер предложения.

2. Линейная последовательность значений переменных в каждом рассказе подвергается тесту на независимость с целью обретения уверенности в том, что эти значения образуют неслучайную последовательность.

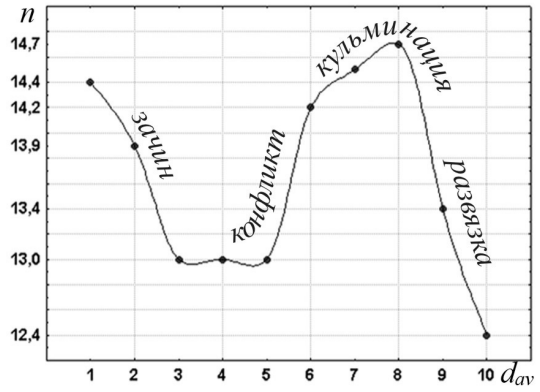
3. Вся последовательность предложений и абзацев разбивается на 10 частей и в каждой из них подсчитывается среднее значение соответствующего параметра.

4. Для каждой переменной строится временной ряд, в котором в качестве независимой переменной выступает номер отрезка текста, а в качестве зависимой переменной — размер абзаца или предложения.

5. Полученные значения средних величин в каждом ряду подвергаются выравниванию с помощью одного из вариантов метода скользящих средних.

6. Для каждого рассказа, для каждого автора и всего корпуса в целом строятся динамические кривые, отражающие изменения переменных по мере развития сюжета, начиная от экспозиции и завязки через развитие действия к кульминации и развязке.

На рисунке представлен динамический контур рассказа А. П. Чехова «Переполох» применительно к размеру предложения.



Динамический ряд среднего размера предложения в рассказе А. П. Чехова «Переполох»

График динамического ряда имеет извилистый, волнообразный контур, коррелирующий с течением «реальных» событий рассказа, а также структурно-композиционными компонентами, отражающими развитие действия. Такой контур достаточно типичен и для усредненного динамического контура русского рассказа первой трети XX века в целом.

В таблице, представленной ниже, приводится перечень основных динамических профилей русского рассказа применительно к размеру предложения.

Ранг	Фигура	Частота	Ранг	Фигура	Частота
1		22	4		5
2		8	5		4
3		7	6		4

Данные таблицы говорят о том, что разнообразие фигур не велико, при этом почти половину всех реализаций дает фигура под рангом 1 — та, которая была получена для рассказа «Переполох». Именно эта фигура является «нормативной» в динамике сюжета.

4. Полученные результаты позволяют перевести в эксплицитную плоскость традиционные филологические представления о динамике содержания художественного произведения, а также построить графическую и квантитативную типологию динамических вариантов развития действия от завязки через кульминацию к развязке.

5. Представляется, что процессы, связанные с человеческой деятельностью и определенным образом параметризованные, могут рассматриваться как тексты или текстоподобные образования с соответствующим динамическим контуром.

Русский язык и научно-технический перевод

Ю. Н. Марчук

Московский государственный университет имени М. В. Ломоносова

marchuk43@migtel.ru

Научно-технический перевод на русский язык; многозначность общеупотребительных слов и терминов; необходимость точного стилистического разграничения текстов

Summary. Three problems of scientific & technical translation into Russian are specified : new meanings of common vocabulary words, polysemy of terms and necessity of more exact stylistic qualification of texts to be translated in the new communication situation of today.

Перевод, в особенности научно-технический перевод, объем которого превосходит все другие виды переводов в мире, играет большую роль не только в научно-техническом прогрессе всякого общества, но также и влияет на развитие и состояние любого естественного языка Нет необходимости доказывать эту аксиому. В концепции академика Ю. В. Рождественского массовая коммуникация как новый этап семиозиса состоит из трех составляющих: массовая информация, информатика и новая реклама [6]. Информатика, понимаемая в данной концепции как научно-техническое информирование, играет в массовой коммуникации и, соответственно, в современном семиозисе, одну из главных ролей. В типологии перевода, которая в настоящее время активно разрабатывается и, как всегда, строится на разных принципах, теория и модели научно-технического перевода также активно разрабатываются, см. например, работу Л. Н. Беляевой [1]. Появляются все новые концепции, типологии текстов, классификации видов перевода в связи с новыми информационными технологиями [5].

При этом много внимания уделяется как типологии терминов (см,. например, [3], так и общеупотребительной лексике в составе научно-технических текстов [2].

В данной работе мы рассмотрим три аспекта современной лексической ситуации в научно-техническом переводе с точки зрения его взаимодействия с русским языком: новые значения общеупотребительных слов, многозначность терминов, взаимодействие терминологических полей и терминологий в разного вида научно-технических текстах. Теоретические импликации новой ситуации еще предстоит осмыслить, а практические проблемы требуют практических решений.

Новые значения общеупотребительных слов в переводе научно-технических текстов распространены довольно широко, по мнению Л. И. Борисовой, которая долгое время специально занималась этим вопросом. Здесь надо отметить два принципиальных подхода к рассмотрению общенаучной (общеупотребительной, общенародной) лексики в таких текстах. Согласно одному из подходов, лексику, отличающуюся от терминологической, можно делить на два слоя : общеупотребительную и общенаучную. Другой подход, которого придерживается Л. И. Борисова и Ю. В. Рождественский, состоит в том, что лексику, отличающуюся от терминологической, можно рассматривать, не проводя раздела между общенаучной и общеупотребительной. По нашему мнению, второй подход более целесообразен, поскольку трудно найти эффективные критерии отделения общенаучной лексики от общеупотребительной в научно-технических, да и в других профессиональных текстах , например, в политических или экономических.

Следует оговориться о том, что в нашем рассуждении научно-технический перевод будет рассматриваться на материале англо-русской языковой пары, наиболее распространенной в современной научно-технической коммуникации, хотя и в других языковых парах проблемы носят такой же характер.

Многозначность терминов также является проблемой перевода. Например, английское слово *pattern* только в словаре авиационно-космических систем имеет следующие значения: *схема, структура, спектр, диаграмма, маршрут, вид, картина.* Конкретный переводной эквивалент может зависеть как от ближайшего, так и от достаточно отдаленного контекста. В связи с этим точный выбор переводного эквивалента может потребовать точной классификации типа научно-технического текста.

Третий аспект — стилистическая типология научно-технических текстов. Здесь применимы методы квантитативной лингвистики [4]. Формальный анализ содержания текстов, как показали опыты по машинному переводу, дает результаты, отличные от «человеческого» анализа стиля и структуры переводимых текстов. Поэтому типология русских научно-технических текстов в формальном плане еще ждет своих теоретических и практических постановок и результатов исследования.

Литература

1. *Беляева Л. Н.* Теория и практика перевода. СПб., 2007.
2. *Борисова Л. И.* Лексические особенности англо-русского научно-технического перевода. М., 2005.
3. *Гринев-Гриневич С. В.* Терминоведение. М., 2008.
4. *Марчук Ю. Н.* Компьютерная лингвистика. М., 2007.
5. Перевод: информационные технологии: Сб. статей / Отв. ред. И. И. Убин М., 2009.
6. *Рождественский Ю. В.* Философия языка. Культуроведение и дидактика. Современные проблемы науки о языке. М., 2003.

Метод контекстного разрешения функциональной омонимии для русского языка

О. А. Невзорова

Татарский государственный гуманитарно-педагогический университет (Казань)

onevzoro@gmail.com

Функциональная (грамматическая) омонимия в русском языке, автоматическое разрешение многозначности, контекстные методы

Summary. The paper discusses the main problems of the method of functional homonymy disambiguation on the basis of contextual rules in Russian. The state-of the-art of lexicographical resources and complicated cases of functional homonymy disambiguation are among the topics debated.

1. Структура метода контекстного разрешения функциональной омонимии

Разрешение функциональной (грамматической) омонимии является одной из актуальных задач обработки текстов. К настоящему времени сформирована основная парадигма методов снятия омонимии, которая включает методы, основанные на правилах; методы машинного обучения, использующие вероятностные модели; гибридные методы. Статистические методы активно развиваются для русского языка прежде всего благодаря проекту «Национальный корпус русского языка», в рамках которого подготовлен размеченный подкорпус русского языка для настройки алгоритмов машинного обучения. Однако, статистические методы для автоматического разрешения многозначности для русского языка пока не достаточно изучены. Отсутствие для исследований доступного размеченного корпуса большого размера, позволяющего получить релевантные статистические данные распределения 500–2000 грамматических тегов, представляет одну из основных проблем. Другим подходом к разрешению многозначности является подход, основанный на правилах. Этот подход является чрезвычайно трудоемким, требует проведения тщательной лингвистической экс-

пертизы каждого типа омонимии. Несмотря на большой исторический возраст, данный метод для русского языка в полной мере не описан в открытой литературе, некоторые принципиальные идеи и реализация представлены в [1]. В [3] даны сравнительные оценки различных модулей разрешения омонимии, построенных на основе статистических методов и метода, основанного на правилах. В целом, оценки для случая полного разрешения функциональной омонимии достаточно близки 97,26% и 96,87%. Можно предположить, что неразрешенные примерно 3–5% относятся к синтаксически сложным случаям и многие авторы сходятся во мнении, что наиболее эффективным является использование гибридных технологий разрешения омонимии. Однако следует отметить, что полученные оценки даны для классификации типов омонимии, принятой в Национальном корпусе русского языка. Эта классификация в ряде конкретных случаев функциональной омонимии расходится с другими классификациями. Предварительно можно отметить, что развитие контекстного метода способствует более четкому выделению основных проблем, связанных, прежде всего, с описанием явления функциональной омонимии в существующих лексикографических источниках; выделением синтаксически сложных случаев разрешения омонимии.

Разработка метода контекстного разрешения функциональной омонимии [2] требует решения следующих задач:

1) лексикографические задачи:
- уточнение набора грамматических характеристик функциональных омонимов;
- построение полной классификации типов функциональных омонимов.

2) вычислительные задачи:
- выделение минимального множества разрешающих контекстов для каждого функционального типа. Формализация контекстных условий.
- для каждого функционального типа построение управляющей структуры обобщенного правила, обеспечивающего максимальную точность распознавания.

2. Основные результаты и проблемы метода контекстных правил

На основе сравнительных сопоставлений различных лексикографических источников, включая словари и корпуса русского языка, задача построения достаточно полного списка грамматических омонимов с уточненными грамматическими характеристиками близка к завершению. Связанная с этой задачей задача классификации типов функциональных омонимов также практически решена. В настоящее время выявлено около 220 классов грамматических омонимов (по оценкам Т. Ю. Кобзаревой в [1] — 57 классов) и построены списки представителей этих классов.

Для каждого типа функциональной омонимии разрабатывается обобщенное правило разрешения омонимии данного типа. Обобщенное правило представляет собой упорядоченную совокупность правил, записанных на специальном формальном языке. Каждое правило внутри совокупности фиксирует некоторый разрешающий контекст, порядок применения правил внутри функционального типа базируется на оценке частотности контекстов. Развитие метода связано с учетом контекстов сложной синтаксической природы, в частности, с анализом однородных групп. Выделение однородной группы позволяет искать разрешающий элемент за границами однородной группы; тем самым, реально увеличивается численный интервал разрешающего контекста. Такого рода правила анализа омонимов в составе однородной группы были включены в состав обобщенных правил различных функциональных типов.

Метод разрешения функциональной омонимии на основе контекстных правил по сути своей базируется на синтаксических моделях. Это обстоятельство определяет и ограничения метода. Приписывание омониму той или иной характеристики части речи осуществляется на основе анализа наличия либо отсутствия в контексте определенной длины слов тех или иных классов. Явления эллипсиса и субстантивации в тексте также представляют сложную проблему метода.

Одним из возможных подходов к разрешению омонимии в сложных (коротких и эллиптических) контекстах является логико-семантический подход к описанию предложений. Нами исследованы логико-семантические интерпретации омонимичных конструкций, относящихся к области неясных и спорных синтаксических явлений, а именно конструкций, находящиеся в тесной смысловой и синтаксической зависимости от окружающего контекста на примере функционального поведения омонимов типа N* / A*. Построена классификация сложных контекстов и предложены логико-семантические интерпретации для правил разрешения омонимии.

Анализ сложных контекстов позволил выделить определенные типы омонимичных контекстов, разрешение которых требует полного синтаксического анализа. Кроме сложных случаев постсинтаксического разрешения можно выделить контексты, не разрешаемые синтаксическими методами, т. е. сохраняющие многозначность.

Литература

1. *Кобзарева Т. Ю., Афанасьев Р. Н.* Универсальный модуль предсинтаксического анализа омонимии частей речи в РЯ на основе словаря диагностических ситуаций // Труды междунар. конференции Диалог'2002. М., 2002. С. 258–268.
2. *Невзорова О. А., Зинькина Ю. В, Пяткин Н. В.* Метод контекстного разрешения функциональной омонимии: анализ применимости // Труды межд. конф. Диалог'2006. М., 2006. С. 399–402.
3. *Сокирко А. В., Толдова С. Ю.* Сравнение эффективности двух методик снятия лексической и морфологической неоднозначности для русского языка (скрытая модель Маркова и синтаксический анализатор именных групп) // Интернет-математика-2005. http://company.yandex.ru/grant/list.xml.

Дискриминантный анализ стилей текстовых произведений

В. В. Поддубный, А. С. Кравцова

Томский государственный университет

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

Тексты, признаки стиля, частоты признаков, ранги частот, дискриминантный анализ, русская проза

Summary. The using of discriminant statistical analysis to decision of the problem of the comparison of the styles of the text products on the base of features frequencies is considered. It is offered the procedure of normalization of frequencies by way of the transition from relative frequencies to its ranks with the following nonlinear transformation of ones into gaussian values. On example of the analysis of texts of russian novels of 19–th century the discrimination of the author's styles on the frequencies of the using of 55 syntactic words is organized.

1. Рассматривается применение дискриминантного статистического анализа к решению проблемы сравнения стилей текстовых произведений на основе частотных признаков. Предлагается процедура нормализации признаков путем перехода от исходных признаков к их рангам с последующим нелинейным преобразованием в нормально распределенные величины. На примере анализа текстов русской художественной прозы XIX века проведена дискриминация авторских стилей текстов по частотам употребления 55 служебных слов.

2. Дискриминантный анализ [2] является одним из мощных инструментов математической статистики, позволяющий исследовать статистические различия классов объектов, относительно однородных внутри каждого класса. Применительно к текстовым произведениям дискриминантный анализ позволяет исследовать степень различия текстовых произведений по авторству, жанру и прочим группирующим признакам при различных наборах признаков стилей текстов (частот употребления служебных слов, наиболее употребительных слов, биграмм и т. п.).

3. При фиксированном (выбранном) наборе признаков стилей текстов каждый текст может быть представлен точкой в многомерном пространстве частот признаков или некоторых (в общем случае нелинейных) функций от них. Произведения одного автора (или одного жанра) группируются в относительно компактные сгустки точек (классы),

в общем случае достаточно сильно перекрывающиеся для разных авторов (или разных жанров). При сильном перекрытии классов задача различения классов (стилей текстов) в многомерном пространстве признаков стилей является достаточно трудной. Дискриминантный анализ позволяет максимально разнести классы друг относительно друга.

4. Пусть $p = \{p_{ij}\}$ — $n \times m$-матрица относительных частот появления j-го признака в i-м тексте ($i = 1,n$, $j = 1,m$, где n — число текстов, m — число признаков). Проранжировав в порядке возрастания величины $\{p_{ij}\}$ по всем текстам (от 1-го до n-го) для каждого j-го признака, получим матрицу рангов $r = \{r_{ij}\}$ признаков (мест признаков среди текстов). При этом рангам совпадающих частот, образующих так называемые связки, припишем средний по связке ранг. Разделив ранги на $n + 1$, приведем их к интервалу $[1 / (n + 1), n / (n + 1)]$. В результате получим матрицу относительных рангов $\{r_{ij} / (n + 1)\}$. Их эмпирическое распределение вероятностей равномерно в единичном интервале. Сопоставим каждому относительному рангу квантиль стандартного нормального распределения уровня этого относительного ранга. В результате такого нелинейного преобразования получим матрицу нормально распределенных величин (нормализованных относительных рангов — НОР) $x = \{x_{ij}\}$ с нулевыми средними и единичными дисперсиями для каждого j-го столбца, причем столбцы будут коррелированны (в общем случае) между собой.

5. Пусть имеется g классов. Вычислив положение центров классов в признаковом пространстве НОР (средние значения координат точек каждого k-го класса), можно подвергнуть оси координат такому линейному преобразованию $y = xV$ (повороту и масштабированию осей), при котором расстояния между центрами классов по отношению к диаметрам классов в новом (дискриминантном) признаковом пространстве станут наибольшими. Дискриминантный анализ предписывает выбирать матрицу V коэффициентов этого преобразования так, чтобы максимизировать отношение $\lambda_l = (V'BV)_{ll} / (V'WV)_{ll}$, $l = 1, 2, \ldots, q$, $q = \min(m, q - 1)$. Здесь штрих — знак транспонирования, $B = T - W$, T — $m \times m$-матрица ковариаций векторов-столбцов матрицы x, W — $m \times m$-матрица внутригрупповых ковариаций векторов-столбцов матрицы x. Из этого критерия оптимизации следует [1]; [2], что m-векторы-столбцы $\{V_l\}$ матрицы V — собственные векторы, соответствующие матрицам B и W и удовлетворяющие уравнению $BV_l = \lambda_l WV_l$, а $\{\lambda_l > 0\}$ — q их первых (в порядке убывания) собственных значений, удовлетворяющих характеристическому уравнению $\det(B - \lambda W) = 0$. Столбцы матрицы $y = \{y_{il}\}$ называются дискриминантными функциями и образуют новое признаковое пространство размерности q (пространство новых факторов), в котором обеспечивается наилучшее разделение (дискриминация) классов. Столбцы матрицы коэффициентов $V = \{V_{jl}\}$ называются факторными нагрузками. При соответствующей нормировке они являются коэффициентами корреляции между каждым новым l-м (дискриминантным) признаком и каждым «старым» j-м признаком пространства нормализованных относительных рангов. Содержательная интерпретация дискриминантных функций (новых, дискриминантных признаков) определяется наборами старых признаков, в наибольшей степени коррелирующими с новыми признаками. Статистическая значимость оставляемой в новом признаковом пространстве l-й дискриминантной функции при гауссовом распределении нормализованных относительных рангов рассчитывается по χ^2-распределению с $\nu_l = (m - l)(g - l - 1)$ степенями свободы, которому при верной нулевой гипотезе подчиняется величина, пропорциональная логарифму Λ-статистики Уилкса [2].

6. В качестве примера приводятся результаты дискриминантного анализа 81 текста крупных произведений художественной прозы (романов, повестей) 12 русских писателей XIX века с использованием в качестве признаков стилей 55 служебных слов. На рис. 1 видно отличное разделение текстов по писателям в пространстве первых двух дискриминантных функций.

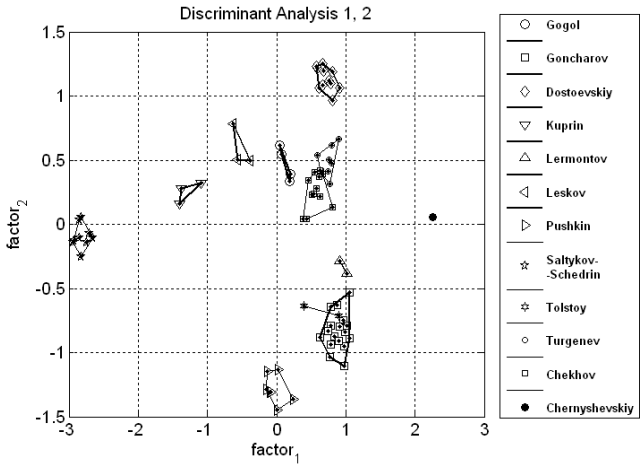


Рис. 1. Дискриминантный анализ текстов русских писателей XIX века по 55 служебным словам.

Литература

1. *Кендалл М. Дж., Стьюарт А.* Многомерный статистический анализ и временные ряды / Пер. с англ. М., 1976.

2. *Ким Дж.-О., Мьюллер Ч. У., Клекка У. Р. и др.* Факторный, дискриминантный и кластерный анализ / Под ред. И. С. Енюкова; пер. с англ. М., 1989.

Математическое моделирование жизненного цикла языкового знака

В. В. Поддубный

Томский государственный университет

pvv@inet.tsu.ru

А. А. Поликарпов

Московский государственный университет имени М. В. Ломоносова

anatolp@philol.msu.ru

Языковой знак, жизненный цикл, полисемия, математическая модель, диссипативный процесс

Summary. The dissipative nonstationary dynamic mathematical model of the life cycle of the language sign is offered. This cycle is based on the interaction of the processes of the sign polysemy growing and of the losing of the earlier gained sign meanings. It is shown that this model satisfies to the variational principle of the least action. The model is presented in continuous and discrete variants. The stochastic expansion of the discrete variant of the model is built. The numerical modeling of the process of the language sign polysemy is made.

1. Жизненный цикл языкового знака от момента его зарождения до момента выхода из употребления связан с взаимодействием двух процессов его развития: процесса роста полисемии знака, приобретения знаком новых, как правило,

все более абстрактных значений, и процесса постепенного выхода из употребления ранее приобретенных значений, начиная с наименее абстрактных. Второй процесс начинается с некоторым запаздыванием по отношению к первому и протекает более медленно. Разность между количеством приобретенных знаком значений и количеством значений, вышедших из употребления к данному моменту времени, составляет размер активной полисемии знака, т. е. количество живущих в этот момент времени значений знака. Максимальное количество возможных значений знака назовем его ассоциативно-семантическим потенциалом (АСП) [3].

2. Естественно предположить, что скорость роста полисемии знака в каждый текущий момент времени пропорциональна запасу потенциала полисемии этого знака, т. е. разности между АСП и количеством значений, порожденных знаком к этому моменту времени. Это предположение в непрерывной модели развития полисемии приводит к дифференциальному уравнению роста полисемии. К моменту рождения знака запас полисемии максимален и равен АСП. Коэффициент пропорциональности при этом может зависеть от времени, возможно, монотонно убывая, оставаясь постоянным или монотонно возрастая. Но всегда этот коэффициент положителен: скорость роста полисемии тем больше, чем больше запас потенциала полисемии. При этом, очевидно, с ростом полисемии скорость роста падает. Процессы с такими свойствами, описываемые дифференциальными уравнениями первого порядка с правой частью, зависящей от состояния процесса и, возможно, времени, и имеющей отрицательную частную производную по состоянию, называются диссипативными [2]; [4]. В случае независимости коэффициента от времени диссипативную динамическую систему, описываемую таким дифференциальным уравнением, называют стационарной [2], в противном случае — нестационарной [4]. Естественно предположить, что и процесс потери значений знака (процесс выхода значений из употребления) является таким же диссипативным динамическим процессом, как и процесс роста полисемии, но с другим (меньшим) коэффициентом пропорциональности, обеспечивающим более медленное его развитие. Эти предположения определяют математическую диссипативную динамическую модель процессов роста и потерь полисемии, а, следовательно, и математическую модель жизненного цикла языкового знака. Из этой модели математически строго следует, что жизненный цикл языкового знака, начинаясь с единственного исходного значения, со временем проходит стадию роста полисемии, а затем, достигнув пика своего развития, утрачивает все большее число значений до выхода из употребления последнего значения и знака в целом. Таким образом, жизненный цикл языкового знака во времени представляется унимодальной (с одним максимумом) асимметричной кривой с положительной асимметрией (смещением пика в сторону начала процесса), спадающую затем до нуля (выхода знака из употребления).

3. Математически строго показано, что жизненный цикл знака, описываемый математической моделью диссипативного динамического процесса (в общем случае нестационарного), подчиняется вариационному принципу наименьшего действия [1]; [5], лежащему в основе формулировки многих законов природы. Применительно к процессу развития полисемии языкового знака принцип наименьшего действия, обосновывающий его диссипативную математическую модель, можно сформулировать так: из всех возможных кривых активной полисемии знака в природе реализуется кривая, минимизирующая интеграл, называемый функционалом действия. Функционал действия — это число, выражающее суммарную на траектории жизненного цикла знака разность между «кинетической энергией» полисемии и ее «потенциальной функцией» [1]; [5]. По определению, «кинетическая энергия» пропорциональна квадрату скорости изменения полисемии, «потенциальная функция» — квадрату запаса ассоциативно-семантического потенциала полисемии. Из решения этой вариационной задачи математически строго вытекают наши естественные предположения о диссипативной динамической математической модели жизненного цикла знака.

4. В дискретном варианте диссипативной математической динамической модели жизненного цикла знака условие пропорциональности скорости роста полисемии запасу его ассоциативно-семантического потенциала выражается не дифференциальными, а разностными уравнениями, определяющими моменты рождения и выхода из употребления значений знака. Интервал времени между рождением текущего и следующего за ним значения знака обратно пропорционален текущему запасу потенциала полисемии. То же — для выхода из употребления значений знака, но со своими параметрами. Моменты рождения и выхода из употребления любого значения языкового знака определяют длительность его жизни. Показано, что в рассматриваемой модели длительность жизни значения монотонно возрастает с его номером (уровнем абстрактности), достигая максимума для самого высокого уровня. С выходом из употребления этого значения жизненный цикл знака заканчивается.

5. Проведено расширение дискретной диссипативной динамической модели развития полисемии языкового знака на стохастический случай, когда коэффициенты детерминированной модели становятся случайными неотрицательными функциями времени, обеспечивающими ту же самую последовательность рождения новых и потери ранее приобретенных значений знака, но в случайные моменты времени. Структура такой стохастической модели остается диссипативной и удовлетворяющей принципу наименьшего действия. Эта модель включает в себя мультипликативный случайный фактор, приводящий к случайным колебаниям моментов последовательного рождения и выхода из употребления значений знака. Детерминированная модель определяет медиану стохастического процесса развития полисемии знака, описываемого стохастической моделью.

6. Проведено компьютерное моделирование детерминированных и случайных процессов развития полисемии знака в соответствии с предложенной моделью.

Литература

1. *Айзерман М. А.* Классическая механика. М., 1980.
2. *Лоскутов А. Ю., Михайлов А. С.* Основы теории сложных систем. М.; Ижевск, 2007.
3. *Поликарпов А. А.* Системно-квантитативный подход в лингвистике // Филологические школы и их роль в систематизации научных исследований. Вестник Смоленского государственного университета. Сер. 1: Филология. Т. 1. Смоленск, 2007. С. 35–59.
4. *Шамолин М. В.* Динамические системы с переменной диссипацией: подходы, методы, приложения // Фундаментальная и прикладная математика. Т. 14. № 3. М., 2008. С. 3–23.
5. *Эльсгольц Л. Э.* Дифференциальные уравнения и вариационное исчисление. М., 1965.

Обнаружение события в русских художественных текстах

С. Б. Потёмкин

Московский государственный университет имени М. В. Ломоносова

potemkin@philol.msu.ru

Summary. Techniques for determining the point of event in a fiction text is proposed. Formal detection of events in the discourse is based on text only without its perception by a reader. Our hypothesis is as follows: descriptions of states are differed by numerous pairs of antonyms of adjectives or adverbs. This paper involves examination of this hypothesis and the results obtained.

Проблема выявления событий в тексте. Событие, элементарная составляющая повествовательного текста, было определено Ю. М. Лотманом как «перемещение персонажа через границу семантического поля». Таким образом, событие заключается в некоем отклонении от законного, нормативного в данном мире, в нарушении одного из тех правил, соблюдение которых сохраняет порядок и устройство этого мира. Определение события как переход между двумя последовательными во времени или причинно-следственно связанными ситуациями покрывает практически все многообразие изменений в любом произведении. Полноценная событийность в нарративном тексте подразумевает выпол-

нение целого ряда дальнейших условий: — **факт** изменения — изменение должно действительно произойти в (фиктивном) мире. С фактичностью связано второе основное условие событийности: — *результат* изменения — должен состояться до конца наррации. Более тонкие свойства события включают следующее: релевантность; непредсказуемость; консекуативность; необратимость; неповторяемость.

Пары антонимов, покрытие, расширение. Можно попытаться составить описание каждого состояния набором признаков, задающих точку в некотором семантическом пространстве. Метод семантического дифференциала использует бинарные шкалы, концы которых отмечены словами-антонимами типа *безопасный — опасный, широкий — узкий, добрый — злой* и позволяет задавать состояние в этих координатах. Антонимические отношения в наибольшей степени свойственны качественным прилагательным и наречиям, в меньшей — существительным и глаголам (главным образом тем, которые содержат в своих значениях качественный признак). Для нас важным является тот факт, что определенное состояние можно описывать набором прилагательных (и наречий), имеющих антонимы, и можно ожидать, что последующие состояния, порожденные произошедшим событием, будет описываться новым набором прилагательных и наречий. Множество антонимических пар (прилагательных) является сильно разреженным относительно множества всех прилагательных языка. Так, во всем произведении (рассказ, повесть) может не встретиться ни одной антонимической пары, либо их количество настолько мало, что описание состояний будет статистически недостоверным. Современные словари антонимов содержат около 10000 антонимических пар и пополнение этого списка требует обширных корпусных исследований. С другой стороны, покрытие текста синонимами значительно плотнее, чем покрытие антонимами. Можно попытаться расширить списки пар антонимов за счет добавления каждому члену антонимической пары списка его синонимов. Механическое расширение списка антонимических пар за счет присоединения синонимов не всегда дает «истинные» в общепринятом смысле антонимы. Например, для прилагательного *веселый* получен следующий список антонимов, встречающихся в том же произведении: *бедный, *бездарный, грустный, жалобный, мрачный, невеселый, *нечистый, *отвратительный, печальный, *святой, сердитый, скучный, траурный, угрюмый, унылый.* (* отмечены сомнительные антонимы, подчеркнуты антонимы, не включенные в исходный словарь антонимов).

Выявление события и эксперименты на текстах русской классики. Суть метода выявления событий заключается в следующем: Для каждого прилагательного или наречия в отдельном предложении отыскиваются и подсчитываются все его антонимы в остальном тексте рассказа, повести или в пределах главы, части крупного произведения. Число антонимов откладывается по оси Y, в зависимости от номера предложения, отложенного по оси X. Полученный график антонимов подвергается сглаживанию и интерпретации (рис. 1).

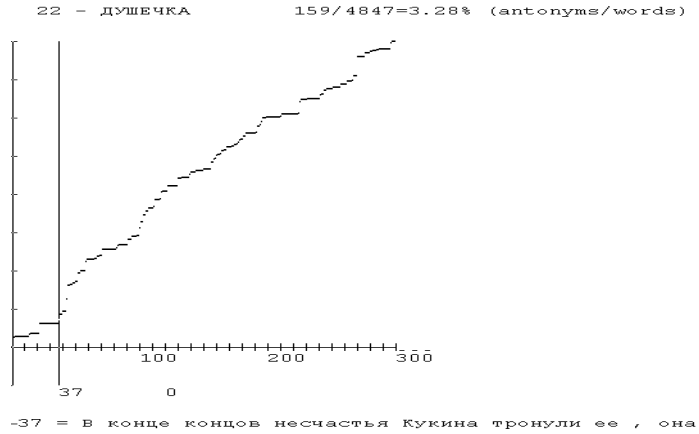


Рис. 1. График числа антонимов в зависимости от номера предложения для рассказа А. П. Чехова «Душечка».

Пользователь может получить номер предложения, на котором происходит существенное изменение числа антонимов и само это предложение, передвигая вертикальную черту вдоль оси X. На графике можно проследить 4 периода, соответствующие 4 периодам жизни Ольги Племянниковой: 1 — (с 37 по 94) — жизнь с антрепренером Кукиным, 2 — (100 по 154) — жизнь с лесоторговцем Пустоваловым, 3 — (163–203) — жизнь с ветеринаром Смирниным и после его отъезда и, наконец, 4 — (232–272) — жизнь ради сына ветеринара, Саши. Эти периоды соответствуют отклонениям кривой графика от среднего значения. Для других рассказов Чехова и повестей Гоголя и также получены графики антонимов и проведена их интерпретация. Так, в «Шинели» Гоголя выявлены 4 существенных разрыва графика антонимов, отождествляемые с событиями перехода между 5 состояниями: 1 — экспозиция бытия Башмачкина; 2 — осознание им необходимости смены шинели; 3 — обретение им цели жизни в связи со строительством новой шинели; 4 — утрата иллюзий и гибель Башмачкина; и 5 — его посмертное существование. В дальнейшем предполагается выполнить анализ других малых форм, в том числе современных русских писателей. В то же время интересно получить результаты для других языков, в первую очередь английского, а также оценить событийность в параллельных текстах — оригинальном тексте и его переводе. Возможно, дальнейшая формализация метода потребует применения методов спектрального и регрессионного анализа полученных данных.

Лексическая синонимика языка А. С. Пушкина: качественный и количественный анализ

С. В. Рымарь, И. В. Кузнецов

Муромский институт (филиал) Владимирского государственного университета

svetlavir@mail.ru

Язык А. С. Пушкина, синонимы, количественный анализ, стилистическая организация

Summary. The paper is dedicated to the description of synonymic raws of A. Pushkin language. Matrix to study quantitative distribution of synonyms according to their belonging to the parts of speech, stylistic indication, raw structure is represented. Indices obtained are compared with the ones, belonging to common language. This allows to make certain conclusions about qualitative originality of A. Pushkin language stylistic organization.

Знание синонимии родного языка, семантический и стилистический анализ функционирования синонимов в художественной речи позволяет не только правильно понимать процесс образования русского литературного языка, определить пути и нормы его дальнейшего развития, более точно воспринимать тексты произведений мастеров художественного

слова, но и обязывает носителей языка к бережному обращению с русским словом. Такие ученые, как В. В. Виноградов, Ш. Балли, А. М. Пешковский, Л. В. Щерба, А. Н. Гвоздев, видели в синонимике фундамент стилистической системы языка.

В. А. Гречко была разработана матрица на материале «Словаря синонимов» под ред. А. П. Евгеньевой (Л., 1975) для изучения количественного распределения синонимов по их принадлежности к частям речи, стилистическим признакам, составу рядов. Использованная методика количественного анализа позволяет сделать определенные заключения о качественном своеобразии стилистической организации современного русского языка, о тенденции ее движения».

Нами была изучена указанная методика количественного анализа синонимов и применена для анализа синонимических рядов языка Пушкина. В данной работе нам важно не только провести количественный анализ синонимических рядов в языке поэта, но и сравнить полученные показатели с общеязыковыми. Думаем, что это позволит сделать определенные заключения о качественном своеобразии стилистической организации языка Пушкина, выявить возможные тенденции развития стилистики языка.

Обратимся к результатам подсчетов и в современном языке, и в языке Пушкина. Показательно общее количественное соотношение синонимов различных частей речи. В общеупотребительном языке по количеству глаголы-синонимы стоят на первом месте по сравнению с синонимами, выраженными другими частями речи. Для сравнения, по данным «Обратного словаря русского языка» (М., 1974), на первом месте по количеству слов стоит имя существительное (56332 слова, или 46,35% к общему объему словаря — около 125 тыс. слов); затем — в порядке убывания — идут глагол (37319 слов, 30,71%), прилагательное (24786 слов, 20,44%), наречие (1916 слов, 1,58%). Количественное распределение синонимов по частям речи в объеме «Словаря синонимов» под ред. А. П. Евгеньевой иное. Общее количество синонимов основных знаменательных частей речи — около 21000 слов; из них глаголов — 8705 слов (около 42%), существительных — 6044 (29%), прилагательных — 4455 слов (21%), наречий — 1611 слов (7,75%).

Приведем результаты подсчетов распределения синонимов по частям речи в языке Пушкина. Общее количество синонимов основных знаменательных частей речи — около 7301 слов; из них существительных — 2833 слов (38,80%), глаголов — 2439 слов (33,41%), прилагательных — 1523 слов (20,86%), наречий — 506 слов (6,93%).

Обратимся к общему количественному распределению синонимов в языке Пушкина по рядам. Наибольшее количество синонимической лексики приходится на двусловные ряды (1700 слов, или 23,28% общего количества синонимов в языке Пушкина). Отмечается значительное количество трехсловных рядов (1524 слов, или 20,87%) и четырехсловных рядов (1108 слов, или 15,18%). Число синонимов в других рядах значительно снижается: пятисловные ряды насчитывают 730 синонимов (9,99%), шестисловные — 492 (6,74%), семисловные — 399 (5,47%), восьмисловные — 288 (3,94%), девятисловные — 225 (3,08%), десятисловные — 150 (2,05%), одиннадцатисловные и более — 685 (9,38%).

При сравнении общего количественного распределения синонимов по рядам в языке Пушкина и в «Словаре синонимов» под редакцией А. П. Евгеньевой можно прийти к таким результатам: в языке Пушкина наибольшее количество синонимической лексики приходится на двусловные ряды, в «Словаре синонимов» — на трехсловные ряды. В отношении других рядов прослеживается следующая тенденция: в языке Пушкина количество синонимов в трех-, четырех-, пяти-, шести-, семи-, восьми-, девяти-, десятисловных рядах уменьшается от ряда к ряду, а в «Словаре синонимов» наблюдается следующая последовательность распределения синонимов по рядам: трех-, четырех-, двух-, пяти-, шести-, семи-, восьми-, девяти-, десятисловные ряды.

В языке Пушкина среднее количество синонимов в ряду — 3,6. Можно отметить, что данный показатель близок и к среднеязыковому количеству, и к среднему количеству синонимов в ряду, зафиксированному в «Словаре синонимов» под ред. А. П. Евгеньевой.

Представляет интерес сравнительный анализ количественного распределения синонимов знаменательных частей речи по рядам в языке Пушкина и в «Словаре синонимов» под ред. А. П. Евгеньевой (см. табл. 1).

Таблица 1. Сравнительная таблица количественного распределения синонимов знаменательных частей речи по рядам.

ЧАСТИ РЕЧИ	Количество синонимов в ряду									
	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**проч.**
Существ.	1257 / 638	1594 / 633	1042 / 400	720 / 240	538 / 192	332 / 105	198 / 104	124 / 72	80 / 90	158 / 359
Глагол	1166 / 632	1862 / 546	1775 / 416	1360 / 195	720 / 132	442 / 175	274 / 96	263 / 90	267 / 30	576 / 127
Прилагат.	508 / 320	1104 / 261	837 / 184	707 / 220	392 / 132	217 / 84	262 / 72	148 / 54	128 / 30	152 / 166
Наречие	110 / 110	265 / 84	264 / 108	244 / 75	104 / 36	135 / 35	138 / 16	87 / 9	80 / -	184 / 33
ВСЕГО	**3041 / 1700**	**4825 / 1524**	**3918 / 1108**	**3031 / 730**	**1754 / 492**	**1126 / 399**	**872 / 288**	**622 / 225**	**555 / 150**	**1070 / 685**

Количественные данные указаны в таблице следующим образом: в числителе — количество синонимов в ряду в «Словаре синонимов» под ред. А. П. Евгеньевой; в знаменателе — количество синонимов в ряду в языке Пушкина. Приведенное выше количественное распределение синонимических рядов языка Пушкина, их первоначальный анализ — это лишь попытка представить организацию лексической синонимики языка поэта. Качественная оценка стилистических процессов в языке Пушкина требует углубленного и разностороннего исследования.

Дистрибуция и реализация словоизменительных аффиксов русского языка: подходы к анализу*

С. Б. Степанова, А. С. Асиновский, А. И. Рыко, Т. Ю. Шерстинова

Санкт-Петербургский государственный университет

stsvet_2002@mail.ru , a.s.asinovsky@gmail.com, aryko@mail.ru, sherstinova@gmail.com

Грамматика речи, многоуровневая лингвистическая разметка, спонтанная речь, фонетика, морфемика

Summary. The paper presents the analysis of distribution and realization of Russian inflexions in spontaneous speech.

Исследование звукового плана выражения словоизменительных показателей в русском языке соответствует задачам, которые ставили перед отечественной лингвистикой Л. В. Щерба и И. А. Бодуэн де Куртенэ, говоря о необходимости опираться на реальное звучание языковых элементов при описании грамматической системы русского языка.

* Исследование выполнено при поддержке гранта РГНФ «Разработка информационной среды для мониторинга устной русской речи» (09–04–12115в).

Для проведения **корпусного** исследования реализаций словоизменительных аффиксов в спонтанной русской речи, в качестве первого шага, для 5 звуковых файлов из подкорпуса «Один речевой день» [2] было осуществлено аннотирование по методу сплошной выборки: вычленялись и транскрибировались **все** встречающиеся в материале флексии существительных, прилагательных, финитных форм глаголов, причастий, порядковых числительных, изменяемых разрядов местоимений (притяжательных, указательных и т. д.).

Выбор для начала обработки именно флексий, обусловлен их особым значением для русского языка. Л. В. Бондарко писала, что в русском языке «основные грамматические показатели — по крайней мере, у существительных и прилагательных — сосредоточены в тех частях словоформы, которые находятся в конечной ее части и поэтому очень часто являются безударными. Безударность, да еще в абсолютном конце слова, — крайне неблагоприятная позиция, поскольку физиологической особенностью русской звуковой системы является вялая артикуляция, приводящая к очень заметным качественным изменениям заударных гласных. Таким образом возникает известное противоречие между грамматическим уровнем языковой системы и свойствами фонетических единиц, призванных выступать в качестве материальной формы грамматического значения. Нужно ли говорить о том, какой интерес представляет это противоречие для любого лингвиста, стремящегося построить правдивое и полное описание живого языка?» [1: 265].

Процедура выделения флексии методом слухового анализа, поддержанного возможностью наблюдать реализацию звучания на спектрограмме и осциллограмме, а также транскрибирование слухового впечатления от флексии проводились в программе Праат (программа доступна для бесплатного скачивания на сайте www.praat.org). Транскрипция осуществлялась в символах МФА.

Затем проводилось аннотирование флексий в программе Элан на уровнях Morphems-gram (грамматическое значение морфемы) и Morphems-orth (орфографическое представление морфемы). Дальнейшая автоматическая обработка полученных аннотаций позволила выявить самые частотные орфографические манифестации флексий, их дистрибуцию в соответствии с грамматическим значениям, их типичные способы реализации.

Так, самой частотной флексией на нашем материале оказалась {а} — почти 20% от всех словоизменительных финалей. Самое частотное грамматическое значение {а} — им. п. имен существительных ед. ч. 2-го склонения — 40% (*мам**а**, пап**а***). Чуть реже встречается глагольное окончание — 30% (ж. род, прош. вр.: *был**а**, делал**а***), значительно реже — род. п. ед. ч. существительных 1-го склонения — 15% (*друг**а***). На долю остальных 9 грамматических значений {а} приходится около 15% ее употреблений в проанализированном материале.

Реализация {а} характеризуется максимальным разнообразием представленных аллофонов (20 различных обозначений). Чаще всего морфема {а} реализуется в нашем материале как [а] гласный среднего ряда (26%) или как [ə] (19%). В 17% случаев {а} была реализована как задний гласный [ɑ] или [ʌ]. Интересно, что таким образом реализуется чаще всего морфема со значением «прошедшее время глагола, ж. р.», а аллофон [æ] практически не встречается в морфеме с этим значением. Однако это различие не связано с противопоставлением разных морфем {а}, оно имеет чисто фонетическую природу. Задний характер гласного во флексиях глаголов прошедшего времени вызван исключительно соседством с сильно веляризованным согласным /l/.

Интересными являются и наблюдения над реализацией финалей, представленных некоторыми орфографическими двухбуквенными сочетаниями.

Речь в данном случае идет о тех явлениях, о которых Л. В. Бондарко писала в уже процитированной работе: «в словах *тихий, серый* нет никаких следов конечного /j/, вообще неустойчивость среднеязычного сонанта проявляется в заударных комплексах, где он оказывается между гласными. В этих случаях единственным следом сонанта оказывается передний характер гласного, который следует за ним» [1: 266]. Именно этот тезис можно подтвердить, наблюдая материал, обработанный к настоящему времени. Из 46 финалей {ый} / {ий} в различных формах /j/ был реализован лишь в одном случае.

Иначе обстоит дело с окончанием {ой}. Чаще всего эта финаль встречается в окончаниях имен прилагательных (им. п., м. род, ед. ч.). Как известно, это окончание всегда ударно. Думается, именно поэтому почти в 30% случаев конечное буквосочетание {ой} реализовано как [oj]. Что касается заударных морфных комплексов, где среднеязычный сонант оказывается между гласными (случаи типа *красн**ые***, *красн**ую***), то, как показывает наш материал, исчезает не только этот сонант, но и само сочетание гласных в подавляющем большинстве случаев стягивается до одного однородного аллофона. Из 99 подобных финалей только в 28 реализовано бифонемное сочетание. Думается, они приходятся на окончания с ударным гласным, что предстоит еще проверить.

В целом можно сказать, что корпусный подход к накоплению, систематизации и — что особенно важно — к анализу материала русской звучащей речи позволяет проверить многие сделанные ранее наблюдения фонетистов, подтвердить или опровергнуть многие научные гипотезы, ответить на многие поставленные ранее вопросы и, может быть, поставить новые, на которые предстоит искать ответам новым поколениям лингвистов.

Литература

1. *Бондарко Л. В.* Фонетика современного русского языка. СПб., 1998.
2. *Asinovsky A., Bogdanova N., Rusakova M., Stepanova S., Ryko A., Sherstinova S.* The ORD Speech Corpus of Russian Everyday Communication «One Speaker's Day»: Creation Principles and Annotation // LNCS / LNAI series. «Text, Speech and Dialogue». TSD-2009. Berlin; Heidelberg., 2009. P. 250–257.

Сила синтаксической связности как критерий определения тематических словоформ в тексте

Н. Г. Чейлытко

Киевский национальный университет имени Тараса Шевченко (Украина)

Natalia.cheilytko@gmail.com

Зона связей словоформы, блок связности словоформ, лексико-тематическая модель текста, лексико-тематическое дерево

Summary. The report focuses on the problem of constructing the lexical-thematic model of the text with the assistance of knowledge about the peculiarities of the distribution of syntactic relations of word-forms in the text.

Доклад посвящен проблеме построения лексико-тематической модели текста с привлечением знаний об особенностях распределения синтаксических связей словоформ в предложениях текста. Ключевыми для представленной в докладе методики являются понятия зоны связей словоформы и блока спаянности словоформ.

Зоной связей словоформы (**ЗЗС**) называем ту часть предложения, в которой реализуются все синтаксические связи словоформы. Под словоформой мы понимаем последовательность буквенных символов между двумя пробелами, другими словами — машинное слово.

Зона связей представляется графически — путем проведения отрезков через все словоформы, которые вступают в синтаксическую связь с анализируемой словоформой. Так на рис. 1 зоны связей всех восьми словоформ в предложении обозначены как вертикальные отрезки (подпись у каждого отрезка указывает на словоформу, для которой была построена данная зона, и на порядковый номер словоформы

в предложении). Точки на отрезках соответствуют словоформам, синтаксически связанным со словоформой, для которой строится зона.

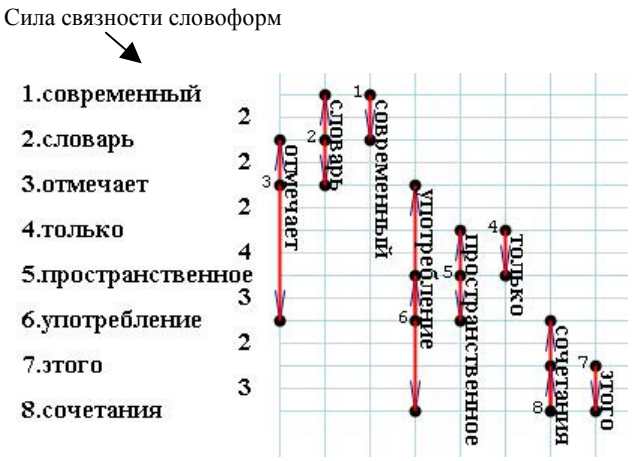


Рис. 1. Зоны связей словоформ в предложении.

Количество отрезков, расположенных в интервале между двумя соседними словоформами, указывает на **силу синтаксической связности** каждой пары словоформ в предложении (например, на рис. 1 сила связности между первой и второй словоформами равняется 2, а между седьмой и восьмой — 3). Тогда **блоком** наибольшей синтаксической **связности** является последовательность словоформ, между которыми сила связности максимальна для данного предложения (на рис. 1 это пара словоформ *только пространственное*, сила связности равна 4). Таким образом, в каждом предложении русского языка можно определить словоформы, которые характеризуются наибольшей степенью синтаксической «спаянности» между собой.

Наличие синтаксической связи между некоторыми элементами предложения свидетельствует также о существовании семантической связи между ними. Соответственно, можно предположить, что блоки синтаксической связности являются, вместе с тем, и блоками семантической связности в предложении. Так, В. И. Перебейнос высказала предположение, что блоки связности в большинстве случаев содержать словоформы, которые особенно важны для смыслового развертывания сообщения [2], а Н. П. Дарчук показала, что выделение блоков связности может стать эффективным инструментом для автоматического определения терминов в научном тексте [1]. Мы, в свою очередь, предположили, что блоки спаянности словоформ могут лечь в основу методики извлечения из текста тех лексико-семантических вариантов, которые наиболее полно отображают его тематику.

Методика предполагает следующие этапы: 1) построение зон связей словоформ для каждого предложения текста; 2) подсчет силы связности словоформ; 3) отбор словоформ, которые попали во все блоки связности текста; 4) лемматизация словоформ; 5) группирование полученных лексем по частям речи, отсеивание служебных лексем; 6) уточнение лексического значения для многозначных лексем (с этого момента исследователь работает с лексико-семантическими вариантами — ЛСВ); 7) присваивание каждому ЛСВ индекса, который указывает на количество раз, когда данный ЛСВ вошел в состав блока связности; 8) группирование ЛСВ одной части речи в лексико-тематическое дерево.

Совокупность лексико-тематических деревьев полнозначных частей речи составляет лексико-тематическую модель текста. Узлы лексико-тематических деревьев (т. е. названия тематических полей, групп, подгрупп), которые заполнены либо бóльшим количеством ЛСВ, либо ЛСВ с высоким индексом, отображают основные темы сообщения. Мало заполненные узлы лексико-тематических деревьев указывают на периферийные темы.

Стоит подчеркнуть, что эффективность методики существенно повышается в случае применения различных средств автоматической обработки текста. Так, приведенный на рис. 1 пример является результатом автоматического построения ЗЗС и подсчета силы связности словоформ на основе дерева зависимостей предложения; лемматизация и сортирование лексем по частям речи стала возможной благодаря работе автоматического морфологического анализатора; а для построения лексико-тематических деревьев понадобилось привлечение электронного тезауруса (идеографического словаря), созданного Н. П. Дарчук [3].

Апробация разработанной автором методики определения тематически значимых ЛСВ на основе блоков связности словоформ, как и созданная лексико-тематическая модель текста, подтвердили теоретическую и практическую значимость понятий зоны связей и блока связности словоформ.

Литература

1. *Дарчук Н. П.* Зона зв'язку слів як засіб автоматичного виділення термінів з тексту // Українське мовознавство. Вип. 17. Київ, 1990. С. 15–20.
2. *Перебейнос В. И.* О единицах текста // Recueil linguistique de Bratislava. Vol. VIII. Bratislava, 1985. P. 146–150.
3. *Darchuk N.* Text-oriented thesaurus retrieval system for Linguistics // Slovko-2009. Materials of Fifth International Conference "NLP, Corpus Linguistics, Corpus Based Grammar Research". Smolenice; Bratislava, Slovakia, 2009.

Формальные и неформальные критерии вычленения ключевых слов из научных и новостных текстов

Е. В. Ягунова

Санкт-Петербургский государственный университет

iagounova_elena@mail.ru, iagounova.elena@gmail.com

Анализ текста, ключевые слова, эксперимент, формальные критерии

Summary. Investigation of key word detection is actually important for both automatic text analysis and discourse analysis. We propose complex investigation method and discuss its applicability for different types of text.

1. Введение

В современном информационном обществе большое внимание уделяется созданию автоматических систем понимания текста, извлечению информации как из отдельно взятого текста, так и из информационных потоков. На настоящее время не решена и гораздо более глобальная проблема: каким образом воспринимает текст человек, какие процедуры он использует для извлечения информации из текста. Извлечение наиболее важной информации из текста может быть смоделировано через процедуры выделения ключевых слов текста. В данном докладе рассматриваются результаты многопрофильного исследования разнообразных процедур извлечения ключевых слов из текста, в этом исследовании объединены существующие подходы и методики — от эксперимента с информантами до вычислительного эксперимента. Одним из важных аспектов работы является анализ зависимости реализуемых процедур выделения ключевых слов от функционального стиля текста. В качестве исходного источника используется батарея экспериментов с носителями языка, в результате которой методом экспертной оценки определяются списки ключевых слов. В ходе вычислительных экспериментов моделируются процедуры, позволяющие на основании формальных и неформальных критериев вычленять ключевые слова.

Одним из существенных положений, лежащих в основе данного исследования, является определяющая роль функционального стиля текста для процедур его анализа [2]. Ранее нами рассматривались предварительные данные по определению ключевых слов для художественного нарратива небольшого объема (повесть, рассказ) [3]. Прикладная направленность данного исследования диктует выбор текстов наиболее востребованных для систем автоматического анализа.

2. Материал и методика

Основным исходным материалом данной работы являются научные (предметной области «корпусная лингвистика») и новостные тексты. Базовая выборка содержит по 10 текстов для каждого функционального стиля (типа).

Первым компонентом цикла является батарея экспериментов с носителями языка, в ходе которой должны быть получены списки ключевых слов (большой, средний и малый наборы по А. С. Штерн и Л. В. Сахарному). Важным аспектом анализа результатов (анкет информантов) является поиск ведущей единицы: словоформы, лексемы, классов условной эквивалентности (подробнее см. [2]).

Второй — основной — компонент исследования представляет из себя вычислительный эксперимент по определению формальных и неформальных параметров выделения ключевых слов, входящих в рассматриваемые списки. Этот компонент включает в себя как анализ (вычислительный эксперимент) самих исследуемых текстов базовой выборки, так и статистическое обследование представительной выборки, однородной исследуемым текстам каждого из стилей. Для возможности статического обследования представительной выборки были созданы коллекции, удовлетворяющие задачам исследования. В качестве формальных критериев, напр., используются такие критерии, как частотность единиц разных уровней в тексте и представительной выборке, степень равномерности распределения этих единиц в тексте и выборках, традиционно используемые в прикладных задачах критерии «уникальности единицы» (прежде всего, TFiDF) и т. д.

3. Обсуждение результатов

В качестве ведущих формальных признаков на данном этапе рассматриваются следующие:

– частота встречаемости в конкретном тексте,
– распределение по тексту (для частотных по тексту):
– равномерность,
– для неравномерных — тяготение к началу / концу текста,
– сопоставление **частота встречаемости в тексте** vs. **частота встречаемости в представительной выборке** vs. **общеязыковая частота встречаемости** (ср. коэффициент TFiDF, традиционно используемый для оценки различительной силы слова текста).

Частота встречаемости определяется для единиц разного типа: лексемы, классов условной эквивалентности, словосочетаний и конструкций.

Для анализа равномерности использованы инструментальные средства, «позволяющие визуализировать плотность встречаемости слова в тексте в зависимости от ширины окна наблюдения. В… спектрограмме по горизонтали откладываются номера вхождения слова в тексте, а по вертикали — ширина окон наблюдения (начиная со значения 1 в самом низу, вхождения слова в данном случае выделяется светло-серым цветом). Если в соответствующее окно наблюдения попадает несколько целевых слов, то оно закрашивается более интенсивным оттенком темного» [1].

В докладе предлагается обсудить данные и сформулировать гипотезы о том,

– какую часть набора КС, выделенных информантами, можно описать через предлагаемый набор формальных признаков,
– какие типы КС можно таким образом описать через набор этих признаков.

Результаты, полученные на материале рассматриваемых научных текстов, характеризуются максимальной однородностью. Подавляющее большинство КС текстов этого типа возможно описать через предлагаемый набор формальных признаков. Гораздо меньшая определенность результатов отличает новостные тексты, что связано с их неоднородностью в отношении структуры текста и его объема. Однако основные типы КС новостных текстов можно описать через предлагаемый набор формальных признаков. В докладе приводится классификация выделяемых типов КС.

В работе уделяется внимание выявлению с помощью предлагаемой методики основных понятий и терминов (как однословных, так и состоящих из двух и более слов).

Полагаем, что полученные результаты актуальны как для теоретической лингвистики текста, так и для прикладных разработок в области автоматических систем анализа текста (автоматического определения КС, информационного и фактографического поиска, рубрицирования текстов и т. д.).

Литература

1. *Ландэ Д. В.* Визуализация статистики вхождения слов // Горизонты прикладной лингвистики и лингвистических технологий: Материалы международной научной конференции 21–26 сентября 2009, Украина. Киев, 2009 (http://ling.infostream.ua/jag/).
2. *Ягунова Е. В.* Вариативность стратегий восприятия звучащего текста (экспериментальное исследование на материале русскоязычных текстов разных функциональных стилей). Пермь, 2008.
3. *Ягунова Е. В.* Визуализация статистики вхождения слов // Горизонты прикладной лингвистики и лингвистических технологий: Материалы международной научной конференции 21–26 сентября 2009, Украина. Киев, 2009.