Министерство образования и науки
Российской Федерации

УНИВЕРСИТЕТ ИТМО

КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА И ВЫЧИСЛИТЕЛЬНЫЕ ОНТОЛОГИИ

Выпуск 2

**Труды XXI Международной
объединенной научной конференции
«Интернет и современное общество», IMS-2018,
Санкт-Петербург, 30 мая – 2 июня 2018 г.**

Сборник научных статей

УНИВЕРСИТЕТ ИТМО

Санкт-Петербург

2018

УДК 800 (075.3)
ББК 81.1
К63

К63 **Компьютерная лингвистика и вычислительные онтологии.** Выпуск 2 (Труды XXI Международной объединенной конференции «Интернет и современное общество, IMS-2018, Санкт-Петербург, 30 мая - 2 июня 2018 г. Сборник научных статей). — СПб: Университет ИТМО, 2018. — 160 с.
ISSN 2541-9781
ISBN 978-5-7577-0584-2

В сборник включены тексты статей, представленные на XXI Международной объединенной конференции «Интернет и современное общество» (Internet and Modern Society - IMS). Работы прошли рецензирование и отобраны в результате конкурсной процедуры. Сборник снабжен авторским указателем.

Издание адресовано научным работникам, преподавателям, аспирантам и магистрантам, изучающим компьютерную лингвистику и вычислительные онтологии, а также междисциплинарные проблемы влияния информационно-коммуникационных технологий на трансформацию социальных отношений в современном обществе.

Информация о конференции «Интернет и современное общество» представлена на сайте объединенной конференции (http://ims.ifmo.ru).

XXI Международная объединенная конференция «Интернет и современное общество» (IMS-2018) проведена при поддержке Российского фонда фундаментальных исследований (проект № 18-07-20031).

УДК 800 (075.3)
ББК 81.1

УНИВЕРСИТЕТ ИТМО

XXI Международная объединенная научная конференция **«Интернет и современное общество»** (IMS-2018)

Санкт-Петербург, 30 мая – 2 июня 2018 года

http://ims.ifmo.ru

Конференция «Интернет и современное общество» (Internet and Modern Society – IMS) проводится в Санкт-Петербурге ежегодно с 1998 года. С 2014 г. конференция проводится как международное научное мероприятие. С 2016 года она объединяет серию различных научных мероприятий и проходит в рамках Недели технологий информационного общества в Санкт-Петербурге.
Конференция является «объединенной», т.к. научная программа конференции объединяет серию специализированных российских и международных научных конференций, симпозиумов, семинаров, круглых столов и других мероприятий, посвященных специальным вопросам развития технологий информационного общества.

Организаторы конференции IMS-2018:
- Университет ИТМО
- Библиотека Российской академии наук

Основные мероприятия Недели технологий информационного общества в Санкт-Петербурге:
- Симпозиум молодых учёных **«Цифровые трансформации: перспективные социально-экономические и гуманитарные исследования»**: 30 июня (рабочий язык – русский) - http://ims.ifmo.ru/ru/pages/24/molodezhnyy_simpozium.htm.
- Мероприятия конференции **«Интернет и современное общество»** (открытие, секции, круглые столы): 31 мая – 2 июня (рабочий язык – русский). Сайт конференции: http://ims.ifmo.ru.
- Международная конференция **«Digital Transformation & Global Society» (DTGS-2018):** 31 мая – 1 июня (рабочий язык – английский). Организаторы: Университет ИТМО и НИУ ВШЭ, Санкт-Петербург. Сайт конференции: http://dtgs-conference.org.
- Международный семинар **«Киберпсихология»** (**Internet Psychology – IntPsy-2018**): 1-2 июня как совместное мероприятие конференций IMS-2018 (рабочий язык – русский) и DTGS-2018 (рабочий язык – английский).
- Международный семинар **«Компьютерная лингвистика»** (**Computational Linguistics – CompLing-2018**): 1-2 июня как совместное мероприятие конференций IMS-2018 (рабочий язык – русский) и DTGS-2018 (рабочий язык – английский).

Все мероприятия проводились в Конгресс-центре Университета ИТМО (Санкт-Петербург, ул. Ломоносова, 9).

ПРОГРАММНЫЙ КОМИТЕТ КОНФЕРЕНЦИИ

Председатель Программного комитета:
Васильев В.Н., докт. техн. наук, чл.-корр. РАН, ректор Санкт-Петербургского национального исследовательского университета информационных технологий, механики и оптики (Университет ИТМО)

Заместители председателя Программного комитета:
Борисов Н.В., докт. физ.-мат. наук, директор Центра дизайна и мультимедиа Университета ИТМО
Чугунов А.В., канд. политич. наук, директор Центра технологий электронного правительства Университета ИТМО

Члены Программного комитета:
Антопольский А.Б., докт. техн. наук, академик РАЕН, главный научный сотрудник Центра по изучению проблем информатики ИНИОН РАН
Бершадский А.М., докт. тех. наук, зав. кафедрой Пензенского государственного университета
Борисов Н.В., докт. физ.-мат. наук, директор Центра дизайна и мультимедиа Университета ИТМО
Бродовская Е.В., докт. полит. наук, заведующая кафедрой МГГУ им. М.А. Шолохова
Будрин А.Г., докт. экон. наук, зав. кафедрой МИК ФТМИ Университета ИТМО
Воеводин Вл.В., член-корр. РАН, заместитель директора НИВЦ Московского государственного университета
им. М.В.Ломоносова
Войскунский А.Е., канд. псих. наук, ст. научный сотрудник Московского государственного университета им. М.В.Ломоносова
Волков Д.В., с.н.с. Института прикладной математики им. М.В.Келдыша РАН, главный редактор журнала "Открытые системы.СУБД"
Горбунов-Посадов М.М., докт. физ.-мат. наук, зав. отделом Института прикладной математики им. М.В.Келдыша РАН
Дятлов С.А., докт. экон. наук, профессор Санкт-Петербургского государственного экономического университета
Захаров В.П., канд. филол. наук, доцент Санкт-Петербургского государственного университета
Казаков В.Г., канд. физ.-мат. наук, декан информационно-технического факультета Новосибирского государственного университета экономики и управления
Каленов Н.Е., докт. техн. наук, директор Библиотеки по естественным наукам РАН
Колпакова Н.В., канд. пед. наук, заместитель директора Библиотеки Российской академии наук
Комалова Л.Р., д-р. филол. наук, ст. научный сотрудник ИНИОН РАН, зам. гл. ред. научного журнала "Человек: Образ и сущность. Гуманитарные аспекты"
Леонов В.П., докт. пед. наук, научный руководитель Библиотеки Российской академии наук

Потапова Р.К., докт. филол. наук, профессор, Московский государственный лингвистический университет
Прокудин Д.Е., докт. филос. наук, доцент Санкт-Петербургского государственного университета
Рогов А.А., докт. техн. наук, заведующий кафедрой Петрозаводского государственного университета
Сморгунов Л.В., докт. филос. наук, заведующий кафедрой СПбГУ
Толстикова И.И., канд. филос. наук, заведующая кафедрой СиГН ФТМИ Университета ИТМО
Федосов А.Ю., докт. пед. наук, профессор кафедры информатики и прикладной математики Российского государственного социального университета
Чугунов А.В., канд. политич. наук, заведующий кафедрой УГИС ФТМИ Университета ИТМО
Юсупов Р.М., член-корр. РАН, директор Санкт-Петербургского института информатики и автоматизации РАН

ОРГАНИЗАЦИОННЫЙ КОМИТЕТ

Сопредседатели оргкомитета
Борисов Н.В., докт. физ.-мат. наук, заведующий кафедрой Санкт-Петербургского государственного университета, директор Центра дизайна и мультимедиа Университета ИТМО
Леонов В.П., докт. пед. наук, научный руководитель Библиотеки Российской академии наук

Члены оргкомитета
Прокудин Д.Е., СПбГУ, Университет ИТМО (зам. председателя оргкомитета)
Скворцова О.В., Библиотека Российской академии наук (зам. председателя оргкомитета)
Чугунов А.В., Университет ИТМО, НП ПРИОР Северо-Запад (ученый секретарь конференции)
Кудрявцева М.В., Университет ИТМО (информационный менеджер конференции)
Белинская М.А., Библиотека Российской академии наук
Карачай В.А., Университет ИТМО
Мбого И.А., СПбГУ, Университет ИТМО
Слободянюк В.Е., СПбГУ
Соловьева Д.В., Университет ИТМО

Информация

Объединенная конференция «Интернет и современное общество»: **http://ims.ifmo.ru**
Международная конференция «Digital Transformation & Global Society» (DTGS): **http://dtgs-conference.org**

Предисловие редактора

В последние годы разработка систем автоматизированной обработки текста и речи стала одной из фундаментальных задач современного информационного общества. Проблемами использования естественного языка в системах автоматической обработки информации занимается компьютерная лингвистика. Создание и использование компьютерных лингвистических ресурсов востребовано в самых разных приложениях. За прошедшие годы в области компьютерной лингвистики были получены значительные научные и практические результаты. Были созданы системы машинного перевода текстов с одних естественных языков на другие, системы автоматизированного поиска информации, системы автоматического анализа и синтеза устной речи и многие другие. Новое большое направление в компьютерной лингвистике связано с Big Data. На сегодня в Интернете хранятся огромные корпусы текстов, из которых нужно вычленять определенную информацию. Лингвистический анализ текстов используется и при создании систем искусственного интеллекта.

И сегодня компьютерная лингвистика – динамически развивающаяся наука. Об ее активности и востребованности свидетельствует большое число национальных и международных исследовательских проектов, крупных международных конференций в этой области, специализированных научных журналов, образовательных проектов, открытие отделений компьютерной и прикладной лингвистики в вузах страны.

Статьи, публикуемые в данном сборнике, представляют собой изложение докладов, сделанных на семинаре «Компьютерная лингвистика и вычислительные онтологии», являющемся частью Международной объединенной конференции «Интернет и современное общество». Они отражают широкую тематику исследований по компьютерной лингвистике и многогранность задач в области автоматической обработки текста и речи и имеют как теоретическое, так и прикладное значение.

Часть докладов была посвящена семантическому и семантико-синтаксическому анализу естественного языка. Речь идет об извлечении семантических отношений из текста, о формализации подходов к установлению связей между понятиями, о выявлении семантических отношений на основе мер ассоциации между лексическими единицами, параметризации семантико-синтаксических связей в тексте, выявлению состоятельных статистик и метрик.

В числе других тем следует назвать формально-языковые модели в поэтике, создание и исследование корпусов текстов, выявление и оценку синтагматической связанности единиц текста, вопросы грамматики предложных конструкций русского языка, исследование политической лексики и диахронические исследования на основе корпусов, тематическое моделирование текстов. Большой интерес вызвали доклады, посвященные исследованию цветообозначений в русском языке на базе представительных корпусов текстов. Важное место во многих докладах занимала проблема оценки качества и достоверности данных, получаемых в ходе исследований.

Семинар показал, что с помощью компьютерной лингвистики многие проблемы решаются на совершенно новом уровне: появляются новые теоретические концепции, повышается качество словарей и грамматик, разрабатываются модули обработки естественного языка для многих практических задач. Конференция продемонстрировала успехи российской компьютерной лингвистики в создании электронных лингвистических ресурсов и разработке систем обработки русского языка.

Однако динамика развития компьютерной лингвистики ставит и новые проблемы и намечает новые рубежи, связанные с ее междисциплинарностью, выходом за пределы собственно лингвистики и информационных технологий. Правомерно, видимо, говорить о новом интегрированном подходе, при котором компьютерная лингвистика взаимодействует с информатикой, когнитологией, психологией. При этом наблюдается ее все более широкое распространение – как инструмента и как метода – на всю сферу

гуманитарных исследований, включая историю, социологию, литературоведение и т.д. Уже сегодня на базе компьютерных технологий и корпусной методологии фактически сформировалась новая наука, культурометрия (culturomics) - исследование культуры человечества, направлений её развития во времени посредством количественного анализа лексических единиц в очень больших корпусах оцифрованных текстов. Социальный и культурный опыт человечества, зафиксированный в текстах, получает инструмент, позволяющий надеяться, что мы научимся автоматически извлекать из текстов знание. Хочется верить, что компьютерная лингвистика вместе с традиционной лингвистикой, психолингвистикой и нейролингвистикой сформируют новую науку – интегрированную эмпирическую лингвистику – которая позволит глубже, чем до сих пор, понять фундаментальную природу языка.

В.П. Захаров

Семантическая структура русских предложно-падежных конструкций

И.В. Азарова, В.П. Захаров, А.Д. Москвина

Санкт-Петербургский государственный университет

i.azarovaspbu.ru, v.zakharov@spbu.ru, moskvina.anya@gmail.com

Аннотация

В публикации рассматривается структура синонимичных и квазисинонимичных семантических отношений для первообразных и производных предлогов русского языка в рамках определенных смысловых рубрик: передачи транзитивных локативных, темпоральных и медиативных значений в структуре предложений и словосочетаний. Приведены статистические характеристики корпусной реализации предложных конструкций, описаны регулярные типы категорий главных и зависимых слов, а также типы семантико-синтаксических конструкций реализации вариантов значения смысловой рубрики.

Ключевые слова: корпусная статистика, семантические рубрики, русские предложные конструкции, предложные значения, транзитив, темпоратив, медиатив

1. Введение

Предлоги в русском языке довольно долго оставались без пристального внимания со стороны специалистов по автоматическому анализу текста: например, они входили в списки «стоп-слов», которые предотвращали их использование в векторных моделях информационного поиска. Если искать ответ на вопрос, почему они по-прежнему игнорируются при информационном поиске, то аргументом в пользу такого подхода будет их низкий показатель «специфичности» (tf-idf): они обладают высокой частотностью практически во всех документах, в которых производится поиск. Однако для семантически ориентированного анализа текста они безусловно важны, поскольку передают четкие семантико-синтаксические отношения между знаменательными словами (*устройство с автономным питанием* vs *устройство без автономного питания*), между характеристиками предиката (*говорить с акцентом* vs *говорить без акцента; переводить с английского* vs *переводить на английский*), между пространственно-временными спецификациями пропозиций (*с 1 января* vs *до 1 января*; *на верхней площадке* vs *под верхней площадкой*). И приведенные выше примеры, и устойчивая корпусная встречаемость предложных конструкций являются четким показателем их важности для построения текстов на русском языке. Вторым важным фактором недостаточного использования предлогов является неясность и неструктурированность их значений, например, для предлога «в» в Викисловаре приводится 21 значение, при этом синонимия предлогов указана частично, неясно, в чем различия близких, но несинонимичных предлогов.

В рамках проекта по квантитативному описанию грамматики предлогов мы начинаем систематическое описание предлогов со значениями из определенных семантических рубрик. В данной статье это конструкции транзитива, темпоратива и медиатива.

Семантические рубрики (термин, близкий к понятию «синтаксема»), представленные в статье, помогают обобщить называемые часто по-разному предложные значения.

2. Инвентарь предложных конструкций

Предлоги в русских текстах неоднородны и разнообразны: есть небольшая группа первообразных предлогов и обширная, с неясными границами, группа производных предлогов, которые мотивированы знаменательными частями речи (существительными, наречиями и глагольными формами деепричастий), возможно, в сочетании с первообразными предлогами. Отсюда корпусная частотность первообразных предлогов обычно завышена за счет того, что часть их используется в структуре производных предлогов. При исчислении последних используются зачастую противоречивые параметры. Во-первых, важно отвлеченное значение существительных (у наречий и деепричастий оно безусловно такое), во-вторых, обязательным признаком производного предлога является непроницаемость линейной последовательности такого грамматикализованного знаменательного слова и первообразного предлога, например *в течение, во исполнение* и др. Косвенным и непоследовательным показателем производного предлога является отличное от формы существительного орфографическое написание: *в течение года* vs *в течении реки; несмотря на непогоду* vs *не смотря по сторонам; во исполнение приказа* vs *в исполнении детского хора.*

Самым важным показателем «предложного» характера сочетаний со знаменательными компонентами будет их полная или частичная функциональная синонимичность первообразным предлогам или «подтвержденным» производным предлогам (предлоги с модифицированной орфографией или просто высокочастотные производные предлоги).

Первообразные предлоги определяются в грамматических источниках весьма нечетко: «Предлог — это служебная часть речи, оформляющая подчинение одного знаменательного слова другому в словосочетании или в предложении и тем самым выражающая отношение друг к другу тех предметов и действий, состояний, признаков, которые этими словами называются» [1, с. 706]. В этом определении опущен важный элемент «синтаксического поведения» предлогов — то, что они оформляют предложно-падежную форму, т. е. могут «оформлять подчинение» только падежных знаменательных слов: в первую очередь, имен существительных, во вторую очередь, местоимений-существительных, числительных и, возможно, прилагательных при их окказиональной субстантивации. Скорректированное определение: «Предлог — служебная часть речи, обозначающая отношение между объектом и субъектом, выражающая синтаксическую зависимость имен существительных, местоимений, числительных от других слов в словосочетаниях и предложениях», представленное в Википедии, может нас еще более запутать указанием на «отношения между объектом и субъектом». Разве только объектные и субъектные позиции могут быть оформлены предложно-падежными конструкциями? Безусловно, обстоятельственные значения превалируют в текстах (и в словосочетаниях, и в предложениях), но ими не исчерпывается весь спектр предложных значений.

В нашем проекте будем считать предлогами «стереотипные способы уточнения падежных значений имен (чаще всего, имен существительных) при выражении валентных позиций знаменательных слов (чаще всего, глаголов и отглагольных дериватов) и/или различных обстоятельственных квалификаторов в предложении». «Стереотипность» значения будет доказываться высокой частотностью выражения определенного значения в корпусе, кроме того синонимия и частичная синонимия между первообразными и производными предлогами будет описана в терминах семантических классов зависимых существительных (оформленных грамматически в виде предложных конструкций) и главных слов, предсказывающих появление предложно-падежных структур.

Первоначально синонимия между первообразными и производными предлогами анализировалась нами по разным словарям, в результате чего были составлены базовые таблицы синонимии (в ряде случаев, естественно, частичной) для предлогов, являющихся объектами анализа в данной статье. При этом еще раз напомним, что значения предлогов во многом являются контекстно-зависимыми.

3. Семантическая структура предложных конструкций в рамках определенной семантической рубрики

Поскольку задачей нашего проекта является создание квантитативной грамматики предложных конструкций, мы используем понятие «семантической рубрики» в качестве обобщенного названия группы значений предлогов. Часто такими рубриками являются названия семантических актантов (или падежей, ролей), однако ввиду разнообразной трактовки этих понятий в литературе и отсутствия четких границ между элементами списка, мы начинаем описание «снизу», от некоторого наиболее частотного базового предлога, вариативность значений которого отражается в группе сопряженных с ним предлогов (в основном, производных), образуя ряды синонимичных или квазисинонимичных конструкций. Поэтому названия семантических рубрик отчасти носят «условный» характер.

Таблица 1. Частотные характеристики предлога «через» и некоторых синонимов

предлог	**значение**	**НКРЯ со снятой омонимией, 6 млн. слов (200 случайных контекстов)**	**НКРЯ газетный, 228,5 млн. слов (200 случайных контекстов)**	**Araneum Russicum веб-корпус 120 млн. (200 случайных контекстов)**
Через	**медиатив**	**ipm 173,53 20,5% (41 употр.) от всех употребл.**	**ipm 221,30 34,5% (69 употр.) от всех употребл.**	**ipm 185,21 32,5% (65 употр.) от всех употребл.**
С помощью	медиатив	ipm 76,63 460 употр.	ipm 113,21 25870 употр.	ipm 228,64 27437 употр.
При помощи	медиатив	ipm 22,32 134 употр.	ipm 27,66 6322 употр.	ipm 82,60 9916 употр.
Посредством	медиатив	ipm 16,49 111употр.	ipm 12,31 2813 употр.	ipm 42,90 5154 употр.
Через посредство	медиатив	ipm 0,99 6 употр.	ipm 0,09 22 употр.	ipm 0,42 50 употр.
Через	**транзитив**	**ipm 245,39 29,5% (57 употр.) от всех употребл.**	**ipm 118,65 18,5% (37 употр.) от всех употребл.**	**ipm 157,71 27,5% (55 употр.) от всех употребл.**
сквозь	транзитив	ipm 119,83 719 употр.	ipm 19,25 4398 употр.	. ipm 23,90 2869 употр
Поперек	транзитив	ipm 12,00 72 употр.	ipm3,15 720 употр.	ipm3,40 408 употр.
Через	**темпоратив**	**ipm 390,96 47% (94 употр.) от всех употребл.**	**ipm 319,95 50% (100 употр.) от всех употребл.**	**ipm 222,25 27,5% (55 употр.) от всех употребл.**
Спустя	темпоратив	ipm 54,67 328 употр.	ipm 82,28 18800 употр.	ipm 52,50 6300 употр.
По истечении	темпоратив	ipm 2,17 13 употр.	ipm 5,81 1328 употр.	ipm 6,88 826употр.
Чрез	темпоратив	ipm 14,00 84 употр.	ipm 0,16 37 употр.	ipm 3,29 395 употр.

Для выявления семантической структуры необходимо выяснить базовое ядерное значение, которое представлено частотным употреблением первообразного падежа, в отдельных случаях производного предлога или характерной падежной формы. Подбор синонимических конструкций производится частично за счет информации, представленной в толковых словарях. Синонимия дополнительно проверяются через выявление характерных конструкций реализации значения в случайных выборочных

совокупностях корпусных контекстов. В случае несовпадения сочетаемости предлогов и семантических классов имен или появления иных ограничений на реализацию значений семантической рубрики они рассматриваются как частичные синонимы, различия в употреблении и значении будут оценены в соответствии с корпусной статистикой (см. табл. 1 и табл. 2).

Таблица 2. Частотные характеристики предлога «по» и некоторых синонимов

предлог	**значение**	**НКРЯ со снятой омонимией, 6 млн. слов (200 случайных контекстов)**	**НКРЯ газетный, 228,5 млн. слов (200 случайных контекстов)**	**Araneum Russicum веб-корпус 120 млн. (200 случайных контекстов)**
По	**медиатив**	**ipm 332,95 7% от всех употребл.**	**ipm 82,57 1% от всех употребл.**	**ipm 237,03 3% от всех употребл.**
Посредством	медиатив	ipm 16,49 111 употр.	ipm 12,31 2813 употр.	ipm 42,90 5154 употр.
Через посредство	медиатив	ipm 0,99 6употр.	ipm 0,09 22употр.	ipm 0,42 50 употр.
По	**транзитив**	**ipm 245,39 16% от всех употребл.**	**ipm 908,35 11% от всех употребл.**	**ipm 869,11 11% от всех употребл.**
По	**темпоратив**	**ipm 390,96 4% от всех употребл.**	**ipm 82,57 1% от всех употребл.**	**ipm474,06 6% от всех употребл.**
после	темпоратив	ipm 810,83 4865 употр.	ipm 1446,41 330504 употр.	ipm 1245,90 149504 употр.
вслед за	темпоратив	ipm 39,00 234употр.	ipm 34,36 7851употр.	ipm 14,53 1744употр.
по истечении	темпоратив	ipm 2,17 13 употр.	ipm 5,81 1328 употр.	ipm 6,88 826 употр.
по завершении	темпоратив	ipm 0,67 4 употр.	ipm 3,68 842 употр.	ipm 2,96 355 употр.
по окончании	темпоратив	ipm 9,33 56 употр.	ipm18,86 4310 употр.	ipm 16,80 2916 употр.

3.1. Структура предложных конструкций медиатива

Медиатив как семантическая рубрика имеет узкую и широкую интерпретацию. В большинстве случаев она рассматривается как определенная семантическая роль в предикативной структуре глагола. В узком смысле медиатив понимается как средство, т. е. вещество или предмет, которые расходуются при выполнении действия или процесса [2]. В более широком понимании медиатив является инструментом и включает их вещественные и отвлеченные реализации [3]. В русском языке инструмент и медиатив, как правило, выражаются формой творительного падежа (*красить стены валиком, рисовать картину красками*), однако в виде предложной конструкции мы видим более сложные варианты их синкретизма.

Предлог «через» используется в рубрике медиатива с высокой корпусной частотой 163 ipm в корпусе «Бокренок» кафедры математической лингвистики СПбГУ, при этом нюансы его значений весьма разнообразны: *гладить брюки через влажную ткань* (инструмент *утюг*); *настроить протокол TCP/IP через вкладку «Конфигурация»; ультразвук воздействует на организм через воздух*; *ставить горчичники через газету*; *отмывать деньги через другие фирмы; протереть творог через дуршлаг; получить кредит через знакомых* и др. Во всех таких случаях четкая обстоятельственная

характеристика затруднительна, а зачастую — амбивалентна. Например медиативная конструкция *пропустить мясо через мясорубку* одновременно указывает также на реальное перемещение объекта (см. транзитив ниже). Первое значение предлога «через» в Викисловаре трактуется как конструкция с транзитивным значением при помощи синонимов «сквозь, поперек», приводится пример *перевести женщину через дорогу,* но ни один из указанных предлогов мы не можем подставить в этот пример, а в конструкции с амбивалентным значением «через», наоборот, можем заменить его на «сквозь»: *протереть творог через/сквозь сито/дуршлаг*, аналогично в конструкциях наблюдения *видеть через/сквозь стекло.*

Предлог «сквозь» рассматривается как производный (отнаречный), хотя употреблений «сквозь» в качестве наречия в корпусе просто нет. Частотность предлога «сквозь» в медиативно-транзитивном употреблении существенно ниже (20 ipm), чем «через». В большинстве случаев «сквозь» вряд ли может быть заменен на «через»: *видеть сквозь дымку / туман / крону деревьев*. В указанных контекстах объекты, вводимые предлогом «сквозь», обозначают не столько то, что способствует осуществлению действия, сколько то, что не смогло помешать ему. Обозначим такое употребление как медиативно-неактивное. Это и другие наблюдения говорят о том, что, возможно, номенклатура предложных значений требует расширения. Менее частотным (7 ipm) является употребление «сквозь» для (обычно) слухового восприятия на фоне каких-то помех: *слышать смех /крик сквозь шум / телефонный звонок / сон.*

Частотные производные предлоги в данной рубрике мотивированы существительным «помощь» в комбинации с первообразными предлогами: *с помощью* (98 ipm), *при помощи* (27 ipm). В качестве наиболее частотного семантического класса выступают обозначения абстрактных понятий (*игры, техники, приема, публикации* и т.д.) и реальных предметов-инструментов или устройств (*зеркала, микроскопа, проволоки* и т.д.), обозначения людей-помощников медиативно-активных (*друзей, секундантов* и т.д.).

Аналогичный низкочастотный предлог на базе лексемы «помощь» — *без помощи*, своеобразный антоним для предлогов *с помощью* и *при помощи.*

Производный предлог «путем» обладает высокой частотностью (51 ipm), присоединяет абстрактные понятия: *путем угроз / обмана / обещаний* и т.д., тяготеет стилистически к официально-деловым текстам. Вообще говоря, особо следует изучить поведение одних и тех же предлогов (предложных значений) в текстах разных функциональных стилей.

Еще одна знаменательная лексема «посредство» мотивирует целый ряд производных предлогов, среди которых есть один частотный *посредством* (19 ipm) и целая группа низкочастотных (*при посредстве, через посредство, благодаря посредству*). Как и в случае с наречием «сквозь», само существительное «посредство» в корпусах или не представлено, или имеет крайне низкую частоту, не сопоставимую с частотностью предлога. Предлог «посредством» присоединяет в равной степени как абстрактные понятия (*посредством воображения / скобочной записи* и т.д.), так и обозначения предметов-инструментов и веществ (*посредством специальных приборов / электромузыкальных инструментов / анилиновой краски / нажатия правой кнопки мыши* и т. д.)

Частотный производный предлог «благодаря» (71 ipm) обладает сходной сочетаемостью с приведенными выше предлогами (*благодаря деятельности / сыну / судье / компьютеру*), его традиционно относят к другой рубрике — каузативу. Амбивалентность инструментатива (и, следовательно, медиатива) неоднократно подчеркивалась [3], однако в нашем проекте мы будем рассматривать каузативные предлоги в виде самостоятельной рубрики. При этом предлог «благодаря» тяготеет к данной рубрике, что подтверждается производным образованием « *благодаря посредству*».

Отдельная задача, которая может решаться «попутно» — выявление нерегулярных (фразеологизированных, метафоризированных) употреблений предлогов. В первую очередь так используются первообразные предлоги и некоторые производные. В

частности для центра нашей рубрики «через» и «сквозь» имеется довольно большое число употреблений «со сдвигом» значений, которые частично упоминаются в толковых словарях:

бить ***через край*** — «слишком сильно»;

обратиться ***через голову X*** — «минуя X в социальной иерархии»;

переступить ***через себя*** **—** *«*преодолеть в себе какое-то чувство»;

пропустить ***через себя*** **—** *«*осмыслить»;

сматывать ***через локоть*** **—** *«*на локоть»;

улыбнуться ***через силу*** — *«*превозмогая себя»;

готов/ хочет /... ***провалиться сквозь землю*** **—** *«*скрыться, испытывая чувство стыда»;

смотреть сквозь пальцы на X **—** *«*делать вид, что не замечаешь X»;

цедить/ говорить ***сквозь зубы*** **—** *«*говорить неотчетливо / нехотя, выражая недоброжелательное отношение к собеседнику».

3.2. Структура предложных конструкций транзитива

Транзитив — один из вариантов локализации пропозиции. В отличие от характеристики местоположения, которая применима практически к разнообразным действиям, состояниям и процессам, эта спецификация чаще всего ассоциируется с «рамочной» конструкций в паре с префиксом *пере-* для глаголов движения и их дериватов: *перейти через дорогу, перевозки нефти через Атлантику* и т. д. Аналогично предыдущей рубрике (медиативу) значение транзитива может быть выражено падежной формой, ср.: *перейти дорогу / площадь / реку через дорогу*, причем, когда возможны такие варианты выражения падежный вариант более частотный. Таким образом, для нас важно обозначить случаи, когда значение транзитива почему-то выражается только предложно-падежной формой. В Викисловаре приводятся 2 значения, которые можно связать с семантической рубрикой «транзитива»: (1) «сквозь, поперёк» (*Он помог слепой женщине перейти через дорогу на другую сторону*); (2) «поверх чего-либо» (*Я легко перепрыгнул через забор*). В примере для первого значения невозможно заменить «через» на приводимые синонимы, что указывает, что данная формулировка неточна. В МАС [4] первое значение «через» сформулированы точно: «1. Употребляется при обозначении какого-л. пространства, места и т.п., поперек которого располагается что-л., с одной стороны которого на другую совершается движение, действие» (*Перейти через улицу. Переправиться через реку. Мост через Неву*). То есть это расположение, движение или действие, в которое вовлекается объект, имеющий сторону меньшего размера. Лексический эквивалент значения при глаголах движения — «пересекая» для приглагольных и «пересекает» для присубстантивных употреблений: *воздух проходит через нос; пройти через вращающиеся двери*; для глаголов перемещения объекта эквивалент — «перемещая»: Предлог «сквозь» синонимичен отчасти: идея поперечника не существенна, важен пространственный характер того, что пересекается: *пройти / пролететь / прохождение) сквозь (атмосферу / толпу / пространство)*. В Википедии в качестве синонима для «сквозь» приводится «насквозь», однако это слово регулярно употребляется при тех же глаголах движения, но обозначает характеристику действия, что подчеркивается совместной встречаемостью с предлогом «через»: *винт проходит насквозь через гриф, мы прошли насквозь через весь холл.*

Употребление «сквозь» и «через» при глаголах, обозначающих перемещение с преодолением препятствия, имеет лексикализацию «проникая»: *росток пробился сквозь асфальт.*

3.3. Структура предложных конструкций темпоратива

Темпоратив — такая же сложная семантическая рубрика, как и локатив, поэтому все варианты ее реализации сразу представить довольно сложно. Предлог «через» обозначает

«по истечении некоторого отрезка времени». Это самое частотное употребление предлога (285 ipm в корпусе «Бокренок» кафедры математической лингвистики СПбГУ), причем оно мало зависит от выражаемой пропозиции, регулярно стоит в позиции детерминанта в абсолютном начале предложения и присоединяет существительные, обозначающие временные интервалы, в сочетании с числительными: *через день / 2 дня / 10 минут / 20 столетий*. Эквивалентный производный предлог — «спустя». Он употребляется с теми же существительными, но помимо предложной позиции (до группы существительного) может стоять после группы как послелог: *спустя несколько лет / несколько лет спустя; спустя три недели / три недели спустя* и т. д.

Производный крайне частотный (825 ipm) предлог «после» толкуется примерно таким же образом «по окончании чего-либо», но в подавляющем числе случаев указывает следование относительно некоторого события (*после обеда / прогулки / инъекции* и т. д.) и даже если присоединяют существительное со значением интервала времени, то, как правило, оно используется в событийном значении: *после своего рабочего дня / десятичасового замачивания*.

Производный предлог «по истечении» (3 ipm) сочетается с существительными по модели «через» (*по истечении года / месяца / двух лет*), «по окончании» (9 ipm) следует модели «после» и лишь небольшое число сочетаний аналогичны «через»: *по окончании месяца / года.* Приводимый в Викисловаре вариант перефразирования «вслед за» используется в основном для указания на несамостоятельное движение или аналогично «после» указывает на завершение события (2 ipm): *вслед за заключением мира*.

Заключение

Семантические рубрики, представленные в статье, помогают организовать довольно расплывчатые предложные значения. Их близость и различие могут быть объяснены путем анализа семантических классов управляющих и подчиненных слов («хозяева» и «слуги»). Но это уже тема отдельной статьи. Вся структура предложных частот и распределения семантических (лексических) классов призваны стать базой для создания квантитативной грамматики предложных конструкций русского языка.

Следующие шаги в реализации проекта — анализ всего множества предлогов с разбиением их по семантическим рубрикам и составление реестра «хозяев» и «слуг» в терминах лексико-семантических классов и грамматических категорий.

Отдельная задача — описание синонимии первичных и производных предлогов как некоторого градуированного пространства в терминах семантических контекстов.

Благодарности. Работа поддержана грантом РФФИ № 17-29-09159 «Квантитативная грамматика русских предложных конструкций».

Литература

[1] Грамматика русского языка. М., 1980.
[2] Всеволодова М.В., Потапова Г.В. Способы выражения временных отношений в современном русском языке. М., 1975.
[3] Мустайоки, А. Теория функционального синтаксиса: от семантических структур к языковым средствам. М., 2006.
[4] Словарь русского языка в 4 томах. Т. 4 / Под ред. А.П. Евгеньевой и Г.А. Разумниковой. Изд. 3-е, стереотип. М., 1988.

Semantic Structure of Russian Prepositional Constructions

I.V. Azarova, V.P. Zakharov, A.D. Moskvina

Saint Petersburg State University

In this paper we present a technique for specification of Russian prepositional constructions. We bring together a number of primary and secondary prepositions expressing roughly similar relations between the main and the subordinate full words. The whole group of prepositional constructions demonstrates several implicit variants of the broad sense rubric, which we nominate the semantic domain. In the paper we describe 3 such domains: temporal, transitive and mediative. For each domain we range appropriate prepositions according corpus statistics, show contextual restrictions and subtle nuances of the dominant rubric meaning, which are expressed in collocations with particular semantic types of lexemes occupying head and daughter positions of prepositional relations. The most ambivalent domain described is mediative. Usually this meaning is connected with a notion of an instrument, however, it is complicated by notions of material, device, and even an assistant and space transition. The temporal domain is described partially as a reference to the time moment, the other temporal relations will be the matter of the future research. The transitive domain is a part of the heterogeneous locative relations, which are to be the developed later.

Keywords: corpus lexical statistics, semantic domain, Russian prepositional constructions, prepositional meanings, transitive prepositional constructions, temporal constructions, mediative constructions

Извлечение семантических отношений для создания предметного тезауруса

М.С. Каряева

Ярославский государственный университет

Mari.Karyaeva@gmail.com

Аннотация

Данная работа посвящена разработке методов для автоматического построения тезауруса. Под тезаурусом обычно понимают словарь концептов с определенной структурой хранения данных и набором семантических отношений. Особое значение наряду с отношением основным отношением, используемым в тезаурусах, синонимией, мы исследовали такие виды семантических отношений как часть-целое, род-вид и отношение ассоциации.

В качестве предметной области для проведения первичных исследований была выбрана поэтология. Под поэтологией понимается группа дисциплин, ориентированных на всестороннее теоретическое и историческое изучение поэзии. Основным объектом изучения предметной области «Поэтология» является стихотворное произведение того или иного автора. На прошлом этапе была создана терминологическая коллекция (32 тыс. терминов), путем извлечения терминов различной длины из оцифрованных источников. В качестве начального этапа для задания предметной области служил словник (1,5 тыс. терминов), составленный вручную экспертами предметной области.

Ключевые слова: тезаурус, семантические отношения, data mining, knowledge acquisition, domain engineering, machine learning, word2vec.

1. Данные

Выбор базы знаний для извлечения семантических отношений является приоритетным по отношению к качеству результата. Источникам для извлечения отношений при построении тезаурусов могут служить тексты [1], словари [2], Википедия [3–6] и результат выдачи поисковых запросов [7–9]. На основании опыта представленного в вышеперечисленных работах в данном исследовании были использованы следующие источники данных:

- Оцифрованные источники по предметной области. Оцифрованные источники включают в себя именно ту терминологическую составляющую, которую не встретить в источниках с общеупотребительной лексикой. Поэтому их использование обязательно с точки зрения включения в тезаурус специфичных и редких терминов предметной области.
- Википедия. Поскольку Википедия имеет упорядоченную структуру, заданную Вики-разметкой, были использованы:
  - категории Википедии;
  - внутренние ссылки;
  - объяснение-определение — первый абзац статьей из Википедии, который расположен сразу после объявления термина (заголовка отдельно взятой статьи).

Выбранная предметная область наглядно демонстрирует редкость и специфичность терминологической базы, таким образом, возможно дать общую оценку использования

Википедии в качестве базы знаний для задачи по извлечению семантических отношений. На примере работы [10] можно сделать вывод о том, что Википедия действительно может быть применима в качестве базы знаний для построения предметно-специализированных тезаурусов для английского языка. Тезаурус по сельскому хозяйству был автоматически сгенерирован на основе статей Википедии с поддержкой семантических отношений. Кроме того, было обнаружено, что качество автоматически сгенерированного тезауруса превосходит тезаурус, разработанный до этого вручную.

Перед использованием русской Википедии был проведен эксперимент для определения границы встречаемости терминов из словника и наличия их в Википедии в виде заголовка статьи. 54% терминов присутствуют в виде заголовка статей Википедии. Такой малочисленный результат говорит о том, что эксперты указали необщеупотребительные и многословные термины. Количество однословных терминов из словника в заголовках Википедии присутствует в 72%, что подтверждает этот факт. Дополнительно аналогичным образом были проверены оцифрованные источники, которые были лемматизированы для проведения эксперимента. Результат присутствия терминов из словника оказался выше по сравнению с Википедией (результаты эксперимента представлены в таблице 1).

Таблица 1. Встречаемость терминов в источниках данных

Источник	**Википедия, заголовки**		**Оцифрованные источники**
	Всего	**в т.ч. однословных**	
Словник	54%	72%	94%

2. Используемые подходы

Основной идеей для создания тезауруса служит подход к комбинированию как ручных (на начальных этапах), так и автоматических методов для формирования связей между терминами тезауруса. Изучение основных закономерностей при построении тезауруса поможет формализовать основные шаги и методики для создания баз знаний и способствовать развитию построения автоматическими методами тезаурусов в других предметных областях.

Кроме того, важно отметить, что проект посвящен автоматизации тезауруса, где термины представляют собой не только однословные концепты, но и многословные. Поскольку очевидно, что однословные термины не в состоянии сформировать полный тезаурус. В данном исследовании многословные и однословные термины будут рассматриваться одновременно, в алгоритмах будут по умолчанию использоваться N-граммы <=4 (то есть длина термина не будет превышать 5 слов).

3. Родо-видовые отношения

Существует два основных подхода к извлечению гипо-гиперонимических отношений: на основе шаблонов и на основе статистических (дистрибутивных) методов. Задачей распознавания иерархических отношений между словами занимаются, начиная с 1990-х годов. Один из первых предложенных методов [11] позволял извлекать отношения между терминами из корпуса текстов с использованием лексико-синтаксических шаблонов для английского языка. Дальнейшее развитие автоматического извлечения родо-видовых отношений происходит за счет использования сочетания лексико-синтаксических шаблонов с алгоритмами машинного обучения. В исследовании [12] лежит идея нахождения пар гипоним-гипероним на основе векторного представления: сначала разности векторов кластеризуются (по мысли авторов, кластеры отражают разные типы иерархических отношений), после этого для каждого кластера обучается отдельная проекция вектора на основе обучающей выборки, полученной из тезауруса. После получения кластерной

структуры и обучения операторов проекции, каждая пара векторов, соответствующих словам, может быть классифицирована как “род-вид” или нет.

Опираясь на опыт коллег, были выбраны следующие подходы для извлечения родо-видовых отношений:

- подход на основе лексико-синтаксических шаблонов;
- подход на основе векторного представления слов.

В работе [13] представлен метод извлечения родо-видовых отношений из толкового словаря русского языка посредством использования лексико-синтаксических шаблонов для русского языка. За основу метода принята идея о том, что в определении к понятию содержится кандидат, обозначающий родовое понятие по отношению к определяемому слову. Часто определение уточняет определяемое слово за счет указания в определении рода с сочетанием уникальных качеств понятия. Для автоматического извлечения родо-видовых отношений был разработан набор шаблонов (извлечение первого существительного в определении в качестве родового признака; наличие слов-индикаторов «вид», «разновидность» и т.д.; использование отсылочных конструкций и т.д.). Точность извлечения составляла 0.65.

Мы улучшили точность извлечения до 0.71 за счет усложнения шаблонов, путем добавления многословного извлечения видового понятия, поскольку описанный выше метод не предусматривает наличие родовых понятий, состоящих из двух или более слов. Расширение представленных в работе шаблонов было выполнено за счет включения в шаблоны следующих правил:

- извлечение одного слова, которое является родом, из определения посредством основного алгоритма;
- формирование однословного или многословного видового термина;
- проверка частотности пары, полученной на шаге 2.

Оптимизированный метод был апробирован на статьях русской Википедии. Для эксперимента из всех статей был извлечен первый абзац, который выступает в качестве определения, а название статьи — в роли родового понятия.

Разработанные шаблоны (N* — существительное в им.п., которое является ядром в родовом понятии; A — прилагательное, P — предлог):

- AN* {Силлепс — стилистическая фигура};
- N*N {Тоническое стихосложение — система стихосложения};
- AAN* {Акцидентный набор — малые наборные рифмы};
- N*AN {Парцелляция — конструкция экспрессивного синтаксиса};
- AN*PN {Дастан — эпическое произведение в фольклоре};
- N*NNN {Риторическое восклицание — прием передачи кульминации чувств};
- N*NPN {Строфа — сочетание строк в стихотворении}.

Алгоритм выполнялся следующим образом: если ни один из данных шаблонов не используется, то родовым понятием служит однословное существительное. Ряд шаблонов, таких как N*AN, N*NAN, был исключен из-за своей непригодности для использования в предметных областях, хотя проверено, что данные шаблоны подходят для извлечения общеупотребительных родовых понятий.

Результаты использования метода с использованием лексико-синтаксических шаблонов приведены в таблице 2. Данный способ не только устанавливает семантические отношения между существующими терминами, но и расширяет количество имеющихся терминов. Поскольку при извлечении лексико-синтаксическими шаблонами термин из описания определения может быть новым относительно существующего списка терминов. Таким образом, терминологическая база пополнилась 3,7 тыс. новыми терминами.

Несмотря на успешные попытки использования word2vec для детектирования родо-видовых отношений для общеупотребительных слов в работе [12], для предметных областей использование данного подхода оказалось неудовлетворительным. Обучение модели word2vec проводилось на корпусе русской Википедии. Общий дамп статей русской

Википедии составляет 16 ГБ. Эксперимент с векторным представлением слов не дал положительных результатов, поскольку в 80% случаев не содержал родо-видовых отношений между кандидатами из ассоциативного ряда и главного слова.

Таблица 2. Извлечение родо-видовых отношений
(подход на основе лексико-синтаксических шаблонов)

Шаблоны	Википедия	Пример
N	29%	Рифма – созвучие
AN	23%	Силлепс – стилистическая фигура
NN	17%	Тоническое стихосложение – система стихосложения
AAN	9%	Акцидентный набор – малые наборные рифмы
NAN	8%	Парцелляция – конструкция экспрессивного синтаксиса
ANPN	5%	Дастан – эпическое произведение в фольклоре
NNNN	3%	Риторическое восклицание – прием передачи кульминации чувств
NNPN	6%	Строфа – сочетание строк в стихотворении

4. Отношение часть-целое

Наличие в тезаурусе отношения «часть-целое» (part-whole), или отношения меронимии (мероним – понятие, которое является составной частью другого понятия), позволяет улучшать качество систем информационного поиска, вопросно-ответных систем, других систем анализа текста, поскольку отношение меронимии обладает свойством транзитивности. То есть части частей являются частями вышестоящего целого. Стоит отметить, что не все термины обладают отношением меронимии, особенно это характерно для литературоведческой терминологии, например, термин «литературный стиль» не имеет частей, однако содержит виды, то есть выступает в качестве гиперонима в родо-видовых отношениях.

Шаблон «часть» при извлечении из предметных источников (не словарей) не является показателем для отношения меронимии. В большинстве случаев, в словах-индикаторах «часть» используется для обозначения доли чего-либо, в отличие от словарей, где данное слово-индикатор раскрывает напрямую отношение части-целого.

Ниже приведены выдержки из предметных источников, где использование шаблона «часть» не применимо для извлечения отношений меронимии. Однако, несмотря на это, велика вероятность наличия термина предметной области после существительного «часть».

- *«Начальная **часть** элегии развивает аналогии из быта…»* [Гаспаров. Избранные труды,1997].
- *«…здесь сложилась трехчастная структура, определяющая жанр гимна: **часть** «призывательная» — с именованием божества и развернутым описанием его; часть «повествовательная»* — с изложением какого-либо мифа о нем;» [Гаспаров. Избранные труды,1997].
- *«Миф — главное средство утверждения события в оде; поэтому чаще всего он занимает главную, серединную **часть** оды.»* [Гаспаров. Избранные труды,1997].
- *«Вот лишь малая **часть** идиом и клише, подвергшихся искажению либо остранению в «Вечерах на хуторе близ Диканьки»»* [Шапир, «Язык-Стих-Смысл русской поэзии 18-20вв.», 2000].
- *«Итак, медитативная **часть** «фабулы» динамична.»* [Тарановский, «О поэзии и поэтике», 2000].

5. Отношение часть-целое

5.1. Гипотеза по извлечению отношения меронимии

Если два термина расположены последовательно <термин 1> <термин 2> и <термин 1> стоит в именительном падеже (или любом другом падеже — проверить), а <термин 2> — в родительном падеже, то они образуют отношения меронимии, где <термин 1> — является меронимом (частью), а <термин 2> - холонимом (целым).

Примеры: концовка литературного произведения: <концовка> <литературного произведения>.

Данная гипотеза была проверена на источниках предметной области. Для реализации эксперимента на подготовленном наборе данных была использована автоматическая разметка терминов, где в качестве терминов выступили термины словника.

Этапы:

- нормализация набора данных (удаление специальных символов, приведение слов в начальную форму);
- приведение терминов в нормальную форму;
- разметка в нормализованном наборе данных нормализованных терминов;
- перенос размеченных терминов в наборе данных в ненормализованный набор данных;
- установка рода для терминов в ненормализованном наборе данных;
- извлечение пары терминов, удовлетворяющих гипотезе.

В таблице 3 представлены результаты применения гипотезы по извлечению меронимии в коллекции документов предметной области. Для оценки результатов была проведена проверка наличия отношения меронимии между парами терминов экспертами предметной области. Экспертам был предоставлен проверочный набор 100 случайно выбранных пар, пары могли быть отнесены к трем категориям: «наличие отношения меронимии» (1), «термины не имеют отношений»(2), «присутствует семантическое отношение, но не отношение меронимии»(3).

Таблица 3.Оценка извлечения отношения часть-целое

Типы конструкций	**Экспертная оценка**		
	1	**2**	**3**
<им.п> <р.п.>	63%	37%	8%

5.2. Критерий проверки

Был разработан критерий проверки автоматического извлечения отношений части-целого: *если объект состоит из частей, то при извлечении количество меронимов должно быть больше, чем один.*

6. Отношение синонимии

Отношение синонимии является одним из основных видов семантических отношений, применяемых в базах знаний, таких как тезаурусы и онтологии. В тезаурусах с помощью отношения синонимии термины группируются в синсеты. Синонимами являются слова близкие по значению, при этом замена одного термина на другой, при наличии между ними отношения синонимии, не изменит смысла содержимого. Отношение синонимии широко используется в задачах NLP для информационного поиска, извлечения смысла, в рекомендательных системах.

Для предметных областей, в отличие от общеупотребительных слов, наличие синонимов значительно меньше. Поскольку дублирование смыслового значения специфичных терминов может привести к неоднозначности восприятия концепта. Несмотря на большое

количество работ по извлечению синонимов, большинство из них направлено на работу с общей лексикой, без использования предметных областей. Таким образом, на наш взгляд не является возможным использовать методы машинного обучения на данном этапе, рискуя качеством при итак немногочисленном наличии синонимии предметной области. Хотя, данный момент требует проведения серьезного ряда экспериментов для доказательства, это будет выполнено в рамках оптимизации и повышения качества тезауруса на следующем этапе.

Для получения высокого качества были использованы лексико-синтаксические шаблоны, которые были применены к толковым словарям предметной области:

- Краткая литературная энциклопедия: В 9 т. – М.: Сов. Энцикл., 1962–1978.
- Литературная энциклопедия: В 11 т. – М.: Ком. акад., 1929–1939.
- Словарь литературных терминов: В 2-х т. – М., Л.: Изд-во Л. Д. Френкель, 1925.
- Квятковский А.П. Поэтический словарь. – М.: Советская энциклопедия, 1966.

Одним из лучших ресурсов по извлечению синонимии оказался «Поэтический словарь» Квятковского. Были использованы конструкции отсылочные или уравнительные конструкции следующего вида:

<Термин 1> — см. <Термин 2>.

<Термин 1> (или <Термин 2>) — …

Извлечено порядка 1,3 тыс. пар с отношением синонимии, экспертная оценка 74% точности.

7. Отношение ассоциации

Отношение ассоциации трудно определяемо математическими стандартами. Отношение ассоциации неиерархично, согласно ГОСТ 7.25-2001 «4.5.6.11 ассоциативное отношение является объединением отношений, не входящих в иерархические отношения или в отношения синонимии. Допускается включать в ассоциативное отношение все виды отношений, кроме синонимии и отношения род-вид». В нашем случае, точнее рассматривать отношение ассоциации, полагаясь не только на исключение по синонимии и родо-видовым отношениям, но и отношению часть-целое.

Для разработки алгоритма по автоматическому извлечению отношения ассоциации достаточно трудно разработать гипотезы, поскольку отношение ассоциации трудно формализуемо. В отличие от родо-видовых отношений или отношения часть-целое, которые явно встречаются в толковых словарях и справочниках при задании определенных лексико-синтаксических шаблонов, определить правила извлечения отношения ассоциации таким образом не представляется возможным.

Википедия является уникальным ресурсом для проведения экспериментов по извлечению знаний. Кроме того, в работе [14] было доказано, что Википедия подходит в качестве базы знаний для автоматического извлечения отношения ассоциации. Существует различные подходы использования Википедии для извлечения отношения ассоциации, например, использование категорий Википедии [15], работу с ссылками [16].

В нашей работе основным источником для извлечения отношений ассоциации служит Википедия. Разработанный нами метод будет настроен на работу с ссылками в статьях. Текст статей Википедии содержит внутренние ссылки на другие статьи Википедии, название ссылки полностью совпадает с заголовком перенаправленной статьи. Данный вид ссылок обозначен специализированной разметкой Википедии, что позволяет явно отличать их от ссылок на другие ресурсы не относящиеся к Википедии. Термины Википедии были представлены в виде вершин графа, а ребрами графа являются ссылки внутри статьи термина. Стоит заметить, что структура, состоящая из более 1 млн. вершин плохо масштабируема и затратна по времени. В связи с этим был применен алгоритм поиска компоненты связности графа, той компоненты, которая содержит термины предметной области. Далее, работа происходила с выделенной компонентой, поскольку ссылки на

ассоциативные термины сохранены в данной компоненте. Задача сводилась к тому, чтобы найти путь между двумя вершинами, если этот путь находился, то термины ассоциативны. Длина пути и количество путей между двумя вершинами являются признаками для ранжирования по мере ассоциативности. Были использованы 2 меры TF-IDF и мера Жаккарда. Последний этап алгоритма включал в себя очистку пар отношения род-вид, часть-целое и отношения синонимии из полученных пар ассоциативных отношений. Результат был оценен экспертами предметной области, см. таблицу 4.

Таблица 4. Экспертная оценка автоматического извлечения ассоциативных отношений

Оценка экспертов 100 случайно выбранных пар	**Наличие заданного отношения**	**Термины не имеют заданного отношения**	**Присутствует семантическое отношение, но не заданное отношение**
	75%	11%	14%

Литература

[1] Shimohata M., Sumita E. Acquiring synonyms from monolingual comparable texts // Proceedings of Second International Joint Conference on Natural Language Processing. 2005. P. 233-244.

[2] Wang T., Hirst G. Extracting synonyms from dictionary definitions // Proceedings of International Conference on Recent Advances in Natural Language Processing. 2009. P. 471-477.

[3] Bohn C., Norvag K. Extracting named entities and synonyms from wikipedia // Proceedings of International Conference on Advanced Information Networking and Applications. 2010. P. 1300-1307.

[4] Milne D., Medelyan O., Witten I. H. Mining domain-specific thesauri from wikipedia: a case study // Proceedings of Winter Simulation Conference. 2006. P. 442-448.

[5] Navarro E., Sajous F., Gaume B. Wiktionary and NLP: improving synonymy networks // Proceedings of Workshop on the People's Web Meets NLP. 2009. P. 19-27.

[6] Weale T., Brew C., Lussier E. F. Using the wiktionary graph structure for synonym detection // Proceedings of Workshop on the People's Web Meets NLP. 2009. P. 28-31.

[7] Chakrabarti K., Chaudhuri S., Cheng T., Xin D. A framework for robust discovery of entity synonyms // Proceedings of the 18th ACM International Conference on Knowledge Discovery and Data Mining. 2012. P. 1384-1392.

[8] Chaudhuri S., Ganti V., Xin D. Exploiting web search to generate synonyms for entities // Proceedings of International World Wide Web Conference. 2009. P. 151-160.

[9] Wei X., Peng F., Tseng H., Lu Y., Dumoulin B. Context sensitive synonym discovery for web search queries // Proceedings of the 18th ACM Conference on Information and Knowledge Management. 2009. P. 1585-1588.

[10] Milne D., Medelyan O., Witten I. H. Mining domain-specific thesauri from wikipedia: A case study // Proceedings of the 2006 IEEE/WIC/ACM international conference on web intelligence. EEE Computer Society, 2006. P. 442-448.

[11] Hearst M. A. Automatic acquisition of hyponyms from large text corpora // Proceedings of the 14th conference on Computational linguistics. Vol. 2. Association for Computational Linguistics, 1992. P. 539-545.

[12] Fu R. et al. Learning semantic hierarchies via word embeddings // Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Volume 1: Long Papers. 2014. Vol. 1. P. 1199-1209.

[13] Киселёв Ю. А., Поршнев С. В., Мухин М. Ю. Метод извлечения родовидовых отношений между существительными из определений толковых словарей // Программная инженерия. 2015. Вып. 10. С. 38—48.

[14]Takahiro H. et al. A thesaurus construction method from large scaleweb dictionaries // Advanced Information Networking and Applications. 2007. AINA'07. 21st International Conference on. IEEE, 2007. P. 932-939.
[15]Voss J. Collaborative thesaurus tagging the Wikipedia way // arXiv preprint cs/0604036. 2006.
[16]Nakayama K., Hara T., Nishio S. Wikipedia mining for an association web thesaurus construction // Web Information Systems Engineering – WISE 2007. 2007. P. 322-334.
[17] Лукашевич Н. В. Тезаурусы в задачах информационного поиска М.: Издательство МГУ, 2011.
[18]Добров Б. В., Лукашевич Н. В. Лингвистическая онтология по естественным наукам и технологиям для приложений в сфере информационного поиска // Ученые записки Казанского университета. Серия Физико-математические науки. 2007. Т. 149, №. 2.
[19]Гельфейнбейн, И.Г. Автоматический перевод семантической сети wordnet на русский язык / И.Г. Гельфейнбейн, А.В. Гончарук, В.П. Лехельт, А.А. Липатов, В.В. Шило // Компьютерная лингвистика и интеллектуальные технологии: Тр. междунар. конф. «Диалог-2003», 11–16 июня, 2003 г., Протвино, Россия. М.: Наука, 2003. С. 193–198.

Extraction of Semantic Relations for Developing a Domain Thesaurus

M.S. Karyaeva

Yaroslavl State University

This research is devoted to the development of methods for the automatic creation of a thesaurus. A thesaurus is usually defined as a dictionary of concepts with a specific structure and a set of semantic relationships. Along with the basic relation synonymy used in thesauri, we investigated such kinds of semantic relations as part-whole, hypernym-hyponym («is-a») relation and associations. We choose a poetology as a domain area for carrying out primary research. By poetology we mean a group of disciplines oriented to a comprehensive theoretical and historical study of poetry. The main object of studying the area is poetic works. At the last stage, a terminological collection was created (32 000 terms), by extracting terms of different length from digitized sources. To detect the area, 1 500 terms were manually by experts.

Keywords: thesaurus, semantic relations, data mining, knowledge acquisition, domain engineering, machine learning, word2vec

Формально-языковая модель рифмующихся слов для автоматического поиска

В.Н. Бойков[1], М.С. Каряева[1], И.А. Пильщиков[2]

[1] Ярославский государственный университет,
[2] Московский государственный университет, Таллиннский университет

boykov_bh@bk.ru, mari.karyaeva@gmail.com, pilshch@yandex.ru

Аннотация

В работе выстраивается формальная классификация типов рифмы по точности и богатству созвучия на основе формально-языковой модели рифмующихся слов. Полученная классификация делает очевидной процедуру автоматического поиска в стиховом тексте рифмующихся слов и выявления классификационного кода рифмы как атрибута спецификации стихового текста по рифмике.

Ключевые слова: автоматический поиск рифмы, информационно-аналитическая система, модель фонетического слова, разметка стиха, созвучие, спецификация стихового текста, тезаурус по поэтологии, точность и богатство рифмы, формальный язык стиха

1. Введение

Поставленная задача возникла в ходе построения информационно-аналитической системы на основе тезауруса по поэтологии [1–5], аналитический блок которой предназначен для изучения стихового текста поэтического произведения с его метаописанием в терминах названного тезауруса.

Под поэтологией понимается группа дисциплин, ориентированная на всестороннее теоретическое и историческое изучение поэзии (как способа организации речи) и стиха (как сегмента поэтического текста). Тезаурус по поэтологии включает наряду с терминами для метаописания стихового текста также собственно научную терминологию поэтологии. Верхний уровень иерархии терминов тезауруса по поэтологии состоит из 10 разделов (предметных подобластей): 1. Стиховедение; 2. Стилистика; 3. Поэтика; 4. Риторика; 5. История литературы; 6. Переводоведение и литературная компаративистика; 7. Текстология; 8. Герменевтика; 9. Теоретические школы и направления; 10. Логика и методология науки.

Проблемой изучения стихового текста является его спецификация: с одной стороны, задание параметров спецификации — некоторой матрицы (формуляра) характерных атрибутов (признаков, свойств, особенностей) стихового текста; с другой стороны, — выявление специфических значений (специфицирование) этих атрибутов для конкретного стихового текста. Тезаурус по поэтологии включает наряду с терминами для метаописания стихового текста также собственно научную терминологию поэтологии. Предварительные итоги задач построения тезауруса по поэтологии и спецификации стихового текста были рассмотрены в [6].

Следует различать поэтическое произведение как эстетическую целостность и стиховой текст произведения как объект изучения в «Стиховедении». В связи с этим существенно различие спецификации стихового текста и лингвостиховедческой разметки стиха, которая отражает, в основном, статистические характеристики стихового текста. Определению таких характеристик посвящена серия работ по автоматизации анализа

русского поэтического текста В.Б. Барахнина и О.Ю. Кожемякиной с соавторами, начиная с [7].

Спецификация стихового текста на основе тезауруса может задавать и выявлять ее параметры по крайней мере на трех уровнях: 1. метрико-строфическая спецификация конкретного стихового текста; 2. поэтико-стилистическая спецификация поэтического творчества конкретного автора; 3. историко-литературная спецификация направлений, школ и периодов на основе статистической близости поэтико-стилистической спецификации различных авторов. В тезаурусе 1-му уровню спецификации соответствует раздел 1. «Стиховедение», подрубрика которого 1.1. «Стих» разбивается на 6 подрубрик 3-го уровня: 1.1.1 Метрика; 1.1.2. Явления начала и конца стиха (строки); 1.1.3. Ритмика; 1.1.4. Строфика; 1.1.5. Рифмика; 1.1.6. Лингвистика стиха.

Основные атрибуты этих подрубрик могут поддаваться автоматическому анализу, и соответствующие задачи по автоматической спецификации стихового текста, — в частности, по рифмике, — решаются в рамках информационно-аналитической системы на основе тезауруса по поэтологии.

2. Аналитическая квалификация рифмы и ее сегментов

Проблемами рифмы начал заниматься еще В.К. Тредиаковский, когда в 1730-х годах затеял реформу русского стихосложения. Всерьез рифма стала изучаться в начале XX века в работах С.И. Бернштейна, Б.В. Томашевского и В.Я. Брюсова, но наиболее глубоко — В.М. Жирмунским в работе «Рифма, ее история и теория» (1923). В 1970–1980-х годах рифме посвящен ряд исследований А.Л. Жовтиса, Ю.И. Минералова, С.И. Шепелевой и работа М.Л. Гаспарова «Эволюция русской рифмы» [8].

С развитием компьютерных методов в интернете появилось почти не обозримое множество ресурсов, посвященных автоматическому поиску рифмы, но в большинстве эти словари и генераторы рифм, предназначенные для поиска рифмы к заданному слову, мало применимы в профессиональном стиховедении. Здесь же рассматривается автоматический поиск рифмующихся слов в стиховом тексте с опорой на разработку классификации типов рифмы.

В данном разделе рассматривается аналитическая квалификация рифмы, т.е. постулируются основные функции рифмы, выделяются ее разновидности в зависимости от положения в стихе и способы рифмовки строк, задается трехчленная структура рифмующегося слова с опорной ударной гласной. Различаются также рифмы по наличию в них словораздела, по наличию гласной или согласной в окончаниях и по количеству слогов в завершении рифмующихся слов. Приводятся некоторые фонетические правила для облегчения поиска рифмы.

Под рифмой понимается созвучие (фонетическая близость) сопоставляемых слов (группы слов) независимо от их положения в стиховом тексте, начиная с ритмически сильных мест (опорных слогов с ударением) в этих словах.

Таким образом, рифму образуют не менее двух рифмующихся слов, где опорные слоги в сопоставляемых словах составляют ударные гласные (подчеркнутые) с прилегающим сочетанием согласных — заударным, предударным или с тем и другим:

Вышел месяц **из** тум**а**на,

Вынул ножик **из** карм**а**на.

Приведенное определение рифмы не является конструктивным и достаточным для построения алгоритма поиска рифмы в тексте, поэтому необходимо исследование формальных принципов генерации рифмы, опираясь на ее функции в стихе.

Важная функция рифмы в области эвфонии стихового текста в том, что она задана как звуковой повтор, выделяющий в явлениях начала и конца стиха анафору и эпифору и проявляющийся в таких явлениях как аллитерация, ассонанс и диссонанс.

В области конструкции стихового текста можно выделить две основных функции рифмы:

- в области метрики подчеркивается выделение строк и подчас полустиший;
- в области строфики определяется композиция строфы и структура произведения, в частности твердые формы.

Конструктивные особенности текста определяются положением рифмы в строке, в строфе и в тексте произведении в целом, что не определяют внутренний характер и структуру рифмы.

В строках могут быть разновидности рифм по положению: 1) начальная, 2) внутренняя, 3) концевая. Строки в тексте выделяются разновидностью 3), иногда разновидностью 1). Разновидность 2) может выделять полустишия, а может и не быть связана с таким членением строки.

В строфах по композиции рифмованных в основном по разновидности 3) строк могут различаться различные виды рифмовки: аавв — смежная (парная); авав — перекрестная; авва — кольцевая или охватная; сплетенная или смешанная рифмовка может иметь множество схем; монорифм (ааа…а) с единственной рифмой в тексте, а также твердые формы с предустановленным порядком рифм (децима, дистих, катрен, октава, рондо, секстина, септет, сонет, триолет и т.п.).

В тексте в целом выделяются кратность повторов рифмы: парная (аа), тройная (ааа), четверная (аааа), многократная (ааа…а); сквозная — одна рифма, проходящая через весь текст, при наличии других рифм.

Далее рассматривается концевая рифма, но практически все выводы будут справедливы и для других разновидностей рифмы по положению в строке.

1. В рифмующихся словах выделяются две части с ударной опорной гласной между ними: предударное сочетание слогов и заударное сочетание слогов, иначе говоря, клаузула.
2. По наличию и отсутствию словораздела в клаузуле хотя бы одного из рифмующихся слов различаются сложная рифма и простая рифма.
3. Если клаузулы рифмующихся слов оканчивается гласным, то имеет место открытая рифма, если согласным — закрытая рифма, если и гласным и согласным — смешанная рифма.
4. По равенству и неравенству числа слогов в клаузулах подразделяются на равносложную рифму и неравносложную рифму.
5. По количеству слогов в клаузулах рифмующихся слов различаются следующие виды клаузул и соответственно равносложных рифм: мужская рифма с нулевой клаузулой, женская рифма с односложной клаузулой, дактилическая рифма с двухсложной клаузулой и гипердактилическая рифма с многосложной клаузулой.

Поиск созвучий сопоставляемых слов, завершающих строку, строго говоря, следовало бы после предварительной фонетической транскрипции текста. Вместе с тем достаточно надежных и доступных ресурсов для такой транскрипции на настоящее время не имеется, а создавать специально соответствующие инструменты для данной задачи вряд ли целесообразно. В поиске рифмующихся слов для трансформирования буквенного текста в буквенный же, приближенный к фонетическому звучанию, можно использовать определенные правила, исходя из основ фонетики [9].

В русском языке из 33 букв алфавита: двумя буквами ъ и ь представлены твердый «ер» и мягкий «ерь» знаки, не обозначающие никаких звуков; 10 буквами обозначены 6 основных гласных звуков; 21 буквой обозначены 36 согласных звуков.

10 гласных букв (а, я, о, ё, э, е, у, ю, и, ы) обозначают 6 основных, произносимых под ударением, гласных звуков ([а], [о], [у], [э], [и], [ы]): [а] — а (край) и я (швея); [о] — о (рот) и ё (ёж); [э] — э (эхо) и е (мел); [у] — у (куст) и ю (тюк); [и] — и (хилый); [ы] — ы (мыло). То есть, для четырёх гласных звуков ([а], [о], [э], [у]) имеется две четверки букв: (а, о, э, у) и (я, ё, е, ю). Под буквами (я, е, ё, ю) подразумевается два звука — согласный [j]

«йот» и соответствующий гласный: я — [ja], е — [jэ], ё — [jo], ю — [jу], однако обозначаются эти йотированные звуки как ([а], [о], [э], [у]).

При рифмовке в рифмующихся словах допускается не только буквальное созвучие опорных гласных, но и попарное созвучие между четверками букв (а, о, э, у) и (я, ё, е, ю), а также между гласными (и) и (ы):

(1) 1) е~э, 2) ё~о, 3) ю~у, 4) я~а, 5) и~ы.

В безударном положении на месте гласных букв оказываются звуки более слабые, чем под ударением. Такое ослабление звучания гласных называется редукцией, так что безударные гласные являются редуцированными.

В безударном положении гласные (и, ы, у, ю) обозначают звуки, менее подверженные редукции и сохраняющие качество звучания (пирог, дыра, тумак, юла). Не различая степени слабости позиции редуцированных гласных, как в предударном положении, так и в заударном, и обозначая знаком [ъ] гласные (а, о, э) и знаком [ь] гласные (я, ё, е), созвучие редуцированных гласных можно представить набором:

(2) 1) и~ы, 2) у~ю, 3) [ъ]∈(а,о,э), 4) [ь]∈(я,ё,е).

Для согласных в русском языке отведена 21 буква (б, в, г, д, ж, з, й, к, л, м, н, п, р, с, т, ф, х, ц, ч, ш, щ), из которых, за исключением (й), выделяются 10 букв для звонких согласных (б, в, г, д, ж, з, л, м, н, р) и 10 букв для глухих согласных (к, п, с, т, ф, х, ц, ч, ш, щ). Из глухих и звонких согласных образуются 6 сближенных по созвучию пар в определенных позициях:

(3) 1) б~п, 2) г~к, 3) д~т, 4) з~с, 5) в~ф, 6) ж~ш.

Четверки согласных звонких (л, м, н, р) и глухих (х, ц, ч, щ) пар не образуют.

Всего в русском языке на 21 согласную букву насчитывается 36 звуков, а 15 согласных — 9 звонких и 6 глухих — (б, п, в, ф, г, к, д, т, з, с, л, м, н, р, х) имеют не выделяемое на письме парное как твердое, так и мягкое звучание, еще 6 согласных не имеют такой парности: 3 твердых согласных (ж, ц, ш) и 3 мягких (ч, щ, й).

Поскольку графическое буквосочетание не всегда соответствует фонетическому звучанию, в ряде сочетаний согласных между гласными в произношении как правило выпадают следующие согласные:

(4) (д) — 1) здн→зн, 2) рдц→рц, 3) дц→ц, 4) дч→ч, 5) жд→ж, например: праздник, сердце, канадцы, зодчий, к дождю;

(5) (т) — 1) стл→сл, 2) стн→сн, 3) нтск→нск, 4) тск→цк, 5) тц→ц, 6) тч→ч, например: счастливый, устный, гигантский, детский, отца, отчет;

(6) (в) — вств→ств, например: чувство;

(7) (л) — лнц→нц, например: солнце;

(8) (з) — зж→ж, например: разжевать;

(9) (с) — сж→ж, например: сжечь;

(10) (ст) — стск→ск, например: пропагандистский.

Нередко при озвучивании происходит трансформация сочетаний согласных:

(11) 1) в глагольных формах — тс, тьс→ц, например: бреется, бриться; 2) зч, жч, сч, ждь→щ, например: заказчик, мужчина, счастье, дождь; 3) гк→хк, например: легкий, мягкий.

В процессе использования автоматического поиска рифмующихся слов могут быть добавлены и другие фонетические закономерности созвучия согласных.

3. Формальная модель стихового текста и рифмующихся слов

Формальная модель стихового текста и выделенная из нее модель рифмующихся слов необходима для последующей предварительной стиховедческой разметки и затем для автоматического поиска рифмующихся слов. Ранее [10, 11] была предложена формально-языковая модель слогового представления строки стихового текста как цепочки слоговых символов:

(12) $b^{m(1)}a_1d^{n(1)}b^{m(2)}a_2d^{n(2)}\ldots b^{m(k)}a_kd^{n(k)}$,

где a означает ударные слоги фонетических слов, состоящих из одного или нескольких лексических слов с одним общим ударением, b и d означают безударные соответственно препозитивные и постпозитивные слоги между ударными слогами, k — количество ударных слогов, n и m длину соответствующих препозитивных и постпозитивных подцепочек безударных слогов.

Подцепочки $b^{m(i)}a_id^{n(i)}$, i=1,2,…k, представляет собой фонетические слова, словораздел между фонетическими словами лежит между подцепочками $d^{m(i-1)}$ и $b^{n(i)}$, i=2,3,…k.

В конечном фонетическом слове $b^{m(k)}a_kd^{n(k)}$ подцепочка $d^{n(k)}$ представляет собой клаузулу.

Для поиска в стиховом тексте рифмующихся слов с концевой рифмой формальную модель конечного фонетического слова из (12) для полноценного слогового представления следует дополнить сочетаниями (кортежами) согласных Δ^p в предударных и Δ^z в заударных сегментах. Здесь нумерация безударных гласных b и d идет от опорной гласной к началу и к концу слова соответственно:

(13) $\Delta^{p(m(k)+1)}b_{m(k)}\Delta^{p(m(k))}\ldots\Delta^{p(3)}b_2\Delta^{p(2)}b_1\Delta^{p(1)}a_k\Delta^{z(1)}d_1\Delta^{z(2)}d_2\Delta^{z(3)}\ldots d_{n(k)}\Delta^{z(n(k)+1)}$.

Если в представлении предударной части конечного фонетического слова ограничиться глубиной в два слога перед опорным гласным a_k, то ее развернутая модель имеет вид:

(14) $\Delta^{p(3)}b_2\Delta^{p(2)}b_1\Delta^{p(1)}a_k$,

где $\Delta^{p(i)}$ — кортеж из p(i) согласных, i=1,2,3.

Развернутое представление заударной части дифференцируется в зависимости от длины n клаузулы d^n. Для простоты изложения не рассматривается малоупотребительная гипердактилическая клаузула при n≥3.

При n=0 заударная часть представлена не имеющей слогов мужской клаузулой в виде кортежа из z(1)≥0 согласных:

(15) $a_k\Delta^{z(1)}$.

При n=1 заударная часть представлена односложной женской клаузулой, состоящей из примыкающих к безударной гласной d_1 двух кортежей согласных длины z(1)≥0 и z(2)≥0:

(16) $a_k\Delta^{z(1)}d_1\Delta^{z(2)}$.

При n=2 заударная часть представлена двухсложной дактилической клаузулой, состоящей из трех кортежей согласных длины z(1)≥0, z(2)≥0 и z(3)≥0, разделенных безударными гласными d_1 и d_2:

(17) $a_k\Delta^{z(1)}d_1\Delta^{z(2)}d_2\Delta^{z(3)}$.

4. Кодификация рифмы по типу созвучий

В стиховедении установилось различение связанных рифмой слов по характеру созвучия в предударных сегментах, выражающему глубину богатства рифмы, и в заударных сегментах, выражающему степень точности рифмы.

Схематически разновидности рифмы выводятся путем попарного сопоставления раздельно и поочередно для предударных и заударных сегментов конечных фонетических слов с клаузулами равной длины и с определенным созвучием опорных гласных. Созвучие кортежей согласных иерархически более приоритетно, чем созвучие безударных гласных, подверженных редукции. Такая частность, как созвучие только некоторых звуков в сопоставляемых кортежах согласных, здесь не учитывается.

Предлагаемая классификация типов рифмы как по богатству, так и по точности в определенном смысле универсальна, поскольку не суть важно, какое созвучие опорных гласных имеется у сопоставляемых слов (ассонанс — созвучие по кодам (1) или буквальное, диссонанс — отсутствие созвучия), типология и соответствующие коды для ассонансной и диссонансной рифмы аналогичны.

В примерах к приведенным ниже схемам созвучия согласных ● и ■ выделено подчеркиванием, созвучие гласных – прописной буквой.

Схемы и коды созвучий предударных сегментов

Созвучие по кодам (2) безударных гласных b_1 или b_2 с одинаковым номером в сопоставляемых предударных сегментах в формуле (14) схематически выделяется заглавной B, отсутствие созвучия обозначается через b.

Если схематически в формуле (14) кортежи согласных $\Delta^{p(i)}$, i=1,2,3, в сопоставляемых предударных сегментах обозначать светлым символом ○, тогда выявленные в результате сопоставления сегментов созвучия кортежей согласных с одинаковым номером выделяется в схеме темным символом ●.

Если дать приоритет в полученной при сопоставлении схеме сочетаниям созвучий кортежей согласных как богатым (●●●, ○●●, ●○●, ○○●), так и бедным (●●○, ○●○, ●○○, ○○○) перед сочетаниями созвучий безударных гласных как богатыми (BB, bB), так и бедными (Bb и bb), то схема богатства созвучий в порядке его убывания представляется вместе с 2-значными кодами, где цифрами (1–8) указано убывающее богатство созвучий кортежей согласных, строчными же буквами (а., б., в., г.) — убывающее богатство созвучий безударных гласных.

Схемы втройне богатых созвучий:

1.а. ●B●B●, со**р**О**к**О**в**ой / **р**О**к**О**в**ой;

1.б. ●b●B●, п**р**я**м**О**т**а / х**р**о**м**О**т**а;

1.в. ●B●b●, **т**А**р**а**к**ан / с**т**А**р**и**к**ан;

1.г. ●b●b●, **фр**а**з**и**р**овка / **фр**е**з**е**р**овка.

Схемы сдвоенно богатых созвучий:

2.а. ○B●B●, б**О**р**О**да / ков**О**р**О**да;

2.б. ○b●B●,бу**л**А**в**а / пах**л**А**в**а;

2.в. ○B●b●, м**О**л**о**т**ьб**а / г**О**л**ы**т**ьб**а;

2.г. ○b●b●, ча**щ**О**б**а / чи**щ**о**б**а.

Схемы раздвоенно богатых созвучий:

3.а. ●B○B●, со**р**О**к**О**в**ой / по**р**О**х**О**в**ой;

3.б. ●b○B●, **п**ан**О**р**а**ма / **п**ил**О**р**а**ма;

3.в. ●B○b●, **з**А**в**а**р**уха / **з**А**в**и**р**уха;

3.г. ●b○b●, **пр**иста**в**ать / **пр**ожи**в**ать.

Схемы богатых созвучий:

4.а. ○B○B●, л**Е**б**Е**д**а** / р**Е**з**Е**д**а**;

4.б. ○b○B●, симп**А**т**я**га / раб**О**т**я**га;

4.в. ○B○b●, пр**О**ще**л**ыга / м**А**ма**л**ыга;

4.г. ○b○b●, бу**м**ага / колы**м**ага.

Схемы отдаленно сдвоенных созвучий:

5.а. ●B●B○, **к**А**т**Ерок / **к**О**т**Елок;

5.б. ●b●B○, **п**о**р**Ешить / **п**е**р**Ежить;

5.в. ●B●b○, **п**О**р**улить / **п**О**р**ошить;

5.г. ●b●b○, **п**е**р**егон / **п**о**р**олон.

Схемы отдаленных созвучий:

6.а. ○B●B○, ш**А**р**О**мыга / т**О**р**О**пыга;

6.б. ○b●B○, чу**д**О**д**ей / бра**д**О**д**брей;

6.в. ○B●b○, г**У**л**л**еван / х**У**л**л**иган;

6.г. ○b●b○, ка**р**апуз / пе**р**егруз.

Схемы вдвойне отдаленных созвучий:

7.а. ●B○B○, **п**О**д**О**ж**дать / **п**О**м**О**г**ать;

7.б. ●b○B○, **пр**ин**И**мать / **пр**ож**И**гать;

7.в. ●B○b○, **п**О**л**оскать / **п**О**д**нимать;

7.г. ●b○b○, **п**оливать / **п**ереждать.

Схемы бедных созвучий:

8.а. ○B○B○, б**А**л**А**мут / п**А**р**А**шют;

8.б. ○b○B○, дом**О**сед / язык**О**вед;
8.в. ○B○b○, п**О**стижение / в**О**лочение;
8.г. ○b○b○, табуретка / сеголетка.

Схемы и коды созвучий заударных сегментов

Созвучие по кодам (2) безударных гласных d_i с одинаковым номером в сопоставляемых заударных сегментах в формулах (15), (16) и (17) схематически выделяется заглавной D, отсутствие созвучия обозначается через d. В тех же формулах кортежи согласных $\Delta^{z(i)}$, i=1,2,3, в сопоставляемых заударных сегментах схематически обозначаются светлым символом □, и тогда выявленное в результате сопоставления сегментов созвучие кортежей согласных с одинаковым номером выделяется в схеме темным символом ■.

Если дать приоритет сочетаниям созвучий кортежей согласных ■ и □ перед сочетаниями созвучий гласных D и d, то степень точности рифмы в порядке убывания раздельно для открытой и закрытой клаузулы без выделения смешанной рифмы согласно формулам (15), (16) и (17) представляется схематически вместе с 3–4-значными кодами. В этих кодах заглавной буквой (М., Ж., Д.) обозначен характер рифмы — мужская, женская и дактилическая, первой цифрой указана ее закрытость 1. или открытость 2., второй цифрой указана степень точности созвучия кортежей согласных для закрытых рифм 1.1. ■■■, 1.2. □■■, 1.3. ■□■, 1.4. □□■, 1.5. ■■□, 1.6. □■□, 1.7. ■□□, 1.8. □□□ и для открытых рифм 2.1. ■■, 2.2. □■, 2.3. ■□, 2.4. □□, строчной же буквой — степень точности созвучия безударных гласных а. DD, б. dD, в. Dd, г. dd.

Точная мужская закрытая рифма:
М.1.1. ■, клад / распад**.**
Неточная мужская закрытая рифма:
М.1.2. □, колпак / тюльпан.
Точная мужская открытая рифма:
М.2.1. стекло / ремесло.
Неточная мужская открытая рифма:
М.2.2. колесо / решето.
Точная женская закрытая рифма:
Ж.1.1.а. ■D■, орна**мЕнт** / фунда**мЕнт**;
Ж.1.1.б. ■d■, вы**рез** / вы**рос**.
Отдаленно точная женская закрытая рифма:
Ж.1.2. а. □D■, стар**Ец** / смал**Ец**;
Ж.1.2. б. □d■, пробле**ск** / отпры**ск**.
Отдаленно неточная женская закрытая рифма:
Ж.1.3.а. ■D□, бо**рО**в / по**рО**х;
Ж.1.3.б. ■d□, во**л**ен / во**л**ок.
Неточная женская закрытая рифма:
Ж.1.4. а. □D□, кит**Е**ль / див**Е**нь;
Ж.1.4. б. □d□, вылет / выгон.
Точная женская открытая рифма:
Ж.2.1.а. ■D, во**лЯ** / до**лЯ**;
Ж.2.1.б. ■d, го**р**е / го**р**ы.
Неточная женская открытая рифма:
Ж.2.2.а. □D, гор**Е** / пол**Е**;
Ж.2.2.б. □d, поле / горы.
Точная дактилическая закрытая рифма:
Д.1.1.а. ■D■D■, ка**мЕнщИк** / зна**мЕнщИк**;
Д.1.1.б. ■D■d■, жа**бЕрн**о**м** / безала**бЕрн**ы**м**;
Д.1.1.г. ■d■D■, жа**л**о**бщИк** / опа**лу**б**щИк**;
Д.1.1.д. ■d■d■, хара**ктерн**о**м** / тра**кторн**ы**м**.

Отдаленно точная дактилическая закрытая рифма:
Д.1.3.а. □D■D■, ябЕ**дн**И**к** / правЕ**дн**И**к**;
Д.1.3.б. □D■d■, берЕ**жн**ы**м** / денЕ**жн**о**м**;
Д.1.3.г. □d■D■, допу**ск**О**м** / отти**ск**О**м;**
Д.1.3.д. □d■d■, палу**бн**о**м** / надо**бн**ы**м**.
Ущербно точная дактилическая закрытая рифма:
Д.1.2.а. ■D□D■, по**л**О**в**Е**ц** / мо**л**О**д**Е**ц**;
Д.1.2.б. ■D□d■, за**р**Е**в**о**м** / жа**р**Е**н**ы**м**;
Д.1.2.г. ■d□D■, пра**в**ил**О**м / пра**в**едн**О**м;
Д.1.2.д. ■d□d■. пра**в**ило**м** / пра**в**едны**м**
Отдаленно и ущербно точная дактилическая закрытая рифма:
Д.1.4.а. □D□D■, яг**О**д**А**м / яв**О**р**О**м;
Д.1.4.б. □D□d■, прад**Е**до**м** / прав**Е**дны**м**;
Д.1.4.г. □d□D■, прадед**А**м / правил**А**м;
Д.1.4.д. □d□d■, практику**м** / прадеда**м**
Отдаленно неточная дактилическая закрытая рифма:
Д.1.5.а. ■D■D□, реа**кт**О**рн**Ы**й** / тра**кт**О**рн**Ы**м**;
Д.1.5.б. ■D■d□, реа**кт**О**рн**ый / тра**кт**О**рн**ом;
Д.1.5.г. ■d■D□, хара**кт**е**рн**Ы**й** / тра**кт**о**рн**Ы**м**
Д.1.5.д. ■d■d□, хара**кт**е**рн**ы**й** / тра**кт**о**рно**м**.**
Отдаленно и усеченно неточная дактилическая закрытая рифма:
Д.1.6.а. □D■D□, пал**У**б**н**Ы**й** / паг**У**б**н**Ы**м**;
Д.1.6.б. □D■d□, бер**Е**ж**н**ый / ден**Е**ж**н**ом;
Д.1.6.г. □d■D□, сваде**бн**Ы**й** / пагу**бн**Ы**м**;
Д.1.6.д. □d■d□, сваде**бн**ый / пагу**бн**ом
Отдаленно и ущербно неточная дактилическая закрытая рифма:
Д.1.7.а. ■D□D□, ко**л**О**т**Ы**х** / е**л**О**ч**н**Ы**м;
Д.1.7.б. ■D□d□, зо**л**О**т**ом / е**л**О**ч**ных;
Д.1.7.г. ■d□D□, уво**л**ен**Ы**м / мо**л**от**Ы**х;
Д.1.7.д. ■d□d□; зо**л**отом / позво**л**еных.
Неточная дактилическая закрытая рифма:
Д.1.8.а. □D□D□, мол**О**д**О**сть / мор**О**к**О**м;
Д.1.8.б. □D□d□, гор**О**дом/ хлоп**О**тных;
Д.1.8.г. □d□D□, уволен**Ы**м / порот**Ы**х;
Д.1.8.д. □d□d□, брошеных / голодом.
Точная дактилическая открытая рифма:
Д.2.1.а. ■D■D, ро**щИцА** / первоз**чИцА**;
Д.2.1.б. ■D■d, ро**жИц**ы / стро**жИтс**я;
Д.2.1.г. ■d■D, поме**щиц**А / мере**щатсЯ;**
Д.2.1.д. ■d■d, поме**щиц**у / мере**щатс**я.
Отдаленно точная дактилическая открытая рифма:
Д.2.2.а. □D■D, рощ**ИцА /** рож**ИцА;**
Д.2.2.б. □D■d, рощ**Иц**е **/** рож**Иц**а**;**
Д.2.2.г. □d■D, месяц**У** /помещи**цУ;**
Д.2.2.д. □d■d, месяцу /помещи**ц**а.
Отдаленно неточная дактилическая открытая рифма:
Д.2.3.а. ■D□D, ве**с**Е**л**О / ме**с**Е**в**О;
Д.2.3.б. ■D□d, ве**с**Е**л**о / ме**с**Е**в**у;
Д.2.3.г. ■d□D, ме**с**ев**У** / ке**с**ар**Ю**;
Д.2.3.д. ■d□d; кесарю / весело.
Неточная дактилическая открытая рифма:
Д.2.4.а. □D□D, яг**О**д**А** / ябл**О**к**О**;

Д.2.4.б. □D□d, рощ**И**ца / род**И**не;
Д.2.4.г. □d□D, календул**А** /умелиц**А**;
Д.2.4.д. □d□d. календулу /умелица.

Отсюда вполне ясен алгоритм получения 128 разновидностей закрытой и 64 открытой гипердактилической рифмы с трехсложной клаузулой, достаточно редкой.

5. Алгоритм поиска рифмующихся слов и кодификации рифмы

Полученная классификация типов рифмы по точности и богатству созвучий в сущности задает простой алгоритм поиска рифмующихся слов, являясь эталоном для процесса сопоставления рифмующихся слов и получения схем и кодов полученной рифмы, основные процедуры которого даются ниже**.**

Процедуры разметки стихового текста

1. Морфологическая разметка оцифрованного текста и нумерация стихотворных строк.

2. Акцентуация слоговых слов и выделение фонетических слов путем присоединения безударных слов ко всем ударяемым знаменательным словам.

3. Для конечного фонетического слова (13) в каждой строке:

3.1. маркировка гласных с выделением ударной гласной $\mathbf{a_k}$ и ее кода в созвучиях (1);

3.2. вычисление длины клаузулы n(k) по числу безударных гласных d после опорной ударной;

3.3. замена обозначений безударных гласных в соответствии с правилами (2);

3.4. вычленение кортежей согласных и их трансформация в соответствии с фонетическими правилами (3)–(11);

3.5. маркировка в заударной части между безударными гласными d кортежей согласных $\mathbf{\Delta^{z(i)}}$, i=1,2,…(n(k)+1);

3.5. маркировка в предударной части между безударными гласными b кортежей согласных $\mathbf{\Delta^{p(i)}}$, i=1,2,3.

4. Группировка и маркировка строк с клаузулами равной длины n(k) и с опорными гласными, созвучными как буквально, так и по коду созвучия (1).

Для реализации указанных процедур используется язык программирования Python и, поскольку разметка стихового текста не может гарантированно исключать неоднозначности, то для их разрешения применяются алгоритмы машинного обучения, а также привлекаются вероятностные модели.

Процедуры поиска рифмующихся слов и получения схем рифмы

5. Выбирается группа строк из процедуры 4 с первой строкой стихового текста Ст.1 и для сопоставления из нее выбирается следующая по номеру строка Ст.(1+x).

6. В клаузулах выбранных строк сравниваются, начиная от опорной гласной к концу строки, первые кортежи согласных. В схеме для рифмы строк Ст.1*Ст.(1+x) после знака = (созвучие опорных гласных) при совпадении кортежей ставится знак ■, при несовпадении — знак □.

7. Сравниваются следующие за первыми кортежами согласных безударные гласные. При созвучии в соответствии с (2) в схеме ставится знак D, при отсутствии созвучия — знак d.

8. Процедуры 6 и 7 последовательно повторяются, пока сравнивать оказывается нечего, после чего следует квалификация рифмы по схеме:

=0 знаков — открытая мужская рифма, =1 знак — закрытая мужская рифма;

=3 знака — женская закрытая рифма, =2 знака — женская открытая рифма;

=5 знаков — дактилическая закрытая рифма, =4 знака — дактилическая открытая рифма.

9. Если в одной из сопоставляемых клаузул сравнивать больше нечего, а в другой еще остался кортеж согласных, то следует вывод о наличии смешанной рифмы и схема оказывается не завершенной.

10. В предударной части выбранных строк сравниваются, начиная от опорной гласной к началу строки, первые кортежи согласных. В схеме для рифмы строк Ст.1*Ст.(1+x) перед знаком = при совпадении кортежей ставится знак ●, при несовпадении — знак ○.

11. Сравниваются следующие за первыми кортежами согласных безударные гласные. При созвучии гласных в соответствии с (2) в схеме перед знаками сравнения кортежей согласных ставится знак B, при отсутствии созвучия — знак b.

12. Затем сравниваются последовательно вторые кортежи согласных и безударные гласные, в результате после сравнения третьих кортежей согласных получается схема из 5 знаков. Однако в предударной части некоторых рифмующихся слов может и не быть слогов — например: грот / крот, явор / дьявол, обликом / облаком. Когда процедура сравнения в предударной части не имеет элементов до полного завершения схемы, то в схеме на не заполненных местах ставятся знаки ○ и b соответственно.

Процедуры кодификации рифмы

13. По итоговой схеме результатов сравнения предударных частей сопоставляемых слов отыскивается соответствующий 2-значный код схемы $X_1.X_2$. (X_1=1,2,…8, X_2=а,б,в,г,) в перечне схем и кодов раздела «Схемы и коды созвучий предударных сегментов», упорядоченном по степени убывания числа совпадений кортежей согласных.

14. По итоговой схеме результатов сравнения заударных частей сопоставляемых слов соответствующие коды отыскиваются в перечнях схем и кодов раздела «Схемы и коды созвучий заударных сегментов», выбираемых по результатам квалификации рифмы процедуры 8.

15. Комбинация кодов рифмы — 2-значного по богатству в предударной части и 3–4-значного по точности в заударной части — приписывается к номерам сопоставляемых слов (строк) как атрибут спецификации стихового текста по рифмике.

Примеры поиска рифмующихся слов и кодификации рифмы

Ст.1/ ***Осень. Бездомные псы,***
Ст.2/ ***отбросы лета в траве.***
Ст.3/ ***У неба же вдосталь красы***
Ст.4/ ***запрокинутой голове.***

Все четыре рифмующихся слова по клаузуле входят в одну группу, по опорной гласной образуются две группы строк (Ст.1, Ст.3) и (Ст.2, Ст.4).

Сверка строк 1 и 3: п **с Ы** / кр а **с Ы**; схема ○b○b●= — богатая мужская открытая рифма, код 4.г.=М.2.1.

Сверка строк 2 и 4: тр [**Ъ**] **в Е** / г о л [**Ъ**] **в Е**; схема ○b○B●= — богатая мужская открытая рифма, код 4.б.=М.2.1.

Ст.1/ ***Над окрестной биосферой***
Ст.2/ ***дивны небеса любые*** —
Ст.3/ ***у воды ненастно-серой***
Ст.4/ ***незабудки голубые.***

Все четыре рифмующихся слова по клаузуле входят в одну группу, по опорной гласной образуются две группы строк (Ст.1, Ст.3) и (Ст.2, Ст.4).

Сверка строк 1 и 3: б и [**Ъ**] сф **Е р** [**Ъ**] **й** — не н а сн [**Ъ**] с **Е р** [**Ъ**] **й**; схема ○b○B○=■D■ — бедная и точная женская закрытая рифма, код 8.б.=Ж.1.1.а.

Сверка строк 2 и 4: **л Ю б Ы** [**Ь**] — г о **л У б Ы** [**Ь**]; схема ○b●B●=■D — сдвоенно богатая и точная женская открытая рифма, код 2.б.=Ж.2.1.а.

Ст.1/ ***Вот она — идиллия,***
Ст.2/ ***вот они — блага:***
Ст.3/ ***из рога изобилия*** —
Ст.4/ ***рога, рога, рога!..***

По клаузуле и по опорной гласной образуются две группы строк (Ст.1, Ст.3) и (Ст.2, Ст.4).

Сверка строк 1 и 3: и д **И л И** [**Ь**] / изо б **И л И** [**Ь**]; схема ○b○b○=■D■D — бедная и точная дактилическая открытая рифма, код 8.г.=Д.2.1.а.

Сверка строк 2 и 4: бл [**Ъ**] **г А** – р [**Ъ**] **г А**; схема ○b○B●= — богатая и точная мужская открытая рифма, код 4.б.=М.2.1.

6. Заключение

Установлено, что процедуры выявления характеристик рифмы по богатству и по точности независимы, однако не очевидна предпочтительность той или иной характеристики.

Полученная классификация типов рифмы по точности и богатству созвучий задает алгоритм поиска рифмующихся слов среди выделенных группировок строк с клаузулами равной длины и с одинаковой опорной гласной путем сопоставления конечных слов и выявления созвучия кортежей согласных и созвучия безударных гласных, отталкиваясь от опорной гласной.

Степень точности определяется близостью созвучий к окончанию заударной части сопоставляемых фонетических слов и фиксируется в 5-местной схеме и 3–4-значном коде. Степень богатства определяется близостью созвучий предударной части сопоставляемых фонетических слов к опорной ударной гласной и фиксируется в 5-местной схеме и 2-значном коде. Комбинация кодов, как и комбинация схем, по богатству и по точности может быть приписана к номерам сопоставляемых строк как атрибут спецификации стихового текста по рифмике.

Процедуры автоматического поиска рифмующихся слов в конечном счете позволяют составить коллекцию рифм (парных, тройных, четверных и многократных) со схемой и кодом спецификации и с идентификацией их принадлежности конкретному тексту конкретного автора.

Работа поддержана Российским фондом фундаментальных исследований, гранты № 16-06-00497 и № 16-07-01180.

Литература

[1] Бойков В.Н. и др. Тезаурус как инструмент поэтологии. / Бойков В.Н., Захаров В.Е., Пильщиков И.А., Сысоев Т.М. // Моделирование и анализ информационных систем. 2010. Т. 17. № 1. С. 5–24. URL: mais.uniyar.ac.ru/sites/default/files/journal/private/17_1__5-24.pdf (дата обращения: 01.03.2018).

[2] Бойков В.Н., Пильщиков И.А. Семантическая модель «Тезауруса по поэтологии» в составе информационно-аналитической системы // Интернет и современное общество: Труды XVI Всероссийской объединенной конференции «Интернет и современное общество» (IMS-2013), (Санкт-Петербург, 9–11 октября 2013 г.). СПб: Университет ИТМО, 2013. С. 273–278. URL: http://docplayer.ru/46693016-Semanticheskaya-model-tezaurusa-po-poetologii-v-sostave-informacionno-analiticheskoy-sistemy.html (дата обращения: 01.03.2018).

[3] Бойков В.Н. и др. Тезаурус по поэтологии как инструмент для информационного поиска и коллекции знаний. / Бойков В.Н., Захаров В.Е., Каряева М.С., Соколов В.А. // Моделирование и анализ информационных систем. 2013. Т. 20, № 4. С. 125–135.

[4] Бойков В.Н. и др. Предметно-ориентированный тезаурус в открытой информационно-аналитической системе. / Бойков В.Н., Захаров В.Е., Каряева М.С., Соколов В.А. // Электронные библиотеки: перспективные методы и технологии, электронные

коллекции (RCDL): Труды 15 Всероссийской научной конференции RCDL'2013 (Ярославль, 2013 г). Ярославль: ЯрГУ, 2013. С. 70–76. URL: http://rcdl2014.jinr.ru/doc/RCDL2014_Proceedings.pdf (дата обращения: 01.03.2018)

[5] Каряева М.С. Лингвостатистический анализ терминологии для построения тезауруса предметной области. // Моделирование и анализ информационных систем. 2015. Т. 22, №. 6. С. 834–851.

[6] Бойков В.Н., Каряева М.С. Поэтология: задачи построения тезауруса и спецификации стихового текста. // Моделирование и анализ информационных систем. 2017. Т. 24, № 6. С. 811–815.

[7] Барахнин В.Б., Кожемякина О.Ю. Об автоматизации комплексного анализа русского поэтического текста // CEUR Workshop Proceedings. 2012. Т. 934. С.167–171.

[8] Гаспаров М.Л. Эволюция русской рифмы // Проблемы теории стиха. Л.: Наука, 1984. С. 3 – 36. Переизд.: Гаспаров М.Л. Избранные труды. М., 1997. Т. 3. С. 290–325.

[9] Болычева Е.М. Основы фонетики // Филология: Русский язык: Фонетика. 2004. URL: http://www.portal-slovo.ru/philology/37379.php (дата обращения: 01.03.2018) .

[10] Бойков В.Н. Контекстно-свободная грамматика одной ритмической модели русского стиха. // Моделирование и анализ информационных систем. 2012. Т. 19, № 4.

[11] Бойков В.Н. и др. Об автоматической спецификации стиха в информационно-аналитической системе. / Бойков В.Н, Каряева М.С., Соколов В.А., Пильщиков И.А. // Аналитика и управление данными в областях с интенсивным использованием данных: Труды международной научной конференции DAMDID/RCDL-2015 (Обнинск, 15–16 октября 2015 г.). С. 144–151.

A Formal Language Model of the Rhyming Words for Automated Search

V.N. Boikov [1], M.S. Kariaeva [1], I.A. Pilshchikov [2]

[1] Yaroslavl Demidov State University,
[2] Lomonosov Moscow State University, Tallinn University

The paper offers a formal classification of rhymes based on the accuracy and richness of rhymes defined according to a formal language model of rhyming words.
The classification suggests a procedure of automated search of rhyming words in a poetic text and formulation of the classifying code of the rhyme as a formal attribute in the specification of a poetic text.

Keywords: automated rhyme search, information-analytical system, phonetic word model, verse mark up, specification of verse text, thesaurus of poetics, accuracy and richness of rhyme, formal language of verse description

Оценка эффективности гибридного морфологического анализатора NLTK4RUSSIAN в работе с текстами социальных сетей и художественных произведений

А.О. Кириллова, А.Г. Мельник, А.Д. Плетнева,
Е.В. Еникеева, О.А. Митрофанова

Санкт-Петербургский государственный университет

alyonkakirillova@gmail.com, i-ve-got-a-mail@inbox.ru,
adpletneva@perm.ru, protoev@yandex.ru, oa-mitrofanova@yandex.ru

Аннотация

В статье описан эксперимент по тестированию гибридного морфоанализатора NLTK4RUSSIAN на материалах соревнований «Dialogue Evaluation» 2017 г. Эксперименты с обучением морфоанализатора выполнены на основе подкорпусов НКРЯ и OpenCorpora. Для тестирования использовались выборки из художественных текстов и текстов социальных сетей. В ходе исследования решен ряд задач, в том числе задача конвертации морфологической разметки из формата Universal Dependencies в формат PyMorphy2, используемый в OpenCorpora. Результаты тестирования гибридного морфоанализатора NLTK4RUSSIAN соответствуют «золотому стандарту» для русского языка.

Ключевые слова: морфологический анализ, разрешение морфологической неоднозначности, NLTK4RUSSIAN, корпусы русскоязычных текстов

1. Введение

Морфологическая аннотация и разрешение морфологической неоднозначности при автоматической разметке текстов — задачи, которые имеют множество решений, отличающихся трудозатратами и стандартами качества. Среди систем, обеспечивающих достаточно точный морфологический анализ текста на русском языке, стоит выделить систему АОТ (http://www.aot.ru), анализатор mystem (https://tech.yandex.ru/mystem/) и парсер TreeTagger (http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger). Существует распространенное мнение, что неоднозначность на одном языковом уровне должна сниматься за счет следующего уровня (в частности, неоднозначность в морфологии должна решаться с использованием данных о структуре синтаксических групп и синтаксических связей между словами в предложении) [1]. Такой метод может быть эффективным, но при этом он является дорогостоящим и требующим больших усилий.

В настоящее время развиваются специализированные модули для снятия омонимии, представляющие три основные группы:

- системы, основанные на правилах (созданных вручную);
- системы, основанные на вероятностных моделях и обученные на размеченных корпусах;
- гибридные системы, использующие как подход на правилах, так и вероятностные модели.

Системы, основанные на правилах, как раз похожи по своей реализации на предлагаемый подход к снятию омонимии с помощью следующего языкового уровня. Такие системы работают линейно. Системы второго типа, т.е. основанные на вероятностных моделях, работают медленнее, но их проще реализовать. Немаловажную

роль здесь играет размер обучающего корпуса. Некоторые системы должны обучаться на корпусах небольших размеров, так как большие корпуса могут повредить работе алгоритма.

Ежегодно в России проводится международная научная конференция по компьютерной лингвистике «Диалог» (http://www.dialog-21.ru/). Цель конференции — обмен идеями между специалистами в области лингвистики, а также представление накопленных знаний и результатов разработанных систем для автоматической обработки текста.

В рамках конференции проводится соревнование «Dialogue Evaluation» (http://www.dialog-21.ru/evaluation/), которое включает в себя тестирование по различным темам, таким как анализ тональности, анализ семантического сходства, машинный перевод и др. В 2017 г. одной из таких тем был морфологический анализ (http://www.dialog-21.ru/evaluation/2017/morphology/). Соревнование морфологических анализаторов проводилось в двух форматах: на закрытой дорожке оценивались алгоритмы, обученные на ограниченном количестве заранее предоставленных данных, а на открытой дорожке участники могли обучать анализаторы на любом материале. На обеих дорожках оценивалось качество частеречной разметки, полной разметки (т.е. определения части речи и значений основных лексико-грамматических категорий) и лемматизации. Для обучения алгоритмов на закрытой дорожке участникам были предоставлены размеченные корпусы со снятой вручную омонимией, а именно подкорпус Национального корпуса русского языка (НКРЯ, http://www.ruscorpora.ru/), подкорпус Генерального интернет-корпуса русского языка (ГИКРЯ, http://www.webcorpora.ru/), данные проекта OpenCorpora (http://opencorpora.org/) и материалы корпуса СинТагРус (http://www.ruscorpora.ru/instruction-syntax.html) [2].

Целью данного исследования является тестирование гибридного морфоанализатора NLTK4RUSSIAN на разных типах текстов. Ранее была проведена оценка качества работы морфоанализатора с привлечением тестовых выборок из НКРЯ и OpenCorpora [3]. На данном этапе представляется необходимым обновить эти результаты и протестировать морфоанализатор на корпусных данных второго соревнования по морфологическому анализу в рамках «Dialogue Evaluation». Эти тестовые материалы включают художественные и новостные тексты, а также текстовые материалы из социальной сети ВКонтакте.

Экспериментальный дизайн данного исследования включает в себя следующие этапы:

- исследование основных режимов работы морфоанализатора NLTK4RUSSIAN;
- создание конвертера, преобразующего теги, используемые для морфологической разметки в Universal Dependencies, в теги формата PyMorphy2;
- обучение морфоанализатора и его тестирование на выбранных данных;
- обработка результатов и оценка качества общей морфологической и отдельно частеречной разметок с помощью метрики «точность средства измерений» (accuracy);
- сравнение полученных данных для использованных корпусов.

2. Морфоаналогический анализатор NLTK4RUSSIAN

2.1. Принцип работы морфоанализатора

Анализатор NLTK4RUSSIAN — лингвистический комплекс для исследования русскоязычных корпусов текстов, разрабатывающийся на кафедре математической лингвистики СПбГУ (https://github.com/named-entity/nltk4russian) [3, 4, 5]. Главная задача — улучшение морфологического анализа с использованием уже существующих инструментов. NLTK4RUSSIAN называется гибридным морфоанализатором, так как объединяет в себе алгоритмы NLTK (http://www.nltk.org/) [6] и PyMorphy2 (http://pymorphy2.readthedocs.org/en/latest/) [7].

NLTK — Natural Language Toolkit — библиотека Python, которая предназначена для работы с текстом на всех уровнях языка. Средства библиотеки позволяют проводить как базовые операции (лемматизация, токенизация, стемминг), так и многоуровневый анализ с применением алгоритмов машинного обучения. NLTK также позволяет проводить морфологический анализ и разметку. В числе инструментов теггеры, использующие словарные данные, теггер на основе скрытых марковских моделей, анализаторы с применением контекстной информации и др. [6]. Важная особенность NLTK заключается в том, что в большинстве случаев используется размеченный корпус, что делает лингвистический комплекс независимым от языка текста для анализа. Одновременно с этим среда NLTK не адаптирована для конкретных языков, она использует универсальную метаинформацию. NLTK4RUSSIAN, таким образом, является адаптацией NLTK к русскоязычным текстам, создаваемой путем интеграции лингвонезависимых алгоритмов NLTK и морфологического анализатора для русского языка.

PyMorphy2 — анализатор русских текстов, который использует морфологический словарь OpenCorpora. Для этой системы не нужен обучающий корпус, все необходимые словоформы описаны в словаре. Для редких или несуществующих форм работает предсказатель, который выводит предполагаемый тег по префиксу и по концу слова. Для совпадающих словоформ или форм с неоднозначностью выводятся все возможные интерпретации, вероятность того или иного варианта рассчитывается не на основе контекста, а по данным словаря. Можно сказать, что NLTK4RUSSIAN предлагает альтернативное решение неоднозначности: если PyMorphy2 выводит несколько вариантов разбора, то теггер пользуется возможностями анализа соседних словоформ от NLTK. Таким образом, NLTK4RUSSIAN объединяет достоинства NLTK и PyMorphy2: текст анализируется с помощью анализатора PyMorphy2, для дизамбигуации вызывается теггер NLTK. При этом система может работать как в гибридном режиме, так и с использованием NLTK и PyMorphy2 по отдельности.

2.2. Режимы работы морфоанализатора

NLTK4RUSSIAN работает в 3 режимах, для переключения каждого из них при вызове морфоанализатора необходимо указать специальный параметр-название "-n":

- гибридный (PyMorphy2 + NLTK);
- PyMorphy2 (PyMorphyTagger);
- униграммный, биграммный и триграммный теггеры NLTK.

Гибридный режим — это режим с использованием PyMorphy2 и биграммного теггера NLTK nltk.tag.sequentials. При вызове анализатора указывается файл с текстом, размеченный корпус для обучения и название файла, который будет создан в ходе работы. Файл с неразмеченным текстом читается построчно. Вызывается функция MorphAnalyzer() от PyMorphy2, при обнаружении неоднозначности запускается биграммный теггер NgramTagger с соответствующим параметром. Режим получил название "PyMorphyContext", при запуске анализатора нужно указать параметр "-n" "pmcontext".

В режиме PyMorphy2 вызывается функция MorphAnalyzer() от PyMorphy2, и, как в гибридном режиме, выполняется разметка на основе словарных данных. Но и в отличие от PyMorphyContext, неоднозначность снимается с помощью оценки вероятности. Параметр рассчитывается на основе данных корпуса OpenCorpora: для всех слов со снятой неоднозначностью вычисляется, сколько раз словоформа была отмечена данным тегом. На основе этих частот и выводится условная вероятность тега (параметр "score"). (http://pymorphy2.readthedocs.io/en/latest/user/guide.html)

Последний режим анализирует текст на основании контекста — предшествующих словоформе тегов слов. Униграммный теггер (UnigramTagger) ищет подходящий вариант разбора в обучающем корпусе, биграммный (BigramTagger) анализирует тег предыдущего слова в тексте и ищет соответствие в обучающем корпусе, триграммный (TrigramTagger)

составляет контекст из разбора слова и двух предыдущих слов и ищет подходящие разборы в обучающем корпусе.

Для вызова морфоанализтора в одном из этих режимов при запуске нужно указать соответственно "-n 1gram", "2gram" или "3gram".

3. Тестирование морфоанализатора NLTK4RUSSIAN

3.1. Исходные данные

В качестве обучающих данных были взяты материалы НКРЯ и проекта OpenCorpora, в качестве тестовых — тексты, которые также являлись тестовыми для участников соревнований «Dialogue Evaluation» 2017 г., а именно корпусы текстов новостного сайта Lenta, социальной сети ВКонтакте и текст романа О. Зайончковского «Петрович» (далее — JZ). Стоит отметить, что данный материал представляет собой тексты с различными жанровыми и стилистическими особенностями.

Обучение морфоанализатора NLTK4Russian выполнено на размеченных текстах со снятой омонимией. Обучающие материалы включают выборки из НКРЯ, а также подкорпус OpenCorpora объёмом 30 тыс. словоупотреблений.

3.2. Предподготовка обучающих и тестовых выборок

На предварительном этапе тестирования была сформулирована задача конвертации разметки в формате Universal Dependencies в формат PyMorphy2, используемый в OpenCorpora. За основу для будущего конвертера был взят имеющийся скрипт для конвертации из pymorphy в Universal Dependencies. Отметим, что при конвертации тегов в другой формат часто возникают несоответствия. В нашем случае большая часть таких несоответствий связана с тем, что в разметке OpenCorpora предлагается выделять 12 падежей (шесть стандартных, звательный, два дополнительных родительных, дополнительный винительный и два дополнительных предложных). Стандарт Universal Dependencies содержит только шесть тегов для падежей, поэтому при конвертации часть тегов в формате PyMorphy2 не восстанавливается.

```
'Case': {
    'Ins':  'ablt',
    'Acc':  'accs',
    'Dat':  'datv',
    'Gen':  'gen1',
    'Loc':  'loct',
    'Nom':  'nomn',
},
```

Рис. 1. Фрагмент кода для конвертации падежных тегов в формат pymorphy

Для файлов с тестовой выборкой потребовалась дополнительная обработка, связанная с извлечением слов для разметки. Эта задача была реализована при помощи специального скрипта. На вход поступает файл с тестовой выборкой (Lenta, VK или JZ), в котором из каждой строки извлекаются имеющиеся словоформы. На выходе мы получаем отдельный файл с извлечёнными формами, с которым далее работает морфоанализатор.

Обучающая выборка (материалы НКРЯ и OpenCorpora) уже содержит разметку, однако эта разметка представлена в формате Universal Dependencies. Поскольку на выходе мы должны получить размеченные тестовые материалы с тегами pymorphy, то и обучающая выборка должна содержать теги нужного нам формата. Поэтому обучающая выборка также потребовала предварительной обработки. С этой целью был разработан

дополнительный скрипт, который обрабатывает файлы с обучающей выборкой, последовательно заменяя в них теги Universal Dependencies на теги PyMorphy2.

Таким образом, на предварительном этапе мы подготовили обучающие и тестовые корпусы к тестированию морфоанализатора.

3.3. Тестирование морфоанализатора и результаты

Обучение и разметка тестовых выборок осуществлялась в соответствии со стандартной процедурой, описанной для компонента train-tagger гибридного морфоанализатора. Морфоанализатор может размечать тестовые корпусы двумя способами: приписывая только частеречные теги или полную грамматическую характеристику данной словоформы. Для полной разметки указывается параметр -full . Соответственно, оценка результатов разметки выполняется по этим двум основаниям при помощи отдельного скрипта eval в составе морфоанализатора.

Таблица 1. Конфигурации для тестирования

№ эксперимента	Обучающий корпус	Тестовый корпус
1	НКРЯ (выборка)	JZ
2	НКРЯ (выборка)	Вконтакте
3	НКРЯ (выборка)	Lenta.ru
4	OpenCorpora	JZ
5	OpenCorpora	Вконтакте
6	OpenCorpora	Lenta.ru

Таблица 2. Точность разметки

№ эксперимента	1gram		2gram		3gram		pymorphy		pmcontext	
	Full	POS	Full	POS	Full	POS	Full	POS	Full	POS
1	88,1%	90,9%	**76,8%**	91,7%	**76,1%**	91,5%	**75,5%**	91,3%	**74,8%**	90,5%
2	89%	91,3%	75,5%	91,1%	72,6%	90,9%	70,2%	90,5%	69,3%	89,3%
3	**90,2%**	**93,8%**	68,3%	**91,9%**	66,5%	**91,6%**	65,3%	**91,5%**	64,8%	91%
4	85,4%	90,6%	76%	91,4%	75,4%	91,3%	**75,5%**	91,3%	**74,8%**	90,5%
5	86,1%	90,2%	71,8%	90,2%	70,6%	90,3%	70,2%	90,5%	69,4%	89,4%
6	86,6%	91,7%	66,6%	91,4%	66%	91,3%	65,3%	**91,5%**	65%	**91,1%**

Выводы

Таким образом, при оценке результатов было выяснено, что полученные данные практически не зависят от выбранного обучающего корпуса (например, результат частеречной разметки корпуса текстов автора JZ с НКРЯ и с корпусом OpenCorpora в качестве обучающего в режиме pmcontext оказался одинаковым — 74,8).

Точность частеречной разметки в среднем оказалась выше точности полной морфологической разметки; для частеречной разметки средний показатель равен 91%, а для полной морфологической разметки среднее значение заметно меньше — 74,2%.

Были протестированы все доступные режимы работы морфоанализатора NLTK4RUSSIAN, которые показали в основном примерно одинаковую точность разметки. Было выяснено, что результаты полной морфологической разметки варьируются в зависимости от жанра текстов: наибольшей точности удалось добиться на текстах автора JZ (76,8%), следом идёт корпус текстов социальной сети ВКонтакте (75,5%); наименьшая точность наблюдается для текстов новостного корпуса (68,3%). В этом ряду выделяются

результаты, полученные при работе в режиме униграммного теггера: точность разметки здесь оказалось вообще самой высокой, причем наибольшая точность наблюдается при разметке новостного корпуса (полная морфологическая разметка — 90,2%, частеречная — 93,8%).

Проведенные эксперименты основаны на исследовании возможностей морфологического анализатора, опубликованном в 2015 г. [3]. В ранней серии экспериментов для обучения использовались подкорпус OpenCorpora со снятой омонимией и две выборки из НКРЯ со снятой омонимией: подкорпуса художественных текстов и публицистики. Для тестирования были взяты выборки из НКРЯ. Показатели точности в эксперименте по частеречной разметке располагались в промежутке от 90 до 95,8%, по полной морфологической — от 79,7 до 86%. Самые высокие результаты наблюдались при разметке публицистических (частеречная разметка) и художественных (полная морфологическая разметка) текстов. В новом исследовании удалось улучшить качество разметки: наибольшая точность частеречной и полной морфологической разметок была получена при разметке новостного корпуса (93,8 и 90,2% соответственно).

Результаты нынешних экспериментов также сравнивались с показателями, полученными в ходе соревнований морфологических парсеров «Dialogue Evaluation» 2017 г. Базовые алгоритмы разметки показали результаты в промежутке от 68,44 до 76, 54% при полной морфологической разметке и от 72,1 до 79,49% при частеречной разметке. Лучший результат на данных соревнованиях был получен морфоанализатором ABBYY (точность частеречной разметки — 97,11%). Показатели остальных участников соревнований, полученные в этом сегменте соревнований (частеречная разметка), варьировались от 71,48 до 93,08%. Данные наблюдения позволяют подтвердить вывод о том, что результаты работы парсера NLTK4RUSSIAN являются состоятельными и удовлетворяют высоким требованиям соревнований морфологических анализаторов.

Исследование поддержано грантом РФФИ № 16-06-00529 «Разработка лингвистического комплекса для автоматического семантического анализа русскоязычных корпусов текстов с применением статистических методов» (2016–2018 гг.).

Литература

[1] Сокирко А.В, Толдова С.Ю. Сравнение эффективности двух методик снятия лексической и морфологической неоднозначности для русского языка // Интернет-математика 2005. М., 2005. URL: http://www.aot.ru/docs/RusCorporaHMM.htm (дата обращения 04.05.2018).

[2] Sorokin A., Shavrina T., Lyashevskaya O., Bocharov V., Alexeeva S., Droganova K., Fenogenova A., Granovsky D. MorphoRuEval-2017: an Evaluation Track for the Automatic Morphological Analysis Methods for Russian // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог–2017». М., 2017. Т. 1, № 16. С. 297–313.

[3] Паничева П.В., Протопопова Е.В., Митрофанова О.А., Мирзагитова А.Р. Разработка лингвистического комплекса для морфологического анализа русскоязычных корпусов текстов на основе Pymorphy и NLTK // Труды международной конференции «Корпусная лингвистика – 2015». СПб, 2015. С. 361–373.

[4] Москвина А.Д., Орлова Д., Паничева П.В., Митрофанова О.А. Разработка ядра синтаксического анализатора для русского языка на основе библиотек NLTK // Компьютерная лингвистика и вычислительные онтологии. Труды XIX Международной объединенной научной конференции «Интернет и современное общество», Санкт-Петербург, 22–24 июня 2016 г. СПб, 2016. С. 44–54.

[5] Паничева П.В., Митрофанова О.А. Интеграция морфоанализаторов для аннотации русскоязычных корпусов текстов // Сборник материалов по итогам XLIII Международной филологической конференции. Секция прикладной и математической лингвистики. СПб, 2014. С. 56–60.
[6] Bird S., Klein E., Loper E. Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. Beijing, 2009.
[7] Korobov M. Morphological Analyzer and Generator for Russian and Ukrainian Languages // Analysis of Images, Social Networks and Texts: 4th International Conference, AIST 2015, Yekaterinburg, Russia, April 9–11, 2015. Communications in Computer and Information Science, Springer, 2015. P. 320–332.

Testing and Assessment of the NLTK4RUSSIAN Hybrid Morphological Analyzer

A.K. Kirillova, A.G. Melnik, A.D. Pletneva, E.V. Enikeeva, O.A.Mitrofanova

Saint-Petersburg State University

This paper describes a series of experiments aimed at testing the hybrid morphological analyzer NLTK4RUSSIAN on the datasets of the Dialogue Evaluation 2017. We provide a detailed explanation of the principles and main operation modes of the system. Experiments on training the system were based on RNC and OpenCorpora subcorpora. Certain problems dealing with the data format were solved in course of research: we performed conversion of Universal Dependencies markup format to PyMorphy2 format. The analyzer was tested on social network subcorpus, news subcorpus and fiction subcorpus. Test results for the NLTK4RUSSIAN morphological analyzer correspond to the gold standard of morphological annotation for the Russian language.

Keywords: morphological tagging, morphological disambiguation, NLTK4RUSSIAN, text corpora

К проблеме создания списка высокочастотных слов и выражений немецкого языка для специальных целей

М.С. Коган[1], А.М. Ярошевич[1], А.Ю. Колотаева[1], В.П. Захаров[2], З. Шрот-Вихерт[3], А. Тильманс[3]

[1] Санкт-Петербургский политехнический университет Петра Великого
[2] Санкт-Петербургский государственный университет
[3] Ганноверский университет имени Лейбница

m_kogan@inbox.ru, amjarr@mail.ru, anna.kolotaeva@mail.ru, v.zakharov@spbu.ru, schroth-wiechert@fsz.uni-hannover.de, anna.tilmans@fsz.uni-hannover.de

Аннотация

Статья посвящена статистическому анализу специальных немецких текстов с целью выявления устойчивых сочетаний. В отличие от английского языка, на материале которого подобные исследования проводятся очень широко, списков высокочастотных слов и выражений в немецком для специальных / научных целей не существует. Исследование было проведено на материале подкорпуса «Электротехника» немецкой части корпуса текстов PhD диссертаций по техническим специальностям, являющихся основой *Kod.ING* корпус, который разрабатывается в рамках совместного проекта Ганноверского университета имени Лейбница и Санкт-петербургского политехнического университета Петра Великого.

В работе представлены списки самых частотных существительных, найденных с помощью функции WordList, и высокочастотных коллокаций, найденных с помощью функции N-Grams корпусного менеджера третьего поколения AntConc. Обсуждаются лексические и грамматические особенности полученных выражений, их сходство и отличия от подобных англоязычных списков и возможности использования полученных результатов в дидактических целях в курсе немецкого языка для специальных целей. В частности, обращается внимание на преобладание предложных сочетаний в полученном списке. Также указывается на ограничения программы AntConc, проявляющиеся при анализе относительно больших корпусов немецких текстов.

Ключевые слова: корпусная лингвистика, устойчивые сочетания, лексические пучки, списки частотных слов и выражений, английский научный дискурс, Kod.ING корпус, немецкий для специальных целей

1. Введение

Работы, посвященные количественной оценке лингвистических данных, велись давно. Еще в 1897-1898 гг. немецким лингвистом Ф. Кедингом (F. Keding) был составлен первый корпус текстов в бумажном виде объемом 11 млн. слов для сравнения частоты распределения букв в словах и выявления их сочетаемости. Особенно активно количественные методы стали развиваться с появлением компьютеров. На их основе стали создаваться частотные словари и проводиться различные исследования как теоретического, так и прикладного характера [1–5].

Следующий шаг в использовании количественных методов в лингвистике был сделан с появлением корпусов текстов. Результаты обработки запросов к корпусу всегда сопровождаются выдачей соответствующих статистических данных. В настоящее время репрезентативные корпусы служат источником для создания словарей того или иного языка.

Один из популярных предметов в корпусной лингвистике — это устойчивые сочетания. Среди них можно выделить 3 основных типа: грамматические сочетания (например, составные предлоги), семантизированные устойчивые сочетания, как свободные, так и идиоматические, и просто n-граммы.

Исследователи считают, что одним из важнейших результатов исследований в области корпусной лингвистики за последние десятилетия является вывод о том, что язык состоит не из отдельных слов, а сочетаний, регулярно встречающихся вместе и использующихся в устном и письменном дискурсе [6]. Носители языка как в устной, так и в письменной речи, тяготеют к использованию одних и тех же линейно организованных сочетаний лексических единиц. Такого рода сочетания не являются фразеологизмами, а представляют собой часто повторяющиеся цепочки слов, не всегда обладающие конструктивной законченностью, но хранящиеся в языковой памяти говорящего как некие «строительные блоки». Например, «per cent of the», «for the first time», «at the end of the», «it has been (shown / observed / argued) that». Наиболее разработана тема таких словосочетаний англоязычными исследователями. Тем не менее, в англоязычной литературе не существует общепринятого термина для определения таких словосочетаний. В рамках нашего исследования мы будем использовать предложенный Д. Байбером (D. Biber) с соавторами термин «lexical bundles», который авторы определяют как «последовательность слов, часто встречающуюся в определенном стиле речи» [7]. В отечественных публикациях линейно организованным сочетаниям лексических единиц, не образующих семантического целого, почти не уделяется внимания и, следовательно, отсутствует общепринятое название для данного явления. Поэтому вслед за Н.В. Денисовой и Е.С. Петровой мы будем использовать термин-кальку «лексические пучки» [8].

2. Списки частотной лексики в английском научном дискурсе

Наиболее широко описано использование корпусов в учебных целях в англоязычной литературе. В английском языке эти исследования в 21 веке проводятся с очень большим размахом и широко используются на практике, например, при создании учебников по английскому языку для специальных целей. Речь идет о созданном в 2000г. частотном списке слов общенаучной лексики — AWL (academic wordlist), который включает в себя 570 гнезд слов (word families) [9] на базе корпуса, составленного из статей из научных журналов и учебников по 28 предметным областям, относящимся к четырем крупным областям знания (естествознание, юриспруденция, бизнес и гуманитарные науки) и содержащего 3,5 млн. слов. Список составлялся на основе трех критериев: *specialised occurrence, range and frequency* [9, p.221] (на данный момент работу цитировали 2670 раз). В 2014 г. Д. Гарднер и М. Дэвис (D. Gardner и M. Davies) опубликовали статью, в которой показали необходимость создания нового списка частотных слов общенаучной лексики английского языка на базе подкорпуса научной периодики (academic subcorpus), содержащего 120 млн. слов — части Корпуса современного американского английского (COCA) [10]. Целью исследователей было выделить зону общенаучной «ядерной» лексики, отсекая «слева» высокочастотные слова, встречающиеся во всех типах речи, и «справа» — узкопредметную лексику. Для этого они использовали 4 критерия (ratio, range, dispersion, discipline measure). В результате был получен новый список частотных общенаучных слов, содержащий 3000 слов, полностью доступный на сайте COCA[1]. В данной работе приведены

[1] Corpus of Contemporary American English <https://www.academicwords.info/>

500 наиболее частотных слов из этого списка [10, p. 317–320]. В 2013 г. Л. Валипори и Х. Нассаи (L. Valipouri и H. Nassaji) опубликовали список частотных слов, встречающихся в научном дискурсе по химии (Chemistry Academic WordList (CAWL)), основанный на анализе корпуса, состоящего из 1185 научных статей по химии [11]. Они обнаружили, что 27,85% слов из их списка (CAWL) отсутствовали в широко известном списке слов общенаучной лексики — AWL, составленном А. Кохед (A. Coxhead) [9].

Масштабный проект по созданию и обработке корпусов на 9 языках (английский, польский, итальянский, шведский, норвежский, русский, китайский, греческий, арабский), описан в [12]. Результатом проекта стали списки частотных слов для 9 языков и 72 языковых пар для изучающих иностранные языки и переводчиков. База данных этого проекта, получившего название KELLY, находится в открытом доступе[2], и авторы призывают преподавателей указанных языков и лингвистов активно использовать этот ресурс в своей работе [12, p.155].

Что касается языка для специальных целей, то по мнению М. Маккарти (M. McCarthy) с соавторами составление обычного частотного списка слов уже может предоставить достаточно информации для выделения характерных особенностей дискурса корпуса [13]. Типичные повторяющиеся сочетания слов («лексические пучки»), выделенные из специального корпуса, дают дополнительное представление о языке определенной области (языке для специальных целей).

В англоязычной литературе, особенно в Великобритании, в последнее время широкое распространение получило понятие formulaic language — лингвистический термин для вербальных выражений, которые устойчивы по форме, но не являются единицами плана содержания и которые тесно связаны с коммуникативно-прагматическим контекстом. Проблемы, связанные с изучением formulaic language, особенно активно исследуются в работах, посвященных проблемам изучения иностранного языка. Исследователи обращают внимание на то, что «лексические пучки» занимают, с одной стороны, очень важное место в научном дискурсе, а с другой — в разных дисциплинах они ведут себя по-разному и используются с разной частотностью [7, 14, 15]. Несмотря на наличие работ по исследованию роли лексических пучков в изучении таких языков, как корейский и испанский, большинство исследований проведено на материале английского языка [16]. В статье, опубликованной в 2008 г., К. Хайлэнд (K. Hyland) приводит списки 50 наиболее частотных выражений состоящих из 4-х слов в 4-х предметных областях: биологии, электротехнике, прикладной лингвистике и деловому администрированию, полученных на основе анализа специальных корпусов, собранных из научных статей, магистерских и докторских (PhD) диссертаций общим объемом 3,5млн. слов. При этом он отмечает, что в корпусе по электротехнике таких четырехкомпонентных лексических пучков значительно больше, чем в других корпусах. Больше половины выражений из каждого предметного списка не встречаются в других корпусах, собранных для исследования, и только 5 самых частотных лексических пучков (*on the other hand, as well as, at the same time, the results of the, in the case of*) встречаются во всех корпусах, а 14 лексических пучков — в 3-х разных областях [15, p.12–13].

В 2010 г. Р. Симпсон-Влах и Н. Эллис (R. Simpson-Vlach и N. Ellis) опубликовали список научных «шаблонов» (Academic Formulas List), включив в него выражения, которые встречаются в научных текстах гораздо чаще, чем в других видах дискурса, и типичны для разных подвидов научного дискурса [17]. В 2015 г. Дж. Фокс и М. Тигчелаар (J. Fox и M. Tigchelaar) предложили список из 99 лексических пучков, которые, по их мнению, целесообразно использовать при обучении будущих инженеров письменному дискурсу. Этот список был составлен на основе анализа корпуса научных статей по инженерным специальностям и разделен на 3 функциональные категории: референтные выражения (Referential expressions), выражения, организующие дискурс, и выражения, вводящие новое

[2] KELLY lists<https://www.npmjs.com/package/kelly-lists>, KELLY DB http://kelly.sketchengine.co.uk>

утверждение (stance expressions) [18]. Широкий обзор исследований, в основном, зарубежных авторов и на материале английского языка, по теме *formulaic language* дан в монографии Д. Вуда (D. Wood) [19].

3. Доступность немецких корпусов для специальных целей

Насколько нам известно, специальных корпусов немецкого языка, которые могли бы стать ресурсом для студентов инженерных специальностей при освоении письменной профессиональной коммуникации на немецком языке, не существует. Основанием для такого вывода является сравнительный анализ работ, посвященных немецкому и английскому языку для специальных целей за 1998–2012 гг., проведенный С. Яворска [20]. Автор ссылается на международный проект *Gesprochene Wissenschaftssprache Kontrastiv (GeWiss)* «Разговорный научный язык в сравнении», содержащий 1,2 млн. слов, свободно доступный онлайн для проведения сравнительных исследований устного научного дискурса на 3-х языках: немецком, английском и польском (https://gewiss.uni-leipzig.de). GeWiss-корпус дополняет такие известные корпусы, как Мичиганский корпус разговорного научного английского (MICASE) и корпус Британского научного разговорного английского (BASE) [20, p. 187]. Что касается обучения письменному научному дискурсу, то С. Яворска ссылается на исследования Г. Грэфен (G. Graefen), которая в обучении студентов письменному немецкому для специальных целей подчеркивала важность сопоставления того, как используется общенаучная лексика в научном дискурсе и обыденной речи, ее метафорический характер в научном дискурсе [21]. Также обращается внимание на важность списков частотных слов и словосочетаний, Примеры подобного рода содержатся в пособии [22], предназначенном для студентов и аспирантов инженерных направлений, которые пишут свои диссертации на немецком языке, не являющимся для них родным. Яворска признает, что использование подходов корпусной лингвистики в изучении общенаучного словаря немецкого языка находятся в начальной стадии [20].

Зная это, доцент Ганноверского университета им Лейбница З. Шрот-Вихерт (S. Schroth Wiechert) решила создать специальный корпус немецкого языка как ресурс для изучающих общенаучный и технический немецкий. Проведенный ею анализ показал, что подобных ресурсов очень мало, они не являются легкодоступными, а находящиеся в них примеры не соответствуют направлению подготовки магистрантов и аспирантов по таким специальностям как, например, *турбостроение, механика жидкостей и газов* или *гражданское строительство.* Ее инициатива получила развитие в рамках программы «Стратегическое партнерство» между Санкт-Петербургским Политехническим университетом Петра Великого (СПбПУ) и Ганноверским университетом им. Лейбница (ГУЛ). С 2014 г. Лингвистический центр ГУЛ и кафедры гуманитарного института СПбПУ работают над созданием трилингвальной платформы, ***D**eutsch, **E**nglish and **R**usskii* (DEaR)-корпус, в названии которого отражены языки (немецкий, английский и русский), на которых написаны диссертации и научные статьи, входящие в корпус [23]. Предполагается, что DEaR -корпус будет доступным онлайн аннотированным корпусом с встроенной поисковой системой (корпусным менеджером). Однако, кроме технических проблем, необходимо решить правовые, которые на данный момент запрещают пользователям за пределами ГУЛ обращаться к материалам немецкой части корпуса — наиболее разработанной и названной the Korpus der Ingenieurwissenschaften *(Kod.ING).*

С учетом этого обстоятельства было решено проанализировать подкорпус немецких диссертаций по электротехнике с использованием собственного корпусного менеджера.

4. Сервис корпусного менеджера

Корпус текстов становится мощным инструментом в руках лингвиста лишь посредством специализированных программных средств. Неотъемлемой частью понятия «корпус

текстов» является система управления текстовыми и лингвистическими данными, которую чаще всего называют корпус-менеджером (или корпусным менеджером). Это специализированная система, включающая программные средства для поиска данных в корпусе, получения статистической информации и предоставления результатов пользователю в удобной форме. Также сюда с некоторой долей условности можно отнести сюда средства подготовки и загрузки текстов в корпус.

Современные корпусные системы позволяют не только формировать конкордансы для заданных слов и частотные списки, но и решать достаточно сложные задачи, такие как выявление коллокаций (устойчивых сочетаний), ключевых слов и словосочетаний, построение лексико-семантических групп и др. Функциональность корпусного менеджера определяется типом данных и конкретной задачей, для которой корпус создается. Это же правомерно и для корпусов, обрабатывающих специальные тексты.

Большинство современных инструментов, используемых корпусной лингвистикой, предлагают множество функций, включая статистические методы, обладают некоторой масштабируемостью для работы с большими корпусами, предлагают многоязычную поддержку и включают в себя дружественный интерфейс. Наибольшим их ограничением является то, что они плохо работают с большими корпусами. Более мощные системы 4-го поколения, такие как corpus.byu.edu, CQPweb, Sketch Engine, Wmatrix предлагают лучшую масштабируемость за счет хранения корпуса в базе данных веб-сервера и предварительной индексации данных для обеспечения быстрого поиска.

Однако несмотря на перечисленные выше преимущества новых корпусных систем, они также имеют ряд ограничений. Они требуют покупки программного обеспечения, наличие собственного сервера и поддержание его работоспособности, установки программного обеспечения на сервер, для чего необходима высокая программистская квалификация. Еще одна проблема заключается в том, что в инструментах 4-го поколения размыты границы между данными и инструментом. Из-за способа хранения данных в индексированной форме на внешнем сервере, пользователи не имеют возможности обратиться к исходным данным непосредственно, по крайней быстро соотнести их с результатами поиска в корпусе.

Анализ показывает, что набор свободно доступных корпусных программ совсем не велик. Наиболее известны такие свободно распространяемые универсальные корпусные менеджеры как XAIRA (BNC), Manatee/Bonito, CQP, DDC, Wordsmith, MonoConc и AntConc.

Мы остановили свой выбор на одном из наиболее эффективных корпусных менеджеров, AntConc, программе, разработанной профессором Университета Васэда (Япония) Л. Антони (L. Anthony) и доступной на его сайте[3]. Выбор данного компьютерного инструмента обусловлен тем, что это свободно распространяемое мультиплатформное программное обеспечение, оснащенное удобным интерфейсом, имеющее множество функций по автоматической обработке текстов в разной кодировке и в разных форматах. AntConc позволяет обрабатывать тексты в кодировках Unicode, iso, cp, koi8, ascii для текстов на европейских языках и в специальных кодировках для текстов на восточных языках. Форматы входных файлов, поддерживаемые в AntConc — это .txt, .html, .htm, .xml. В состав AntConc входят статистический модуль и модуль машинного обучения. AntConc допускает обработку как «сырых», так и размеченных текстов. Возможно подключение списков стоп-слов, списков лемм, списков отрицательных ключевых слов и пр.

С помощью AntConc можно производить следующие операции:

- просмотр файла с текстом;
- построение конкорданса для целевого слова в пределах контекстного окна;
- построение графиков к конкордансу;

[3] Laurence Antony’s AntConc, a freeware corpus analysis toolkit for concordancing and text analysis <http://www.laurenceanthony.net/software/antconc/>.

- выделение ключевых слов по двум критериям log-ikelihood и χ2 в анализируемом корпусе с выдачей результатов о ранге и частоте;
- выделение коллокатов целевого слова на основе двух коэффициентов ассоциации (MI и log-likelihood) в пределах контекстного окна;
- построение частотного списка словоформ и/или лемм для обрабатываемого корпуса с указанием ранга и абсолютной частоты;
- построение частотного списка кластеров для заданных слов;
- выделение N-грамм с целевым словом в пределах контекстного окна и построение частотного списка выделенных N-грамм.

Именно последние две функции были использованы нами для получения результатов для написания данной статьи.

5. Материал и инструмент исследования

Выбранный для анализа корпуса немецких диссертаций инструмент, AntConc, относится к корпусным менеджерам 3-го поколения [24]. Их главным недостатком является то, что они не могут справиться с очень большими корпусами, содержащими больше 100 млн. слов [24, p. 152]. Однако, как мы указали в предыдущем разделе, выбор корпусного менеджера 4-го поколения в данный момент не представлялся нам возможным,

В системе AntConc имеется два инструмента, позволяющих выявить наиболее частые N-граммы: «Кластеры» (The Clusters Tool или просто Cluster) и N-граммы (The N-Grams Tool). Первый из них фактически суммирует результаты, полученные в инструменте Concordance Tool или Concordance Plot Tool, подсчитывая частоты левых и правых кластеров (окружений, контекстов) вокруг заданного слова. При этом имеется возможность задать длину кластера и другие параметры настройки. Второй инструмент позволяет сканировать весь корпус и вычислять частотность для кластеров длины N. Все параметры настройки, доступные в инструменте «Кластеры», также доступны в инструменте N-грамм.

Очень привлекательным выглядит поиск в режиме *N-Gram* с указанием минимального и максимального размера лексического пучка, минимальной частоты единицы в корпусе и количества текстов, в которых она встречается (range). К сожалению, для поиска по всем текстам Kod.ING-корпус оказался слишком велик. Для проведения поиска с использованием инструмента «Clusters/N-Grams» нами был использован стационарный компьютер с 64-битной операционной системой Windows 8 с частотой 4-ядерного процессора 3,40 МГц и объемом оперативной памяти 16 Гб. Неоптимальное использование программой AntConc оперативной памяти в процессе обработки массива данных приводит к переполнению последней и последующему аварийному прекращению работы программы. Решить проблему можно, если обрабатывать только небольшие группы файлов, или искать компьютер с большим объёмом оперативной памяти. Для обработки всех файлов корпуса нужно очень большой объем оперативной памяти. Мы установили, что для обработки каждого подкорпуса в отдельности используется 95–99% оперативной памяти указанного компьютера.

6. Получение лексических пучков для высокочастотных слов Kod.ING-корпуса

6.1.Получение лексических пучков с помощью функции *Cluster*

Получение лексических пучков с помощью функции *Cluster* предполагает генерирование частотного списка слов всего корпуса текстов диссертаций, выделение из него самых частотных существительных и дальнейший поиск частотных лексических пучков на их основе или на основе сложного слова, составной частью которого является

слово из списка. Мы проанализировали списки из 100 самых частотных слов в целом Kod.ING корпусе и в каждом подкорпусе.

Проанализировав разные подкорпусы, можно сделать вывод, что после вспомогательных частей речи самыми частотными значимыми словами являются существительные. Общеупотребительных существительных, которые входят в 100 самых употребляемых слов, больше чем общенаучных. Самыми частотными словами оказались такие существительные, которые можно будет встретить в диссертационных текстах по разным специальностям, например: *Daten* (данные), *Abbildung* (изображение), *Kapitel* (глава), *Vergleich* (сравнение), *Untesuchung* (исследование). К самым частотным общенаучным существительным стоит отнести: *Simulation* (симуляция), *Berechnung* (измерение), *Verfahren* (процесс), *Temperatur* (температура). Также присутствуют слова, которые представляют пласт общеупотребительной лексики: *Einfluss* (влияние), *Anzahl* (количество), *Bereich* (область).

В своих работах Г. Грэфен предлагает использовать для составления списков для изучения в курсе немецкого языка распространенные существительные, глаголы и прилагательные в сочетании с их наиболее частотными коллокациями [21]. Мы проверили, являются ли выделенные ею слова *Analyse* (анализ) and *Aspekt* (аспект) высокочастотными в нашем корпусе. В режиме поиска *Cluster/N-Grams* поисковый запрос формируем следующим образом: *Analyse, указав размер кластера Min.2 — Max.5 и минимальную частотность 10. Такой запрос позволяет найти сложные слова с корнем «Analyse» и коллокации с ними. Результаты представлены в таблице 1.

Таблица 1. Наиболее частотные лексические пучки Kod.ING-корпуса, отобранные с использованием сложных слов с корнем «Analyse»

№	Частотность (в корпусе)	Лексический пучок
1	46	Praxisanalyse / Gegenüberstellung
2	42	Praxisanalyse / Gegenüberstellung Zwischenergebnisse
3	20	Feinanalyse und
4	19	Feinanalyse und technische
5	19	Feinanalyse und technische Realisierung
6	11	Sensitivitätsanalyse der
7	10	Bildanalyse-Objektmodellen
8	10	Eigenwertanalyse zyklischer
9	10	Qualitätsanalyse von
10	10	Qualitätsanalyse von gdv
11	10	Qualitätsanalyse von gdv bei

Как и следовало ожидать, выделить сочетания существительных с глаголами автоматически не удалось. Причина состоит в том, что из-за синтаксических особенностей немецкого языка расстояние между существительным и относящимся к нему глаголом часто бывает больше 5–6 слов.

6.2. Получение лексических пучков с помощью функции N-Grams

Так как поиск по всему Kod.ING-корпусус помощью этой функции оказался невозможным, то мы ограничились поиском по подкорпусу из 35 PhD диссертаций на немецком языке в области электротехники. Поиск проводился по следующим параметрам: длина лексического пучка 2–4 слова, минимальная частота встречаемости — 10 раз во всех текстах; распределение (range) — 5, что означает, что искомые пучки должны встречаться не менее чем в 5 разных текстах подкорпуса. Главным критерием отбора лексических пучков являлась их частотность в корпусе. Оборотная сторона применения этой простой метрики состоит в том, что для нахождения «нужных выражений» приходится вручную

обрабатывать длинные списки лексических пучков, найденных программой как удовлетворяющих поисковому запросу.

Поясним, какие лексические пучки мы искали. М. Маккарти (McCarthy) с соавторами указывают, что найденные в корпусе лексические пучки могут состоять:

1) из случайного набора слов, часто встречающихся вместе (*so dass die, die durch die*);

2) из синтаксически неполных, но «осмысленных» выражений: (*zwischen den beiden, auf Basisder, in bezug auf*) и

3) из семантически и прагматически «законченных» выражений: (*im Rahmen dieser Arbeit, ist in Abbildung dargestellt)* (примеры взяты из Kod.ING-корпуса*).* (В своей монографии авторы приводят примеры лексических пучков каждой категории на материале английского языка: as *are to my, this one for* (1), *to be able to, a lot of the* (2) и *on the other hand* and *as a result* (3) [13, p. 61]). Нас интересовали выражения, относящиеся ко второй и к третьей группе. Всего при указанных параметрах нами было идентифицировано 50 выражений. Они представлены в таблице 2.

7. Обсуждение и выводы

Подавляющее большинство выражений представляют собой предложные конструкции: *предлог + существительное*. В немецком более распространены предложные конструкции, которые по структуре представляют собой связку глагол + предлог, которая называется «управление глаголов» и является грамматической доминантой в сочетаемости предлогов с другими частями речи.

Обычно конструкции с предлогами являются составными частями более крупных конструкций. В традиционной лингвистике сочетаемость предлогов, как правило, описывается с точки зрения их грамматических (падежных) функций, однако семантике предлогов и их функциональной роли до настоящего момента должного внимания не уделялось. И совсем не исследованным остается вопрос о семантике и сочетаемости предлогов в специальных текстах. Корпусных исследований, посвященных предлогам, крайне мало. Исчерпывающие формальные описания функционирования предлогов в составе конструкций отсутствуют.

Необходимость корпусно-статистического описания немецких предлогов ставит перед лингвистами сложную задачу. В связи с тем, что глагол и предлог в немецком предложении могут находиться не рядом, что обусловлено рамочной конструкцией, при автоматическом поиске таких сочетаний находится лишь незначительная их часть. Среди трудностей следует назвать также синонимию предлогов и вариативность конструкций. Для исследований необходимо использовать несколько корпусных источников, т.к. типы конструкций и их частотные характеристики для одного и того же предлога в тематически разнородных текстах могут не совпадать.

В предложных конструкциях с существительными предлог может находиться в предпозиции (***in*** *der Regel, bei der Modellierung*) или постпозиции к существительному (*abhängig von, Beispiel für*). Как известно, в немецком языке предлоги управляют следующим после себя словом, т.е. после предлогов zu/von последующее слово будет находиться в дательном падеже, а после предлогов auf/für- в винительном. Маркером рода и падежа является артикль, поэтому он и является такой частотной частью речи в немецких предложениях и широко представлен в корпусе.

В большинстве частотных лексических пучков присутствует артикль, который выполняет несколько функций:

- определять род главного существительного предложного сочетания, например: ***Der*** *Zugriff auf, in* ***der*** *Literatur;*
- являться частью следующего слова, которое не входит в данный лексический пучок, например: *Einfluss auf* ***die*** *,Auf Basis* ***der****, Parameter für* ***die.*** Хайленд также включает в свои списки высокочастотных выражений в разных предметных областях английского

языка выражения, заканчивающиеся определенным артиклем: *due to **the**, in terms of **the**, as a result of **the*** [15, p.7].

– Отдельно можно выделить такие предложные конструкции, как предлог+сущ.+предлог*: i**m** Bereich von; i**n** Form von.* В целом, преобладающими в постпозиции являются сочетания существительных с предлогами*von* и *zu*: сущ. + ***von,*** сущ + ***zu***, а в препозиции – с предлогами *bei* и *in (im*): ***bei***+ сущ. ***in/im*** + сущ.

Таблица 2. Наиболее частотные лексические пучки Kod.ING- корпуса, отобранные с использованием функции *N-Grams*

№	Частотность (в корпусе)	Трехкомпонентные лексические пучки	Двух и четырех компонентные лексические пучки	Частотность (в корпусе)
1	417	in dieser Arbeit	im Rahmen dieser Arbeit	219
2	336	in diesem Fall	zum Zeitpunkt	71
3	219	in der Regel	Beispiel für	69
4	225	mit Hilfe der	Abhängig von	58
5	221	die Anzahl der	Unter Berücksichtigung	58
6	179	auf diese Weise	Vor allem	57
7	179	aus diesem Grund	Mit einem Durchmesser von	21
8	168	mit Hilfe von	Differenz zwischen	10
9	164	im Vergleich zu	Verfahren für	10
10	143	in Bezug auf	für die Herstellung von	10
11	127	In der Literatur		
12	120	in Abhängigkeit von		
13	137	In Abbildung dargesellt		
14	130	Einfluss auf die		
15	113	Stand der Technik		
16	104	im Gegensatz zu		
17	103	auf Basis der		
18	75	in Richtung der		
19	74	in Form von		
20	73	die Verwendung von		
21	72	unter der Annahme		
22	70	es ergibt sich		
23	66	für die Berechnung der		
24	64	der Zugriff auf		
25	62	als Funktion der		
26	58	im Bereich von		
27	56	im folgenden werden		
28	53	im folgenden wird		
29	46	bei der Entwicklung		
30	43	im folgenden Abschnitt		
31	38	auf der Oberfläche		
32	30	bei der Simulation		
33	20	bei einer Frequenz		
34	19	bei der Analyse		
35	18	Parameter für die		
36	18	in der Datenbank		
37	18	bei der Untersuchung		
38	17	bei der Implementierung		
39	17	bei der Modellierung		
40	13	in der Größenordnung		

Еще одним признаком научного стиля является большое количество пассивных конструкций с вспомогательным глаголом *werden,* например*:im Folgenden **werden***. Но

выявить конструкцию целиком автоматически не получается из-за большого расстояния между вспомогательным и смысловым глаголом.

Проведенный анализ позволяет сделать следующие предварительные выводы.

1) В немецком языке, как и в любом флективном языке с изменяющимися формами существительных, правильное употребление требуемой формы слова является непростой задачей для студентов инженерных специальностей. Имея информацию о частоте употребления определенных форм в письменном научном дискурсе, можно целенаправленно уделять больше внимания именно им при обучении студентов письменной практике для научных целей.

2) В состав большинства предложных сочетаний с существительными входит артикль. С употреблением артиклей связано большое количество ошибок. Поэтому возможность проанализировать их употребление в частотных лексических пучках является очень ценной. Наблюдая, в каких случаях на практике используется определенный артикль, а в каких неопределенный, как в "*der* direkte Vergleich" и "*ein* direkter Vergleich" можно с помощью расширенного контекста эмпирически понять, в чем разница в их употреблении.

3) Автоматический поиск по текстам корпуса с использованием программы *AntConc* не позволяет учесть некоторые особенности немецкого синтаксиса при формировании запросов. Например, личные формы глагола в немецком языке могут состоять из нескольких частей (один или два вспомогательных глагола и смысловой глагол), которые не всегда занимают в предложении место рядом с существительным, к которому непосредственно относятся, будь оно подлежащим или дополнением, что не позволяет легко находить в корпусе сочетания существительных с глаголами. В связи с этим, необходима ручная обработка полученного конкорданса. Аналогичной обработки требуют и глаголы с отделяемыми приставками. Техническим решением может стать использование корпусного менеджера с более гибким языком запросов и учет особенностей немецкого синтаксиса и нужных функций для анализа корпуса при разработке корпусного менеджера *HanConc*, встроенного в разрабатываемый *Kod.ING*-корпус.

Полученные результаты позволяют надеяться, что при дальнейшем исследовании Kod.ING-корпуса могут быть получены новые интересные данные по частотности и сочетаемости слов разных предметных областях, которые могут быть использованы в дидактических целях.

Исследование поддержано программой сотрудничества «Стратегическое партнерство», проекты DAAD (Deutscher Akademischer Austauschdienst) № 56268450 (2013-2016) и № 57271274 (2017-2018).

Литература

[1] Алексеев П.М. Статистическая лексикография. Л.: Изд-во ЛГПИ, 1975.
[2] Арапов М.В. Квантитативная лингвистика. М.: Наука, 1988.
[3] Головин Б.Н. Язык и статистика М.: Просвещение, 1971.
[4] Пиотровский Р.Г. Информационные измерения языка. Л.: Наука. Ленинградское отделение, 1968.
[5] Фрумкина Р.М. Статистические методы изучения лексики. М.: Наука, 1964.
[6] Römer U. The inseparability of lexis and grammar: Corpus linguistic perspectives // Annual Review of Cognitive Linguistics. 2009.Vol. 7. P. 141–163.
[7] Biber D., Conrad S., Cortes V. If you look at ...: Lexical bundles in university teaching and textbooks // Appl. Linguist. 2004. Vol. 25, № 3. P. 371–405.
[8] Денисова Н.В., Петрова Е.С. Маркеры ряда: кластерность и вариантность (на материале английского языка) // Вестник Челябинского государственного университета. 2008. Т. 30. С 44–49.
[9] Coxhead A. A new academic word list // TESOL Quarterly. 2000. Vol. 34, № 2. P.213–238.

[10]Gardner D., Davies M. A New Academic Vocabulary List // Applied Linguistics. 2014. Vol. 35, № 3. P. 305–327.
[11]Valipouri L., Nassaji H. A corpus-based study of academic vocabulary in chemistry research articles // J. of English for Academic Purposes. 2013. Vol. 12, № 4. P. 248–263.
[12]Kilgarriff A, Charalabopoulou F., Gavrilidou M., Johannessen J. B., Khalil S., Johannessen K., Sofie J., et al. Corpus-based vocabulary lists for language learners for nine languages // Language Resources and Evaluation. 2014. Vol. 48, № 1. P. 121–163.
[13]McCarthy M., O'Keeffe A., Cartner R. From Corpus to Classroom. UK, Cambridge: CUP, 2007. 315 p.
[14]Cortes V. Teaching lexical bundles in the disciplines: An example from a writing intensive history class // Linguistics and Education. 2016. Vol.17. P. 391–406.
[15]Hyland K. As can be seen: Lexical bundles and disciplinary variation // ESP.2008. Vol. 27, № 1. P. 4–21.
[16]Hyland K. Bundles in Academic Discourse // Annual Review of Applied Linguistics. 2012. № 32. P. 150–169.
[17]Simpson-Vlach R., Ellis N.C. An Academic Formulas List: New Methods in Phraseology Research // Applied Linguistics. 2010. Vol. 31, № 4. P. 487–512.
[18]Fox J., Tigchelaar M. Creating an engineering academic formulas list // The Journal of Teaching English for Specific and Academic Purposes. 2015. Vol. 3, № 2. P. 295 –304.
[19]Wood D. Fundamentals of Formulaic Language: An introduction. London: Bloomsbury. 2015. 205 p.
[20]Jaworska S. Review of recent research (1998–2012) in German for Academic Purposes (GAP) in comparison with English for Academic Purposes (EAP): cross-influences, synergies and implications for further research // Lang. Teach. 2015. Vol. 48, № 2. P. 163–197.
[21]Graefen G. Die Didaktik des wissenschaftlichen Schreibens: Möglichkeiten der Umsetzung // GFL-journal. 2009. № 2-3. S. 106–127.
[22]Schroth-Wiechert S. Deutsch als Fremdsprache in den Ingenieurwissenschaften: Formulierungshilfen für schriftliche Arbeiten in Studium und Beruf. Berlin: Cornelsen Verlag. 2011. 160 p.
[23]Gärtner T., Schroth-Wiechert S., Kogan M. A trilingual platform for academic technical writing // Коммуникация в поликодовом пространстве: лингво-культорологические, дидактические, ценностные аспекты: Материалы междунар. Научн. конф. СПб: Изд-во Политехн. ун-та, 2015. С. 67–68.
[24]Anthony L. A critical look at software tools in corpus linguistics // Linguistic Research. 2013. Vol. 30, № 2. P. 141-161.

On Problem of Creating a German Engineering Academic Words and Formulas List

M.S. Kogan [1], A.M. Yaroshevich [1], A.Y. Kolotaeva [1,] V.P. Zakharov [2,]
S. Schroth Wiechert [3], A. Tilmans [3]

[1] Peter the Great Saint Petersburg Polytechnic University,
[2] Saint Petersburg State University,
[3] Leibniz Universität Hannover

The article deals with statistical analysis of German texts in order to find reoccurring lexical bundles. Unlike the English language there are no academic formulas and words frequency lists in German academic discourse. The authors analyzed the *Elektrotechnik* subcorpus of the Kod.ING corpus, which contains PhD dissertations in German in different engineering fields. The Kod.ING corpus is being developed at Leibniz University of Hannover (LUH) as a part of a joint project with St. Petersburg Polytechnic University.

The article presents lists of the most frequent nouns and lexical bundles found with *WordList* and *N-gram* functions correspondingly of the AntConc freeware corpus analysis toolkit for concordancing and text analysis. The paper discusses lexical and grammatical peculiarities of the selected lexical bundles, highlighting similarities with English academic formulas and words frequency lists. In particular, the dominance of prepositional phrases in the list has been noticed. Limitations of AntConc program for analysis of relatively large German corpora are described and perspectives of applying the findings in teaching German for specific purposes are considered.

Keywords: corpus linguistics, lexical bundles, academic formulas and word list, English academic discourse, Kod.ING corpus, German for specific purposes

Вариативность представления имен политических деятелей при диахронических исследованиях на основе корпусов текстов

А.Ц. Масевич[1], В.П. Захаров[2]

[1] Санкт-Петербургский государственный институт культуры
[2] Санкт-Петербургский государственный университет

andmasev@mail.ru, v.zakharov@spbu.ru

Настоящая публикация продолжает ряд публикаций, посвященных диахроническим исследованиям частотного поведения политической лексики в текстах книг. В определенной степени она носит методический характер. В ней рассматривается, каким образом форма представления собственного имени и статуса политического деятеля отражаются на частотном поведении этой лексики и как она должна быть учтена при осуществлении исследований.

Ключевые слова: корпусная лингвистика, имена политических деятелей, частотность, вариативность, система Google books Ngram Viewer

1. Введение

Настоящая публикация продолжает ряд публикаций, посвященных диахроническим исследованиям частотного поведения политической лексики в текстах книг. В предыдущих публикациях [1–3] мы показали на нескольких примерах, как изменение частоты встречаемости лексических единиц в текстах книг отражает реальные историко-культурные процессы, и описали несколько моделей частотного поведения единиц политической лексики [4, 5]. Настоящая публикация носит в определенной степени методический характер. В ней рассматривается, каким образом форма представления собственного имени и статуса политического деятеля отражаются на частотном поведении этой лексики и как она должна быть учтена при осуществлении исследований.

В любой поисковой системе стоит задача точной формулировки информационной потребности, т. е. перевода содержательного пользовательского запроса, сформулированного на естественном языке, в поисковое предписание на запросном языке данной системы. В случае тематических запросов это далеко не тривиальная задача. В современных поисковых системах бестезаурусного типа содержание запросов, так же, как и содержание документов, выражается с помощью слов естественного языка. Однако понятия, лежащие в основе тематического запроса, в языке могут выражаться с помощью различных слов и словосочетаний. Поскольку тема запроса и ее аспекты суть имена понятий, и мы не знаем, каким способом это понятие будет выражено в искомых документах, то необходимо в запросе «развернуть» все гнездо близких по смыслу слов и словосочетаний, описывающих это понятие (синонимичные выражения, видовые термины и т. п.). Когда объектом поиска являются имена собственные, то задача формулирования запроса, казалось бы, не представляет особой трудности. На самом деле это не совсем так. Проблемой представления имен давно занимаются специалисты по информационным, в особенности библиотечным системам. В электронных каталогах крупных библиотек существуют системы авторитетного контроля, учитывающие вариативность элементов библиографического описания, в том числе личных имен [6–8].

При поиске по именам собственным следует учитывать омонимию имен и фамилий, а также разные способы именования персон (Сталин, товарищ Сталин, И.В. Сталин, Иосиф Виссарионович Сталин и т. п.), причем эти способы могут зависеть от страны, от соответствующего исторического периода или вообще быть сугубо индивидуальны. Эти вариации и исследуются в данной статье.

Наше исследование проводилось на основе трех корпусов системы Google Books Ngram Viewer ― русского, британского английского и американского английского.

Система Google Books Ngram Viewer [9] детально описана нами в нескольких предыдущих публикациях [5, 10]. Она позволяет строить графики встречаемости слов и коллокаций за выбранный временной период. На горизонтальной оси графика показываются годы, входящие в заданный временной период. По вертикальной оси откладывается выраженная в процентах относительная частота встречаемости в корпусе заданной N-граммы (от одного слова до пяти) в соответствующем году. Относительная частота встречаемости N-граммы за определенный год подсчитывается следующим образом: число употреблений N-граммы в данном году делится на общее число словоупотреблений в корпусе в этом же году, результат умножается на 100 [10, с. 307].

При построении графика имеется возможность задать ряд условий, позволяющих делать графики более наглядными, сопоставлять кривые поведения нескольких N-грамм, выявлять наиболее частотные словосочетания с данной словоформой, их поведение во времени и др.

2. Методологические замечания

Значение относительной частоты встречаемости некоторой лексической единицы за данный год зависит от числа текстов, изданных в этом году, которые введены в базу данных Google Books. Здесь перед нами встают методологические трудности. Во-первых, мы не знаем, какая часть опубликованных в данном году текстов введена в базу данных, нам известно только общее число текстов ― примерно 590 тыс. текстов для русского корпуса. Во-вторых, мы не знаем, какая часть опубликованных текстов должна быть введена в базу данных, чтобы исследование было статистически достоверным.

Тем не менее, есть уверенность, что при аккуратно выстроенной методике исследования результатам, полученным на данных системы Google Books Ngram Viewer, можно доверять. В подтверждение сказанного приведем несколько соображений.

1. Система предполагает определенный «уровень достоверности» данных за счет того, что графики могут быть построены лишь для тех N-грамм, которые встречаются в корпусе не менее 40 раз [10, с. 307].

2. В системе имеется механизм сглаживания [10, с. 307–308], позволяющий нивелировать годовые колебания в наполнении корпуса.

3. Результаты, полученные на данном корпусе другими исследователями, такими, в частности, как Ю.С. Масленникова, В.В. Бочкарев, В.Д. Соловьев, Т.И. Галеев [11-13] хорошо коррелируют с данными лингвистической науки.

4. Данное исследование, является частью большой работы, начатой авторами три года назад. В ходе этой работы построено около 500 графиков динамики частотного поведения различных слов и коллокаций, относящихся к политической лексике. Этот материал доступен в Интернет [14]. И полученные результаты показывают очевидные корреляции с реальными историческими событиями.

5. Мы провели сравнение частотного поведения лексических единиц в Google Books Ngram Viewer и в Национальном корпусе русского языка (НКРЯ), сервис «Графики» (http://ruscorpora.ru/ngram.html). Сравнение графиков, построенных разными системами для десяти частотных существительных русского языка (*год, человек, время, дело, жизнь, день, рука, раз, работа, слово*) [15] показало, что соотношение частотности разных единиц и модели их частотного поведения в большинстве случаев сходны или отличаются незначительно. Затем такое же сравнение было проведено и на некоторых терминах

политической лексики. Имеются случаи и существенных различий, но они, как правило, легко объяснимы: это случаи редких лексических единиц, когда объем НКРЯ (76 тыс. текстов) оказывается явно мал.

Таким образом, мы утверждаем, что частотное поведение лексики по данным системе Google Books Ngram Viewer является достоверным индикатором культурно-исторических явлений и процессов, хотя и нуждается в дополнительных методологических исследованиях.

3. Вариативность имен политических деятелей в разрезе хронологии и географии

Рассмотрим варианты именования монархов Российской империи в пред- и послереволюционные годы (1820–1920). При этом, разумеется, была учтена дореволюционная орфография [10].

На рисунке 1 видно, что более частотными способами обозначения особы монарха были выражения «Государь», «Император», «Его Величество».

Рис. 1. Сопоставление изменения частоты слов «Государь» и «Императоръ» и имен пяти императоров за период с 1820 г. по 1920 г.

На рис. 1 видно, что при сопоставлении кривых для слов «Государь» и «Императоръ» и кривых имен императоров последние практически сливаются с горизонтальной осью, т.е. чаще указание на императора шло без имени.

Система Google Books Ngram Viewer позволяет выявлять для каждого выражения 10 наиболее частотных словосочетаний и строить графики их встречаемости за определенный период. При этом существует возможность задать часть речи в правой или левой позиции (рис. 2).

Таким образом, были выявлены наборы глаголов наиболее употребительных с данными выражениями. Во всех трех случаях это, в основном, глаголы в прошедшем времени, такие как «изволил», «отправился», «возвратился», «благоволил», «занимался», «изъявил», «пожаловал», «послал», «согласился», «утвердил», «приказал», реже употребляются формы настоящего времени «изволит», используется также инфинитив «повелеть [соизволил]).

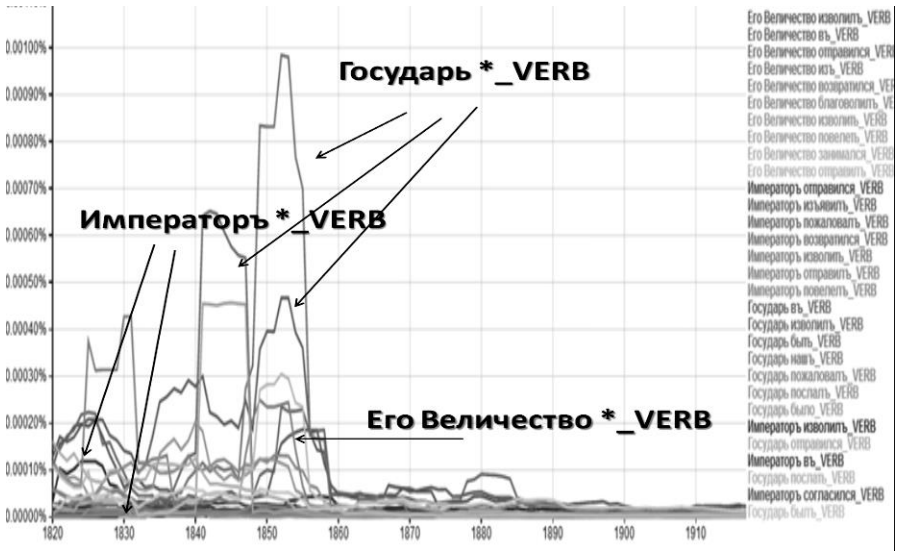

Рис. 2. Динамика частотного поведения выражений «Государь», «Его Величество» и «Императоръ» с глаголами справа (период 1820 по 1917 г.г.)

Может возникнуть опасение, что частота употребления слова «государь» завышена за счет часто используемого в XIX веке светского обращения «милостивый государь».

Рис. 3. Частотное поведение N-грамм «Государь» и «Милостивый Государь» с 1820 г.

Однако, как видно на рис. 3, частотность выражения «милостивый государь» настолько ниже частотности слова «Государь», что никак не искажает график последнего.

Рассмотрим частотное поведение имен руководителей СССР в книгах, зафиксированных в Google books Ngram Viewer (рис. 4).

Пики числа упоминаний трех руководителей СССР (Сталин, Хрущев, Брежнев) всегда приходятся на время их правления. Иначе выглядит кривая имени «Ленин». Этот политик посмертно стал сакральным символом коммунистической идеи и в этом качестве упоминается в текстах книг. Для имен трех советских политиков, долгое время занимавших пост руководителя страны, форма кривой частотного поведения имеет определенные черты сходства. Подъем, достижение пика, затем спад и период забвения, продолжительность которого зависит от того, сколько прошло времени с момента прекращения политической

деятельности до середины восьмидесятых годов двадцатого века. В случае Сталина можно говорить не столько о пике, сколько о плато продолжительностью почти десять лет. Эта модель ранее описана нами в [5].

Рис. 4. Частотное поведение имен руководителей СССР (1920–2000)

Благоприятным для нашего исследования является то обстоятельство, что фамилии политических лидеров страны хоть и не являются редкими, но не слишком распространены. Подобное исследование с именем, например, Медведев, было бы невозможным.

Рис. 5. Частотное поведение имен руководителей СССР и РФ (1950–2005)

На рис. 5 показаны кривые частотного поведения имён последних пяти руководителей СССР и первых двух руководителей РФ. Обращает внимание, что кривые имен «Хрущев», «Брежнев», «Андропов», «Черненко» следуют описанной нами выше модели, которую условно назовём «сталинской». Модель для имен «Горбачёв» и «Ельцин» не имеет резкого спада после завершения политической деятельности лица. Рассмотрим далее формы

представления трех политиков СССР и двух политиков России. Для этого был использован тег формирования десяти самых частотных биграмм с фамилией политика и существительным в левой позиции. Из полученных наборов биграмм мы отобрали те, кривые которых различимы на рисунках.

Самой частотной формой представления И.В. Сталина в текстах печатных документов являлось выражение «товарищ Сталин» (рис. 6). Заглавная «Т», вероятно, не носит специального характера, а связана с позицией биграммы в предложении. Значительно реже использовалась форма с полным именем и отчеством.

Рис. 6. Три наиболее частотных биграммы с именем «Сталин»

Н.С. Хрущева, как видно на рис. 7, чаще всего обозначали полным именем. Варианты выражения «товарищ Хрущев» встречаются реже. Отметим также, что в конце девяностых появляется форма «Никита Хрущев»

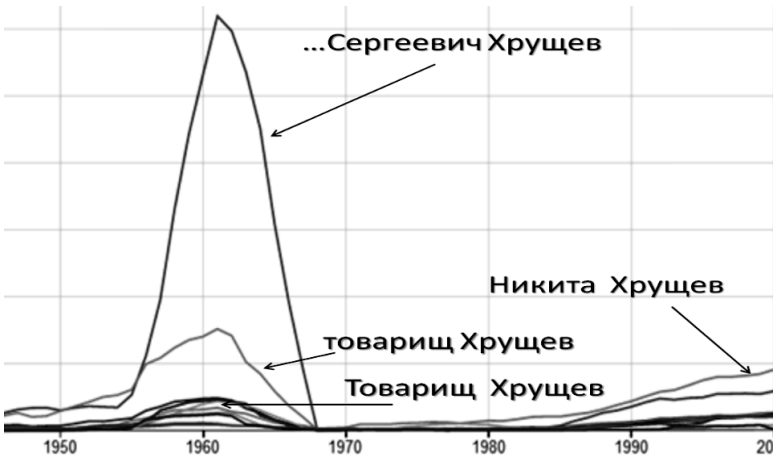

Рис. 7. Четыре наиболее биграммы с именем «Хрущев»

Биграмма «секретарь Брежнев», очевидно, является частью выражения «Генеральный секретарь Брежнев» (рис. 8). В отношении же Сталина и Хрущёва среди самых частотных биграмм указание на занимаемую должность отсутствует.

Рис. 8. Четыре наиболее частотных биграмм с именем «Брежнев»

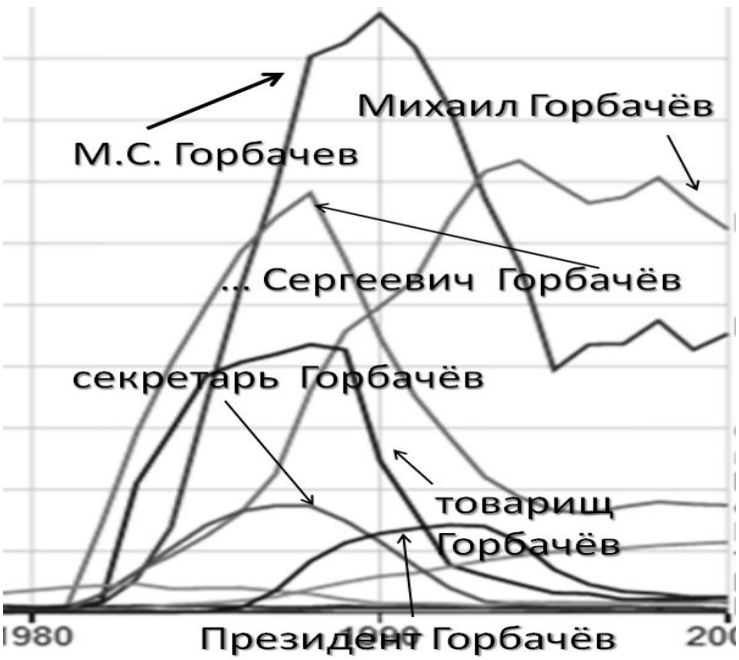

Рис. 9. Наиболее частотные существительные с именем «Горбачев»

Что касается первого и последнего президента СССР, то у него наиболее частотна форма с инициалами. Ее употребляемость выше, чем формы с полным именем. Появляется форма «президент Горбачев». При этом достаточно частотны биграммы «секретарь Горбачев» и «товарищ Горбачев» (рис. 9).

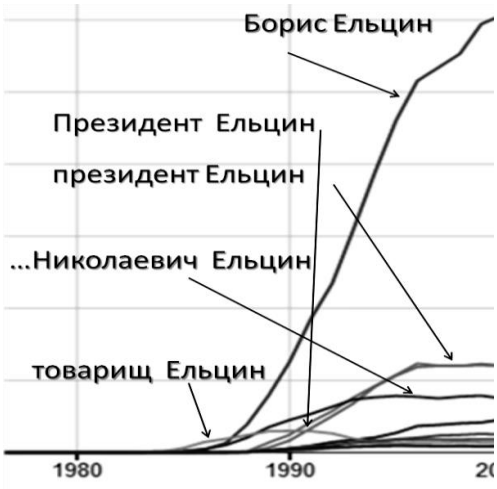

Рис. 10. Наиболее частотные биграммы с именем «Ельцин»

Интересно, что Б.Н. Ельцин чаще всего обозначается как «Борис Ельцин» (рис. 10). Кривые биграмм «Президент Ельцин» и «президент Ельцин» практически сливаются. Вполне различима кривая биграммы «товарищ Ельцин».

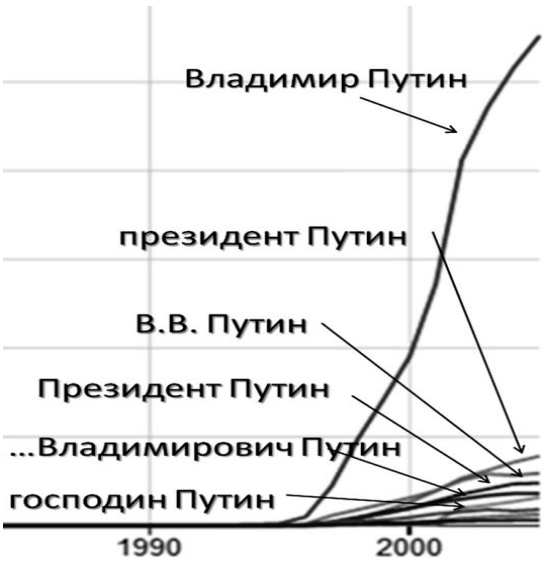

Рис. 11. Наиболее частотные биграммы с именем «Путин»

Полноценно проследить частотное поведение биграмм с именем «Путин» невозможно, поскольку корпус охватывает временной период только до 2008 года, Из рис. 11 видно, что наиболее частотной является форма «Владимир Путин», значения частотности прочих форм очень близки между собой. И, наверно, не будет ошибкой сказать, что именование публичных деятелей по имени и фамилии является особенностью русского языка нового времени.

Покажем одну особенность функционирования имен политических деятелей в функции субъектов в структуре предложения, а именно, сочетаемость их с глаголами (рис. 12).

Рис. 12. Сопоставление частотного поведения N-грамм имен глав СССР с глаголами в правой позиции словоформы «Государь» с глаголами в правой позиции

Для графика на рис. 12 были применены теги для формирования наборов биграмм с глаголами в правой позиции. В каждом из пяти образовавшихся наборов выбрана одна наиболее частотная N-грамма. Самый частотный глагол для «Государя» — «изволилъ», который предполагает после себя другой глагол в инфинитиве, напр., «изволилъ повелеть». Это была, очевидно, официальная словесная формула в Российской империи. Самый частотная биграмма с именем «Ленин» — «Ленин писал». Это легко объясняется — Ленин упоминается не как политик, а как теоретик, непререкаемый авторитет и, кроме того, как сакральный символ, и ссылка делается на его письменное наследие. Что касается трех других советских лидеров, то высокая частотность их имен приходится на периоды правления, и в текстах приводятся ссылки на их выступления — отсюда глагол «говорил».

Рассмотрим далее представление и частотное поведение имен зарубежных лидеров в корпусах американского английского и британского английского языков.

На рис. 13–15 показаны графики частотного поведения имен президентов США в разном представлении.

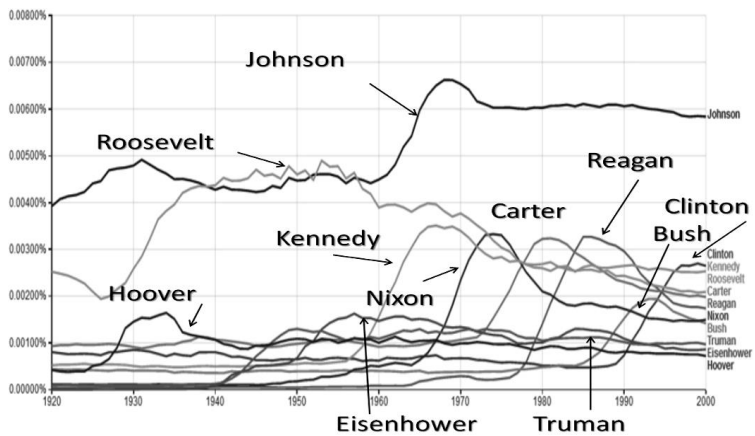

Рис. 13. Частотное поведение имен президентов США в корпусе американского английского языка Google Books (представление — фамилия)

При таком представлении, безусловно, повышается полнота поиска, однако согласно известной закономерности информационного поиска снижается его точность. Так, фамилия президента, заменившего в 1963 году убитого Джона Кеннеди, Джонсон — одна из самых

распространенных фамилий в англоязычных странах. На рис. 14 видно, что фамилия «Johnson» имеет самую высокую частотность, и соответствующая кривая имеет сравнительно мало подъёмов и снижений. Тем не менее, во время президентства Линдона Джонсона (1963–1969) кривая его фамилии дает выраженный подъём.

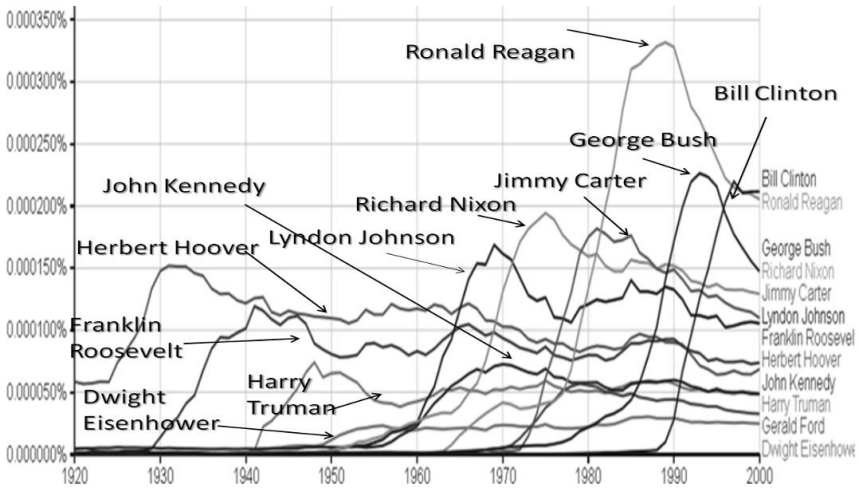

Рис. 14. Частотное поведение имен президентов США в корпусе американского английского языка Google Books (представление — имя, фамилия)

Однако для сравнения частотности упоминаний каждого президента целесообразно выявить данные, наиболее точно им соответствующие. Кривые на рис. 14 в основном изоморфны кривым на рис. 13, но при этом меняется соотношение между ними. Так на рис. 13 пик кривой фамилии Reagan находится примерно на одном уровне с пиками кривых фамилий Kennedy, Nixon, Carter, а пик фамилии Bush значительно ниже этого уровня. На рис. 14 пик кривой имени Ronald Reagan значительно выше пиков кривых Richard Nixon и Jimmy Carter. Пик кривой George Bush также выше этих двух пиков.

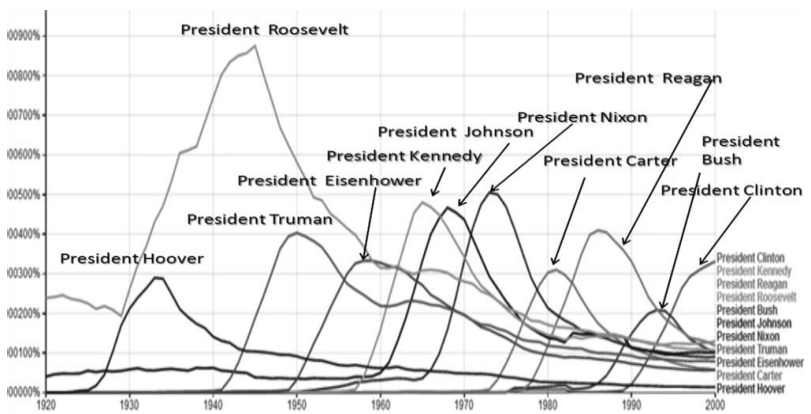

Рис. 15. Частотное поведение имен президентов США в корпусе американского английского языка Google Books (представление – статус, фамилия)

У большей части кривых после окончания срока президентства соответствующих персон отмечается некоторое снижение, которое в некоторых случаях выражено сильнее (Ronald

Reagan, George Bush), а в других слабее (Herbert Hoover, Franklin Roosevelt, Harry Truman, Dwight Eisenhower). Заметим, что более резкое снижение кривых характерно для исторически более поздних президентов.

Сочетание слова «President» (NB: с заглавной буквы) и фамилии лица дает серию изоморфных кривых, подъёмы, пики и снижения которых соответствуют срокам президентства (рис. 15). В отличие от рис. 14 снижения более выражены. По-видимому, в американской традиции слово President с заглавной буквы используется только в отношении действующего президента. На данном графике отчетливо видно, что пик кривой President Roosevelt намного выше, чем остальных кривых.

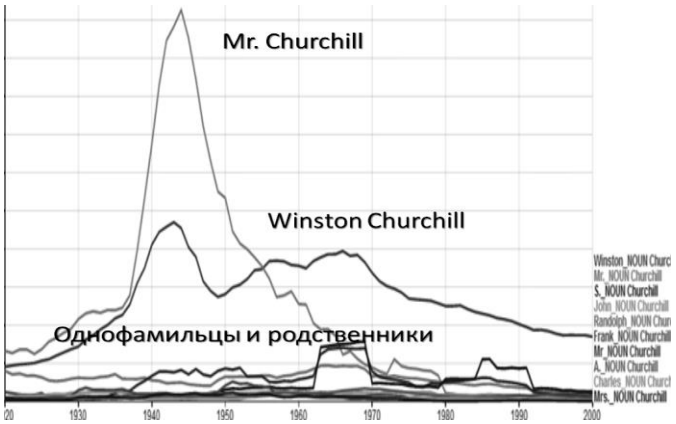

Рис. 16. Десять наиболее частотных биграмм с именем «Churchill» и существительными в левой позиции в корпусе британского английского

На рис. 16 показаны кривые частотного поведения биграмм с фамилией Churchill. Обращает внимание пик в середине сороковых годов биграммы Mr. Churchill. Высокий уровень такой формы представления вполне объясним. Черчилль был в это время премьер-министром. И это был разгар войны. Известно, что королева Елизавета II посвятила в рыцари премьер министра Черчилля во время второго срока его премьерства в 1953. Это обстоятельство отразилось в поведении кривой «Sir Winston Churchill» (рис. 17).

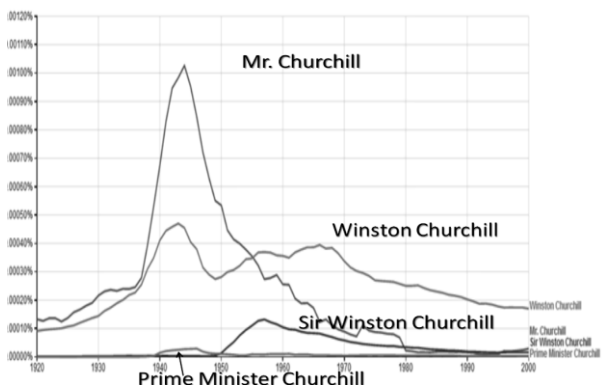

Рис. 17. Четыре варианта обозначения Уинстона Черчилля в корпусе британского английского языка

В число отобранных системой биграмм на рис. 16 вошли имена также предка премьер министра первого герцога Мальборро Джона Черчилль (John Churchill), сына премьер

министра Рандольфа Черчилля (Randolph Churchill), американского композитора Фрэнка Черчилля (Franc Churchill), а возможно и другого Фрэнка Черчилля — персонажа романа Джейн Остин «Эмма», британского поэта сатирика XVIII века Чарльза Черчилля (Charles Churchill) и, наконец, супруги премьер-министра, кривая которой соответствует сочетание «Mrs Churchill». Все эти кривые, однако, плохо различимы, т.к. расположены в нижней части графика, ближе к горизонтальной оси.

Пики кривых «Prime Minister Churchill» и «Sir Winston Churchill» значительно ниже пиков остальных двух. Кривая «Winston Churchill», на наш взгляд, наиболее точно отражает биографию этого политического деятеля. Снижение упоминаний в середине 1940-х, по-видимому, соответствует поражению на выборах в 1946 году, а последующий подъем - второму сроку пребывания на посту.

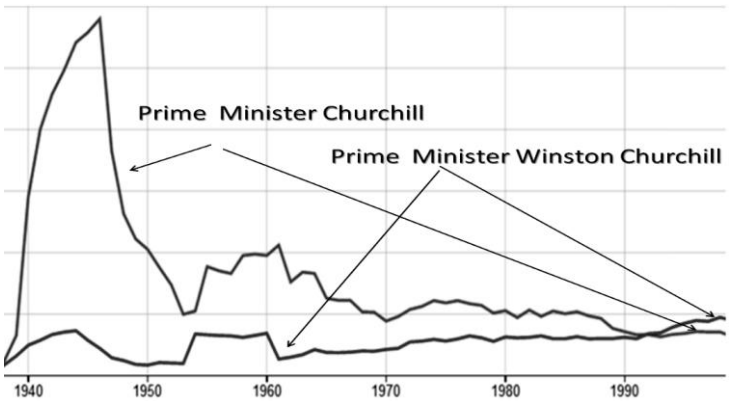

Рис. 18. Сопоставление частотного поведения N-грамм «Prime Minister Churchill» и «Prime Minister Winston Churchill»

Кривые N-грамм, содержащих название должности и имя лица в двух вариантах сходны по конфигурации с кривой «Winston Churchill» (рис. 17). В кривой «Prime Minister Churchill» подъемы и снижения частотности выражения выражены сильнее, чем в кривой «Winston Churchill», что, по-видимому, связано с включением в N-грамму названия должности, ведь именно с ней и связаны колебания кривой. Еще одно наблюдение состоит в том, что практически на всем протяжении выбранного периода кривая «Prime Minister Churchill» выше, чем кривая «Prime Minister Winston Churchill» и только в середине 1990-х частотность N-граммы с именем и фамилией становится чуть выше, чем кривая только с фамилией.

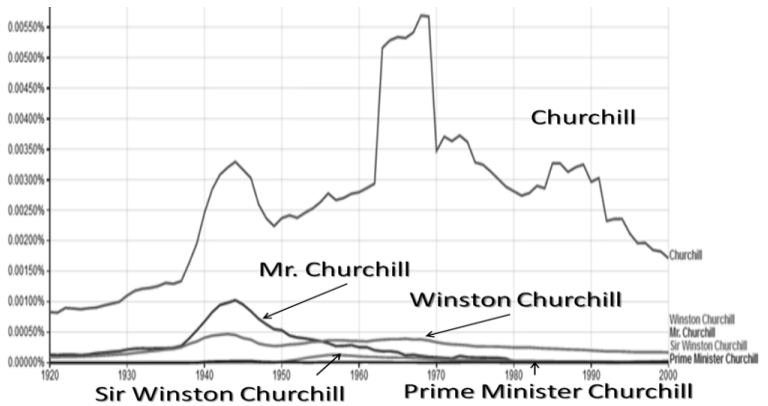

Рис. 19. Сопоставление кривых разных вариантов представления лица с кривой фамилии Churchill.

Понятно, что частотность фамилии «Churchill» значительно выше частотности всех форм (рис. 19). Однако заметим, что поведение кривой «Churchill» в целом, особенно до 1960 г., сходно с поведением кривых для других обозначений этого лица.

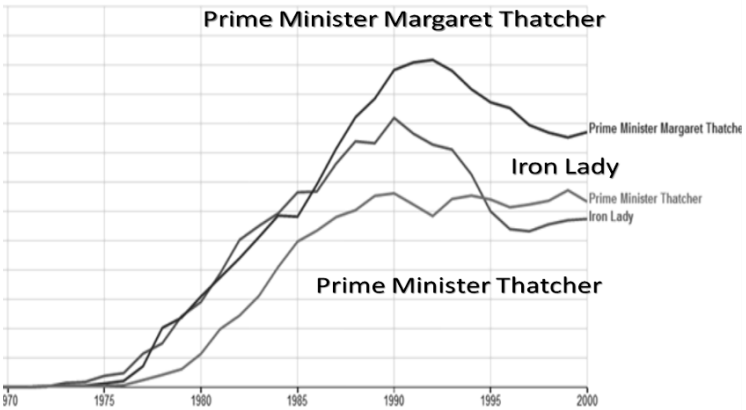

Рис. 20. Частотное поведение N- грамм с указанием должности с полным именем должности с фамилией, а также прозвища

Рассмотрим теперь несколько графиков частотного поведения имени другого британского премьер-министра — Маргарет Тэтчер (рис. 20–22).

Частотность форм обозначения лица в данном случае отличается от случая с Черчиллем, а именно, статус и полное имя более частотно, чем статус и фамилия, а частотность прозвища также довольно высока, причем частотность его снижается после прекращения полномочий и это более выражено, чем двух других N-грамм (рис. 20).

Кривые на рис. 21 показывают, что частотность формы «Mrs Thatcher» значительно выше других биграмм. Это, видимо, особенность британского английского языка или британского политического лексикона. Об этом же говорят и графики на рис. 22.

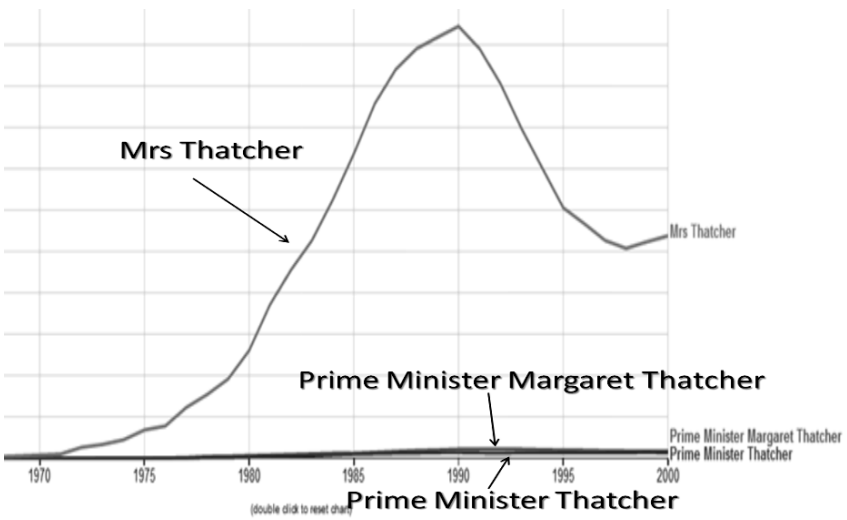

Рис. 21. Частотное поведение N- грамм с указанием должности с полным именем должности с фамилией, а также формы «Mrs Thatcher»

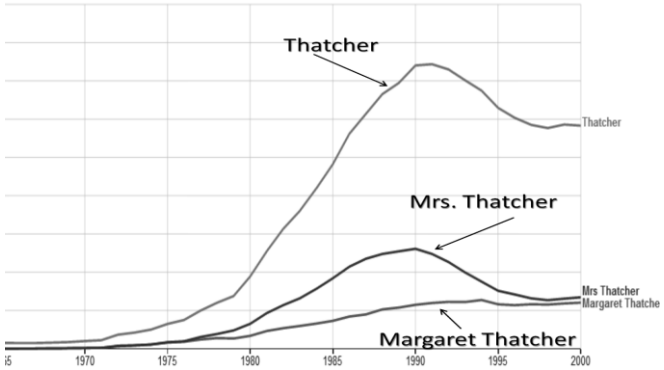

Рис. 22. Частотное поведение N- грамм «Margaret Thatcher», «Mrs Thatcher» и «Thatcher»

Ниже представлены графики частотного поведения фамилий нескольких британских премьер-министров (рис. 23–26).

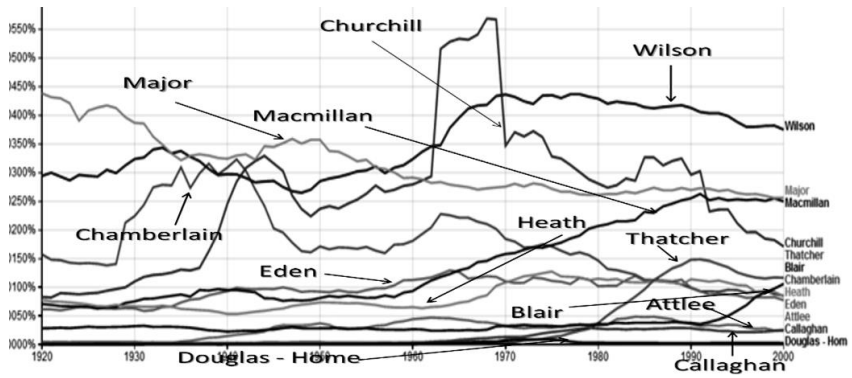

Рис. 23. Частотное поведение имен премьер министров Великобритании в корпусе британского английского языка Google Books (представление — фамилия)

Как видим, такие имена как «Wilson», «Major», являясь распространенными английскими фамилиями и, соответственно, не отражают историческую реальность, связанную с высокопоставленными носителями этих фамилий (рис. 23).

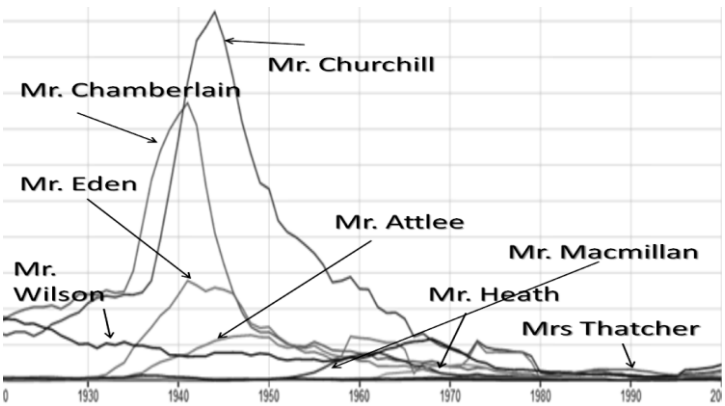

Рис. 24. Частотное поведение имен премьер министров Великобритании в корпусе Британского английского языка Google Books (представление Mr./ Mrs. фамилия)

Можно предположить, что высокий пик биграммы «Mr. Churchill» в 1944 г. связан с незаурядной личностью этого политика, его ролью во второй мировой войне (рис. 24). Вероятно, роль (позитивная или негативная) Невила Чемберлена в предвоенной политике была значительной, и потому пик его кривой – второй по высоте. Обращает внимание, насколько меняются соотношения между кривыми. Кривые имен политиков второй половины XX века значительно ниже, чем кривые первой половины. Можно было бы предположить, что это связано с неспокойной историко-политической ситуацией — войнами, возникновением национал-социализма, усилением коммунистического режима и пр., однако, нижеследующие графики не показывают такой корреляции.

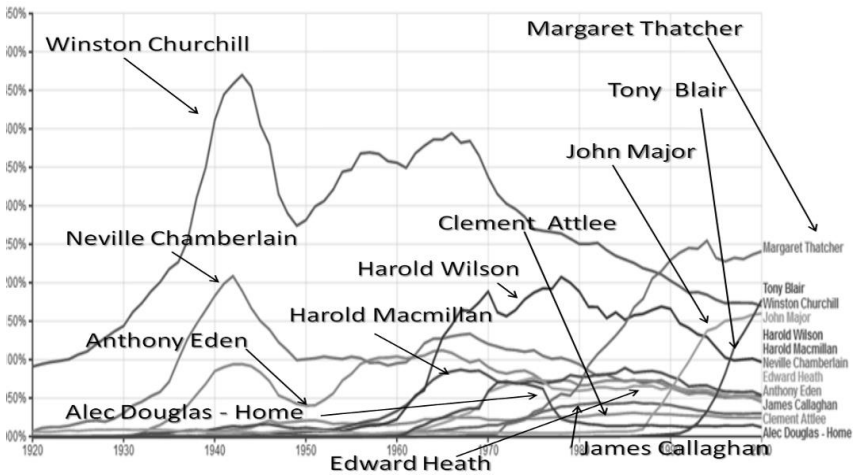

Рис. 25. Частотное поведение имен премьер министров Великобритании в корпусе Британского английского языка Google Books (представление – имя, фамилия)

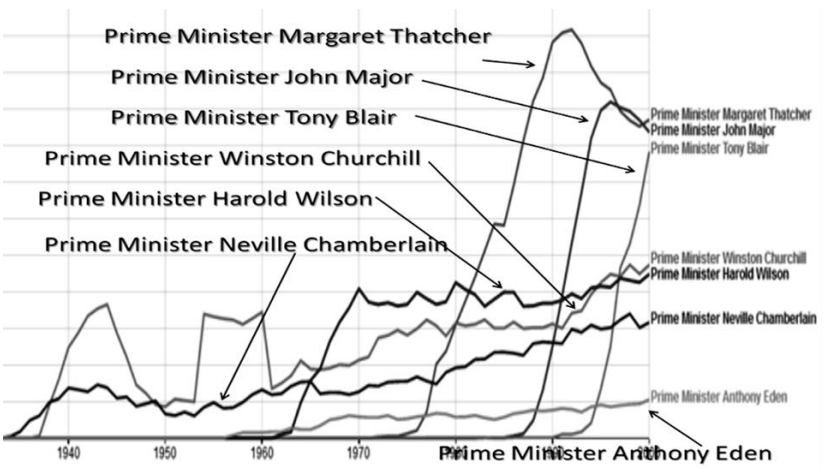

Рис. 26. Частотное поведение имен премьер министров Великобритании в корпусе Британского английского языка Google Books (представление – должность, имя, фамилия)

При указании имени и фамилии соотношения между кривыми снова меняются (рис. 25). Высота кривой биграммы «Winston Churchill» до середины 1980-х является наибольшей среди других кривых. Частотность биграммы «Margaret Thatcher» с конца 1970-х до 2000 г.

становится наиболее высокой, превышая в это время даже «Winston Churchill», что, впрочем, естественно для действующего премьер-министра. На рис. 26 отмечается значительный подъем кривых «John Major» и «Tony Blair» (форма представления – должность, имя, фамилия), которые на рис. 25 практически сливались с горизонтальной осью. Для них, как мы видим, характерно сочетание с личным именем и с должностью.

При таком представлении лидеров Великобритании мы видим картину, противоположную той, которая представлена на рис. 25. Пики кривых британских премьер-министров конца XX века значительно выше кривых премьеров начала века. Следует, однако, помнить, что эта форма представления (см. рис. 20, 22) наименее частотна среди всех рассмотренных форм.

Заключение

Частотное поведение имен политических деятелей отражает политические процессы, политические режимы, а в отдельных случаях и факты биографии политических деятелей.

При этом, как показано в настоящей статье, частотное поведение лексической единицы, означающей политического деятеля, зависит от способа представления: какая форма имени — полная или краткая — используется, указывается ли пост или должность политического деятеля.

В Российской империи имя императора указывалось намного реже, чем его пост, который обозначался лексическими единицами «Императоръ», «Государь», «Его Величество». При этом по нашим данным во второй половине XIX века встречаемость этих слов в текстах русских книг значительно падает (рис. 2). Существует определенный набор глаголов, по-видимому, строго регламентированный, который употреблялся для описания действий императора (рис. 12).

В СССР кривые числа упоминаний имени лидера в печатных документах имеют подъем, некоторое плато или пик, затем после смерти или прекращения полномочий лидера некоторый период малой частотности, затем во второй половине 1980-х гг. рост частотности (Сталин, Хрущев, Брежнев, Андропов, Черненко, см. рис. 4, 5). Эта модель уже была описана нами ранее [5]. Несколько иначе выглядит кривая имени Горбачев. Частотность этой фамилии снижается лишь незначительно. Сходной модели следуют кривые имен «Ельцин» и «Путин», хотя в отношении их возможности исследования ограничены — графики могут быть построены только до 2008 года.

Частотность вариантов представления российских и советских лидеров имеет следующие особенности: у советских лидеров редко упоминается должность. В силу, вероятно, того что названия их должностей довольно длинны. Лидеры СССР обозначаются словом «товарищ» и далее фамилией. Такое обозначение наиболее частотно для И.В. Сталина. Н.С. Хрущёв и Л.И. Брежнев чаще обозначались именем, отчеством и фамилией, чем словом «товарищ» (рис. 7, 8, 9). У Брежнева, тем не менее, встречается и указание на должность.

При представлении М.С. Горбачева самыми частотными формами являются имя и фамилия, а также фамилия с инициалами. Используется также форма «президент Горбачев» но также сравнительно часто, особенно в начале правления, встречается и «товарищ Горбачев».

Самые распространённые формы представления Б.Н. Ельцина и В.В. Путина — имя и фамилия, у Ельцина встречается форма со словом «товарищ», а у Путина «господин».

В англоязычных странах имеется проблемы широко распространённых имен. Такие имена как «Johnson», «Wilson», «Major» и т. п. не могут отражать, какие-либо исторические и биографические события или процессы в силу распространенности этих фамилий. В русском корпусе такую же проблему может представить фамилия «Медведев».

Рассмотрены три варианта графиков частотного поведения имен президентов США (рис. 14–16). В первом варианте (рис. 14) использованы только фамилии. Заметно, что

распространённая фамилия Johnson имеет более-менее пологую кривую с некоторым подъёмом в годы президентства Линдона Джонсона. Кривые имен президентов второй половины XX века практически изоморфны.

График, где кривые построены по имени и фамилии президента (рис. 15), имеет следующие особенности. Кривые имен исторически более ранних президентов, достигнув определенной точки, не снижаются или снижаются незначительно. Кривые более поздних президентов достигают более высоких пиков частотности, но за пиками, как правило, годами окончания президентства, следует выраженное снижение.

На графике, построенном на биграммах из слова «President» и фамилий (рис. 15), виден ряд изоморфных кривых, каждая из которых точно соответствует периоду правления соответствующего президента. Можно утверждать, что эта форма обеспечивает высокую точность поиска.

На двух примерах рассмотрены формы обозначения премьер-министров Великобритании. Существует форма, состоящая из фамилии. Эта форма обеспечивает полноту поиска, но в то же время может дать некоторый информационный шум, иначе говоря, отразить в результатах поиска однофамильцев и родственников (рис. 17).

В британской традиции премьер-министр страны обозначается несколькими способами: Mr / Mrs и фамилия, имя и фамилия, Prime Minister имя и фамилия, Prime Minister фамилия (рис. 17–23). Наши данные показывают, что форма Mr / Mrs и фамилия более частотна в период пребывания политика на посту премьер министра (рис. 17–23).

При построении графиков на основе разных вариантов именований мы видим изменение соотношений кривых в зависимости от выбранного варианта (рис. 24–27).

Таким образом, при проведении диахронических исследований частотности имен политических деятелей следует учитывать варианты их представления в печатных документах.

После пилотного исследования, позволяющего выявить различные варианты именования политических деятелей, должно последовать методическое решение относительно конкретных запросов и конкретных имен. Возможно, например, включать характерные варианты в запрос как члены дизъюнкции. Другой вариант — при выраженном изоморфизме кривых использовать наиболее частотную N-грамму.

Система Google Books Ngram Viewer является мощным инструментом для диахронических исследований как отдельных языков, так для сравнительных межъязыковых исследований. При этом она имеет и существенные недостатки. Необходимы дополнительные методологические разработки для уточнения связи между объемами корпусов и точностью результатов исследований. Мы, тем не менее, считаем, что Google Books Ngram Viewer — на сегодняшний день является единственным эффективным инструментом для исследований подобного рода.

Литература

[1] Захаров В.П., Масевич А.Ц. Опыт корпусно-ориентированного историко-культурного исследования исторической и политической лексики // Библиосфера. 2016. №2. С.47–56.

[2] Масевич А.Ц. Захаров В.П. Методы корпусной лингвистики в исторических и культурологических исследованиях // Компьютерная лингвистика и вычислительные онтологии: сб. научн. статей. Труды XIX Международной объединенной научной конференции «Интернет и современное общество» (IMS-2016), Санкт-Петербург. СПб: Университет ИТМО, 2016. С. 24–43.

[3] Масевич А.Ц., Захаров В.П. Диахроническое исследование лексико-семантического поля «враги»// Труды международной конференции «Корпусная лингвистика – 2017». СПб.: Изд-во С.-Петерб. ун-та, 2017. С. 248–254.

[4] Масевич А.Ц., Захаров В.П. Семантические трансформации политической лексики // Сборник научных статей XIX Объединенной конференции «Интернет и современное общество» IMS-2017. СПб: Университет ИТМО, 2017. С. 107–120.
[5] Масевич А.Ц., Захаров В.П. Модели частотного поведения русской политической лексики XX века // Вестник Новосибирского государственного университета. Серия: Лингвистика и межкультурная коммуникация. 2017. Т. 15, № 2. С. 30–46.
[6] Masevich A., Ostrovskaya A. V. The Authority for Names of Persons of Eighteen and Nineteen Century Russia in the Institute for Studies in Russian Literature: a Utopian Project // The Scholar and Database: paper presented on 4 November 1999 at the CERL conference hosted by the Royal library, Brussels / Ed. by Lotte Hellinga. London, 2001. P. 79–90.
[7] Zakharov V.P., Masevich A.C., Pimenov E.N. Authority Control as a Linguistic Support Element of an Automated Library System // International Cataloguing and Bibliographic Control. 1996. Vol. 25, №. 4. P. 84–86.
[8] Вершинина Л.П., Вершинин М.И., Масевич А.Ц. Построение модели поиска в электронном каталоге библиотеки на основе нечеткого отношения сходства // Библиосфера. 2013. №2. С.44–81.
[9] Michel J-B. et al. Quantitative Analysis of Culture Using Millions of Digitized Books science // Science. 2011. Vol. 331. P. 176. DOI: 1126/Science.1199644.
[10] Захаров В.П., Масевич А.Ц. Диахронические исследования на основе корпуса русских текстов Google Books Ngram Viewer // Структурная и прикладная лингвистика. 2014. Вып. 10. С. 303–327.
[11] Galeev T.I., Solovyev V.D. Methods of Application of Modern Text Corpora in the Study of the Morphological System of Russian Verbs Unification of I Productive (irregular) Class of Verbs. Quantitative Model Based on Google Books // Modern Journal of Language Teaching Methods (MJLTM), (Dec. 2016). P. 177–180.
[12] Bochkarev V., Solovyev V., Wichmann S. Universals Versus Historical Contingencies in Lexical Evolution // Journal of the Royal Society Interface. 2014. Vol. 11: 20140841. DOI: 10.1098/rsif.2014.0841.
[13] Масленникова Ю.С., Бочкарев В.В., Соловьев В.Д. Вероятностная модель для оценки объема лексикона по данным корпуса Google Books Ngram. 2017. С. 255-260.
[14] Захаров В. П., Масевич А. Ц. Лингвистическая картина российской истории XX века: корпусное исследование. URL: https://www.academia.edu/29209763/ Лингвистическая_картина_российской_истории_XX_века:_корпусное_исследование_ Linguistic_portrait_of_20th_century_Russian_history_a_corpora-based_study (дата обращения: 22.04.2018).
[15] Ляшевская О. Н., Шаров С. А., Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М.: Азбуковник, 2009.

The Variability of the Representation of Politicians' Names in Diachronic Studies on the Base of Text Corpora

A.Ts. Masevich[1], V.P. Zakharov[2]

[1] Saint-Petersburg Institute of Culture, [2] Saint-Petersburg State University

The paper continues a number of publications dealing with diachronic studies of frequency behavior of political terms using Google Books Ngram Viewer. In the preceding papers we have demonstrated how changing of frequency of certain lexical units reflects actual historical and cultural processes and described some models of frequency behavior of political terms in the massive of published texts. The present publication is written from the methodological point of view. We consider which way the form of personal name and position of a person can influence the frequency behavior of the lexical unit and how this influence should be considered in diachronic studies. The study covers a large period — from the middle of 19th century until the end of 20th century. We consider frequency behavior of more than 30 proper names of state leaders of Russian Empire, the Soviet Union, the Russian Federation, USA and UK in the texts in Russian, British and American versions of English. In the obtained material, we made certain observations, which can be useful in forthcoming diachronic studies.

Keywords: corpora, diachronic studies, political names, political terms, methodology, Google Books Ngram Viewer

Формализованный подход к установлению связи и роли понятий

С.В. Микони

Санкт-Петербургский институт информатики и автоматизации РАН

smikoni@mail.ru

Аннотация

Эвристическому и экспертному подходу к установлению соотношения понятий в онтологических моделях противопоставляется формализованный подход. Метод определения роли каждого понятия в паре сопоставляемых понятий зависит от конкретного вида связи. Рассматриваются три группы связей: родовидовые, агрегатные и функциональные. Для каждой из групп предлагается использовать свой метод анализа. Логико-лингвистический анализ, применяемый для анализа родовидовых связей, заключается в извлечении существенных признаков из определений сопоставляемых понятий и в теоретико-множественных операциях над их содержаниями и объёмами. Приводится алгоритм установления ролей понятий, находящихся в отношении вид-род. Для установления роли понятий в рамках агрегатных и инструментальных связей предлагается привлекать закономерности системного анализа. Разновидности функциональной связи различаются степенью влияния факторов, сопутствующих сопоставляемым понятиям. Степень влияния предлагается устанавливать методами корреляционного анализа. Показана ситуативность роли анализируемого понятия, меняющейся при его применении в рамках другого вида связи.

Ключевые слова: понятие, отношение, связь, роль, род, вид, часть, агрегат, целое, обобщение, агрегация, инструментальная связь, связь-влияние

Введение

Установление типа связи в каждой паре понятий и их роли в этой связи востребовано при построении онтологических моделей различного назначения. Математической моделью связи двух понятий $a, b \in A$ из множества A является пара $(a, b) \in R$, принадлежащая некоторому антисимметричному отношению $R \subseteq A \times A$ [1]. В силу антисимметрии всех рассматриваемых в работе отношений роли понятий в паре (a, b) различаются (род-вид, цель-средство и т.п.).

В отсутствие формальных моделей задача определения типа связи и ролей понятий решается эвристически, либо экспертным путём. Примером эвристического подхода может служить шкала уровней качества систем с управлением [2]. Качество этой модели было исследовано в работе [3]. В работе [4] эвристическому подходу к установлению связи между понятиями противопоставлен экспертный подход. Однако практические знания экспертов и их интуиция не являются гарантией правильных решений. Проблема усугубляется перегруженностью русского языка иностранными терминами, имеющими неоднозначную трактовку.

Уменьшению степени субъективизма в решении рассматриваемой проблемы должна способствовать более высокая степень формализации связей между понятиями. За объекты формализации примем связи, используемые в семантических сетях и изучаемые в рамках искусственного интеллекта [5] и системного анализа [6]. Формализацию

нахождения связей и роли в них сопоставляемых понятий предварим их исчерпывающим описанием.

1. Связи и роли понятий

1.1. Отношение *вид-род*

Отношение *вид-род* реализует *связь-обобщение* (*generalization*). Обобщение представляет собой переход от частных понятий к общему путём исключения различающих их особенностей. Простейшим примером является обобщение мужчин и женщин понятием *человек* там, где признак пола не существен, например, «в выборах приняло участие *N* человек», поскольку мужчины и женщины в России обладают одинаковым избирательным правом.

Обратной обобщению является операция *конкретизации* понятия, применяемая в тех случаях, когда требуется учесть особенности видов. Конкретизация родового понятия в видовые требует привлечения системообразующего признака, называемого *основанием деления* [7]. Применительно к примеру с мужчинами и женщинами таковым признаком является пол, имеющий два значения *мужской* и *женский*. Они называются *видовыми отличиями*. *Видовые* понятия, порождённые из родового понятия с применением противоположных видовых отличий, называются *координатными* понятиями [7].

Понятие, обладающее признаками обоих координатных понятий, называется *собирательным*. Применительно к рассматриваемому примеру им является ныне принятый так называемый *третий пол*. Понятие, представляющее собой результат объединения видовых понятий, порождённых разными родовыми понятиями, называется *межвидовым* понятием. Привлечём к предыдущему примеру родовое понятие *лыжные гонки*, порождающее гонки *без ружья* и *с ружьём* (биатлон). Объединяя видовые понятия разного происхождения: *женщина* и *биатлон*, получаем межвидовое понятие *женский биатлон*.

1.2. Отношение *часть-агрегат*

Отношение *часть-агрегат* реализует *связь-агрегацию* (лат. *aggregatio* присоединение). Под термином *агрегат* (лат. *aggregatus*, соединенный, собранный) понимается совокупность элементов, образующих систему или её часть. Трактовка самого агрегата как части системы придаёт этому термину смысл незавершённости по отношению к системе. Согласно такому пониманию для описания агрегата A, собираемого из k частей A_i, $i=\overline{1,k}$, достаточно выражения:

$$\bigcup_{i=1}^{k} A_i = A \qquad (1)$$

Часть A_i находится с агрегатом в отношении включения: $A_i \subset A$. Например, сборный дом представляет собой *агрегат* по отношению к своей части — крыше. Части искусственного агрегата содержатся в сформулированном для него перечне.

1.3. Отношение *часть-целое*

Отношение *часть-целое* реализует *связь-композицию* (лат. *compositio* — составление целого из частей). По своей сути процессы *композиции* и *агрегации* совпадают. В обоих случаях выполняется сборка из частей, но в отличие от агрегата понятие *целое* характеризуется дополнительно свойствами *целостности*, не присущими его частям [8]. Таким образом, целое помимо условия (1) должно отвечать условию (2):

$$A = \{ A_i \mid Pr(A \setminus A_i \cup A_i)\} \qquad (2)$$

Аргумент $A \setminus A_i \cup A_i$ предиката *Pr* означает необходимость дополнения агрегата $A \setminus A_i$ до образования целого, обладающего свойством целостности. С учётом дополнительного условия (2) *композиция* представляет собой частный случай *агрегации.* Согласно выражению (2) понятие *a*, описывающее *A*, не является целым без любой части A_i, $i = \overline{1,n}$.

В технике составление целого из частей называют *сборкой*, а в науке — *синтезом*. Отношение *целое-часть* реализуется в *перечислительном* определении понятия. Например, неделю можно определить через перечисление входящих в неё дней.

С точки зрения такого свойства целого, как *воспроизводство населения*, целым следует считать, объединение двух частей человечества: женщин и мужчин, учитывая тот факт, что третий пол не принимает участия в воспроизводстве населения Земли. С точки зрения потребляемых человечеством ресурсов третий пол также будет отнесён к народонаселению Земли. А это означает, что присвоение понятию категории целого относительно и зависит от сформулированной цели.

Обратной *композиции* является операция *декомпозиции* (расчленение целого на части). В технике она называется *разборкой*, а в науке — *анализом*. Примером декомпозиции понятия *неделя* является перечисление имён составляющих её семи дней.

1.4. Отношение *элемент-класс*

Отношение *элемент-класс* реализует *связь-принадлежность***.** Принадлежность элемента *x* множеству *X* определяется относительно свойства *Pr*, присущего всем его элементам [9]:

$$X = \{x \mid Pr\,(x)\} \tag{3}$$

Истинность одноместного предиката *Pr*(*x*) свидетельствует о том, что элемент *x* обладает заданным свойством *Pr*. Таким образом, предикат *Pr*(*x*) представляет собой *логическое правило* формирования множества *X*. Оно характеризует факт принадлежности элемента *x* множеству *X* и на этой основе называется его *характеристической функцией*: $Pr\colon X \to \{0, 1\}$. Значение 0 (ложь) означает, что $x \notin X$, а значение 1 (истина) означает, что $x \in X$. Свойству множества (класса) *X*, формализуемому одноместным предикатом *Pr*, соответствует существенный признак, извлекаемый из *определения* понятия.

Если элементы множества *X* должны обладать *одновременно* $n>1$ свойствами, отнесение к нему элемента *x* осуществляется по *n* логическим правилам:

$$A = \left\{ x : \bigwedge_{j=1}^{n} Pr_j(x) \right\}. \tag{4}$$

Пример. *Приспособление для сидения* принадлежит к классу *стул*, если оно удовлетворяет всем существенным признакам множества {*Приспособление для фиксации сидячей позы, Плоская поверхность, Ножки, Спинка*}.

1.5. Отношение аргумент-функция

Отношение *аргумент-функция* реализует *функциональную* связь или *связь-влияние* (*influence*). Представим влияние факторов $a_1,\ldots,a_i,\ldots,a_n$ на фактор *a* как функциональную зависимость $a = f(a_1,\ldots,a_i,\ldots,a_n)$. Различают следующие виды зависимости.

а) **Причина-следствие** (каузативная связь) — зависимость *a* от a_i настолько сильна, что можно пренебречь зависимостью *a* от других аргументов. В этом случае фактор a_i можно считать *причиной* возникновения *a*, а сам фактор *a* — её *следствием*. Иными словами, фактор a_i является *необходимым* и *достаточным*: если событие a_i произошло, то обязательно произойдет событие *a*, и наоборот, если случилось второе, то обязательно этому предшествовало первое (обратная функция). Эту связь между *a* и a_i можно рассматривать как *функцию одного аргумента* $a=f(a_i)$. Причинно-следственная связь

между двумя факторами является приблизительным описанием реальности, когда наличием взаимодействий с другими факторами можно пренебречь.

Пример. За вспышкой молнии *обязательно* последует гром. Он слышен, если пренебречь возможной блокировкой слухового канала (стена и пр.). Обратная функция: если прогремел гром, была молния. Поскольку возникновение молнии ***необходимо*** и ***достаточно*** для того, чтобы прогремел гром, она рассматривается как причина, а гром — как её следствие.

б) **Суммирующая** связь. Она имеет место в том случае, когда зависимость a от любого из аргументов является *необходимой*, но *не достаточной*. Событие a может произойти только при наличии помимо наличия i-го фактора a_i (основной причины) совокупности остальных n–1 факторов (сопутствующих причин), т.е. $a = f(a_1,\ldots,a_i,\ldots,a_n)$. Для выявления этого вида связи необходимо обеспечить полноту совокупности причин, сопутствующих основной причине и оценить степень их влияния на исследуемое понятие [10].

Пример. Для выпадения осадков нужна туча (*основная причина*), но её *не достаточно*. Нужны *сопутствующие причины* (низкое давление, слабый ветер и пр.).

в) **Альтернативная** связь**.** Она характеризуется зависимостью a от любого из n факторов в отдельности. Иными словами, событие может быть вызвано разными причинами. Для выявления этого вида связи следует найти все возможные причины возникновения события.

Пример. Высокая температура тела может быть вызвана различными заболеваниями.

г) **Транзитивная связь.** Эта связь осуществляется через промежуточные факторы. Фактически она реализуется последовательным выполнением суммирующей связи. После принятия одной из причин за основную причину следует последовательная реализация сопутствующих причин. Примером может служить транзитивная связь между исходными данными и результатом алгоритма, связанными цепочкой промежуточных данных.

1.6. Отношение *средство-цель*

Отношение **средство-цель** реализует ***инструментальную*** связь**,** поскольку средство играет роль инструмента в достижении цели. Роль средства двойственна. Являясь *средством* реализации цели верхнего уровня, оно является целью для нижнего уровня иерархии. Достижение цели может осуществляться *разными средствами*, что соответствует альтернативной связи. В общем случае отдельно взятое средство недостаточно для реализации поставленной цели, что соответствует суммирующей связи.

Пример. *Цель* — доставка груза. *Средство* доставки — автомобиль, самолёт (альтернативная связь). Для доставки на дальнее расстояние необходимо привлекать и то, и другое средство (суммирующая связь).

Рассмотрим аналитические формы влияния на примере функции трёх аргументов $a=f(a_1, a_2, a_3)$, представленные в табл. 1.

Таблица 1. Варианты зависимости

Номер	a_1	a_2	a_3	$f_1(a)$	$f_2(a)$	$f_3(a)$
0	0	0	0	0	0	0
1	0	0	1	0	0	1
2	0	1	0	0	0	1
3	0	1	1	0	0	0
4	1	0	0	1	0	1
5	1	0	1	1	0	0
6	1	1	0	1	0	0
7	1	1	1	1	1	0

Если первый аргумент a_1 принять за причину, пренебрегая влиянием факторов a_2 и a_3, функция f_1 отражает связь *причина-следствие*: $a = f_1(a_1, a_2, a_3) = a_1$ с необходимостью причины a_i ($a = f(a_i)$) и достаточностью следствия a ($a_i = f^{-1}(a)$).

Функция f_2 отражает суммирующую связь, как зависимость следствия a от всех трёх аргументов: $a = f_2(a_1, a_2, a_3) = a_1 \wedge a_2 \wedge a_3$, а при n аргументах:

$$a = f_2(a_1,...,a_i,...,a_n) = \bigwedge_{i=1}^{n} a_i$$

Функция f_3 отражает альтернативную связь, как зависимость следствия a от каждого аргумента в отдельности:

$$a = f_3(a_1, a_2, a_3) = \overline{a_1} \cdot \overline{a_2} \cdot a_3 \vee \overline{a_1} \cdot a_2 \cdot \overline{a_3} \vee a_1 \cdot \overline{a_2} \cdot \overline{a_3}.$$

После преобразования формула выражается через функцию альтернатива, причём любой из дизъюнктивных членов выражения является избыточным:

$$a = f_3(a_1, a_2, a_3) = \overline{a_1} \cdot (a_2 \oplus a_3) \vee \overline{a_2} \cdot (a_1 \oplus a_3) \vee \overline{a_3} \cdot (a_1 \oplus a_2).$$

2. Формализация родовидовых связей

Любое понятие обозначается именем (термином) и характеризуется содержанием и объёмом [7]. Под *содержанием* понятия будем понимать совокупность существенных признаков, которыми обладают объединяемые в понятие сущности, а под *объёмом* — совокупность сущностей, обладающих этими признаками. Имя отвлечённого понятия будем обозначать латинской прописной буквой (курсив), содержание понятия — символом C, а его объём — символом V. Характеризуемое ими понятие будем помечать аргументом соответствующего символа, например, содержание понятия a обозначается символом C(a), а его объём символом V(a). Совокупности признаков и характеризуемых ими сущностей представляют собой множества. В силу конечного числа известных существенных признаков содержание понятия a задается в перечислительной форме, а его объём, представляющий собой открытое множество (класс) — в описательной форме:

$$C(a) = \{C_1(a),\ldots,C_j(a),\ldots,C_k(a)\},$$
$$V(a) = \{a_i \mid \forall\, C_j(a) \in C(a)\},$$

Определим отношения и теоретико-множественные операции между понятиями b и d, с содержаниями:

$$C(b) = \{C_1(b),\ldots,C_i(b),\ldots, C_k(b)\},$$
$$C(d) = \{C_1(d),\ldots,C_j(d),\ldots, C_m(d)\}.$$

Содержание понятия a, объединяющего содержания понятий b и d, должно включать все признаки $C_i(b)$ и $C_j(d)$, представляющие каждое понятие b и d в отдельности:

$$C_a = \{C_1(b),\ldots,C_i(b),\ldots, C_k(b), C_1(d),\ldots,C_j(d),\ldots, C_m(d)\}. \quad (5)$$

Согласно формуле (5) содержание понятия a шире, чем содержание исходных понятий b и d, что соответствует отношениям включения:

$$C(b) \subset C(a),\ C(d) \subset C(a), \quad (6)$$

выражаемым через операцию объединения:

$$C(a) = C(b) \cup C(d). \quad (7)$$

Определим теоретико-множественные операции и отношения между понятиями b и d с объёмами V(b) и V(d). Между их объёмами имеют место обратные отношения включения:

$$V(b) \supset V(a),\ V(d) \supset V(a). \quad (8)$$

Они выражаются через операцию пересечения:

$$V(a) = V(b) \cap V(d) \quad (9)$$

Отношения (6) и (8) и их следствия иллюстрируют закон двойственности содержания и объёма понятия [7]: чем обширнее набор признаков, составляющих содержание понятия,

тем уже класс объектов, удовлетворяющих им, и, наоборот, чем у́же содержание понятия, тем шире его объём. Отношения совместимости между понятиями *a* и *b* выражаются через их содержание следующим образом:

1) $C(a) \cap C(b) = \varnothing$ — понятия *a* и *b* несравнимы;
2) $C(a) = C(b)$ — понятия *a* и *b* равнозначны (по содержанию);
3) $C(a) \subset C(b)$ — понятие *a* подчинённое, а *b* подчиняющее;
4) $C(a) \cap C(b) \neq \varnothing$ — пересекающиеся понятия (имеют общие признаки).

Родо-видовая связь понятий естественно выражается через содержание понятий. Если исходное понятие *a* принять за *родовое*, а понятие *b* — за *видовое отличие*, то содержание получаемого на их основе *видового* понятия *ab* определяется на основе формулы (7):

$$C(ab) = C(a) \cup C(b) \quad (10)$$

Ранг ρ понятия *ab* на единицу больше ранга исходного понятия *a*. Величина ρ определяется количеством видовых отличий, которые привлекаются для образования требуемого вида понятия. Ранг ρ обладает свойствами метрики, позволяя количественно оценивать степень родства понятий.

Объём видового понятия *ab* определяется на основе формулы (9), двойственной формуле (7):

$$V(ab) = V(a) \cap V(b) \quad (11)$$

Согласно формуле (11) объём видового понятия (*ab*) меньше объёма родового понятия *a* (а его содержимое больше).

Сущность наследования признаков [7] состоит в том, что предметы, входящие в объём делимого понятия, распределяются по группам. Делимое понятие рассматривается при этом как родовое, и его объём разделяется на соподчиненные виды. Основанием деления является признак, значениями которого являются видовые отличия, например, цвет со значениями *красный, жёлтый, зелёный*.

Видовое понятие ab_i(æ) с видовым отличием b_i, полученным по основанию деления æ, называется *координатным*. Его содержание соответствует формуле (10):

$$C(ab_i)_{æ} = C(a) \cup C(b_i)_{æ} \quad (12)$$

Объём координатного понятия ab_i(æ) выражается через объём родового понятия с применением формулы (11):

$$V(ab_i)_{æ} = V(a) \cap V(b_i)_{æ} \quad (13)$$

Объёмы координатных понятий ab_i(æ) и ab_j(æ), $i, j = \overline{1,k}$, $i \neq j$, полученных по одному основанию деления æ, не совпадают:

$$V(ab_i)_{æ} \cap V(ab_j)_{æ} = \varnothing. \quad (14)$$

Например, если за основание деления понятия *контроль* принять способ размещения средства контроля относительно объекта контроля (вне или внутри последнего), то объёмы полученных видовых понятий *внешний контроль* и *внутренний контроль* не совпадают (пересечение их является пустым). Данный пример иллюстрирует также полное деление понятия *контроль* по выбранному основанию деления, так как последнее не порождает других членов деления, кроме приведённых выше. Деление исходного понятия *a* по основанию æ на *два* видовых понятия называется *дихотомическим*. Примером неполного деления понятия *контроль* по основанию деления *объект контроля* являются видовые понятия *параметрический* и *функциональный контроль*. Помимо параметров и функций устройства *объектами контроля* могут быть внешний вид устройства, его габариты и другие свойства.

Координатные понятия ab_i и ab_j, полученные по *одному* основанию деления æ, могут объединяться в агрегат, образуя *собирательное* или смешанное понятие ab_{ij}(æ) с содержанием:

$$C(ab_{ij}) = C(ab_i)_{æ} \cup C(ab_j)_{æ}. \quad (15)$$

Примером наложения голубого и зелёного цветов является жёлтый цвет. Лицам, обладающими признаками обоего пола, присваивается третий пол.

Координатные понятия $ab_i(æ)$ и $ab_j(η)$, полученные по *разным* основаниям деления, образуют *межвидовое* понятие $ab_{ij}(æ,η)$ с содержанием:

$$C(ab_{ij})_{æη} = C(ab_i)_{æ} \cup C(ab_j)_{η} \quad (16)$$

и объёмом:

$$V(ab_{ij})_{æη} = V(ab_i)_{æ} \cap V(ab_j)_{η} \quad (17)$$

Примером межвидового понятия является *внешний параметрический контроль*. Видовое отличие *внешний* является значением признака *способ размещения средства контроля* (внешний или внутренний), а видовое отличие *параметрический* является значением признака *объект контроля* (параметр или функция).

Утверждение 1. В параллельной классификации каждое межвидовое понятие, порождаемое из любых двух координатных понятий, полученных по *различным* основаниям деления, имеет непустой объём [7].

Согласно формуле (14) пустой объём даёт пересечение объёмов координатных понятий, полученных по одному основанию деления. Однако условиям утверждения 1 соответствует $æ \neq η$ и $æ \neq φ(η)$, что доказывает его справедливость, т.е.

$$V(ab_i)_{æ} \cap V(ab_j)_{η} \neq \varnothing.$$

Утверждение 2. Межвидовое понятие $ab_{ij}(æ,η)$, порождаемое координатными понятиями $ab_i(æ)$ и $ab_j(η)$, $æ \neq η$ и $æ \neq φ(η)$, представимо последовательной родо-видовой решеткой [7].

Раскроем формулу (16), определяющую содержание порождаемого понятия $ab_{ij}(æη)$ через родовое понятие $a_ρ$ и видовые признаки, с помощью формулы (15):

$$C(ab_{ij})_{æη} = C(ab_i)_{æ} \cup C(ab_j)_{η} = (C(a) \cup C(b_i)_{æ}) \cup (C(a) \cup C(b_j)_{η}) = C(a) \cup C(b_i)_{æ}) \cup C(b_j)_{η}.$$

Полученное выражение характеризуют последовательный процесс порождения понятия $ab_{ij}(æη)$ на основе родового понятия $a_ρ$ и видовых отличий $b_i(æ)$ и $b_j(η)$, причем очерёдность их привлечения безразлична. Отсюда следует, что порождение межвидового понятия $ab_{ij}(æη)$ можно выразить с помощью последовательной родо-видовой решетки.

3. Алгоритм определения роли понятий, находящихся в родовидовой связи

За исходные положения (аксиомы) принимаются определения понятий. Их качество напрямую влияет на результаты алгоритма [11]. Отношение *род-вид* реализуется в *родовидовом* определении понятия, т.е. в определяемое понятие выражается через понятие более общей категории. Каждое философское и научное направление вырабатывает и использует свой набор категорий, как предельно общих понятий. К общенаучным категориям относятся: предмет (entity), свойство (attribute), состояние (state), процесс (process), событие (event), отношение (relation).

Вид понятия (*родовое, видовое, межвидовое, собирательное*) определяется по отношению к другому понятию при наличии родовидовой связи между ними. Родовидовая связь между парой понятий (a, b) устанавливается путём сопоставления их содержаний $C(a)$ и $C(b)$ с применением теоретико-множественных операций. Для выполнения операций необходимо нормализовать по смыслу словосочетания, используемые в качестве существенных признаков (привести к «общему знаменателю»).

Определение вида понятия осуществляется в следующей последовательности.

1. Выявление родовидовой связи

Если $C(a) \cap C(b) = \varnothing$, то родовидовая связь ***отсутствует***.

2. При $C(d) = C(a) \cap C(b) \neq \varnothing$ определяется категория признаков, вошедших в $C(d)$.
3. Если нет признака $D_j \in C(d)$, более общего, чем остальные, то понятия a и b принадлежат *разным* предметным областям, а признаки $D_j \in C(d)$, $i, j = \overline{1,k}$,

представляют собой ***видовые отличия***. Имеет место сходство (аналогия) по видовым отличиям $D_j \in C(d)$, $i, j = \overline{1,k}$.

Пример. Понятия *производственный контроль* и *производственный брак* с общим видовым отличием *производственный* имеют отношение к производству, но принадлежат разным предметным областям.

4. Если есть признак $D_j \in C(d)$, более общий, чем остальные, то понятия a и b находятся в ***родовидовой связи***.
5. Если $C(a) = C(d)$, то понятия a и b находятся в отношении *род-вид*.
6. Если $C(b) = C(d)$, то понятия b и a находятся в отношении *род-вид*..
7. Если $C(a) \neq C(d)$ И $C(b) \neq C(d)$, то сравниваются видовые отличия, принадлежащие множествам $C(a)\backslash C(d)$ и $C(b)\backslash C(d)$.
8. Если $C(a)\backslash C(d) \cap C(b)\backslash C(d) = \varnothing$ и $|C(a)\backslash C(d)| = |C(b)\backslash C(d)| = 1$, то видовые отличия $D_i \in C(a)\backslash C(d)$ и $D_j \in C(b)\backslash C(d)$ образуют пару значений (D_i, D_j).
9. Если $D_i = D_j$, то понятия a и b ***равноценны***.
10. Если $D_i = D_j$ являются значениями *одного* основания деления (системообразующего признака), то оба видовых понятия группируются относительно этого основания деления, иначе они относятся к *разным* группам.
11. Если $C(a)\backslash C(d) \cap C(b)\backslash C(d) = \varnothing$ и $|C(a)\backslash C(d)| > |C(b)\backslash C(d)|$ (множество $C(a)\backslash C(d)$ содержит более одного видового отличия) и видовые отличия $D_i, D_j \in C(a)\backslash C(d)$ являются значениями *одного* основания деления, то понятие a является ***собирательным***, а если они относятся к разным основаниям деления, то понятие a является ***межвидовым***.

4. Относительность роли понятия

При рассмотрении связи между понятиями в паре (a, b) каждому из них присваивается своя роль (вид, род, часть, целое и т.п.). Следует подчеркнуть, что эта роль фиксирована только в пределах конкретного отношения. При смене отношения понятие приобретает одну из ролей, свойственную рассматриваемому типу отношения.

Для выявления роли понятия в системе родовидовых связей используются существенные признаки, извлекаемые из определений сопоставляемых понятий, и логические операции над ними. Выявление ролей части, агрегата и целого осуществляется с применением закономерностей иерархичности и целостности системного анализа [8]. Роль понятий в функциональных связях зависит от степени влияния на исследуемое понятие сопутствующих факторов, выявляемой с применением корреляционного анализа.

Покажем изменение ролей на примере уже приводившегося понятия *женщина*. В случае родовидовой связи, понятию *женщина*, как и мужчина, присваивается роль *видового* понятия с видовым отличием *пол* по отношению к родовому понятию *человек*. Понятия *женщина* и *мужчина* рассматриваются как *части* при объединении их в *агрегат*. Примером является их объединение в агрегат *покупатели* в магазине. Их цель — покупка товаров — не придаёт агрегату *покупатели* нового свойства. А вот семья, образованная женщиной и мужчиной, безусловно, обладает новыми свойствами, присущими *целому*. Несмотря на то, что в обоих примерах понятия *женщина* и *мужчина* играют роль *части*, небезразличным является способ их объединения, дающий в результате либо агрегат, либо целое. Роль *элемента* по отношению класса *женщина* выполняет индивид, обладающий признаками женского пола.

Выявление степени влияния женщины на мужчину в связях влияния требует анализа конкретных ситуаций, хотя часто проще всего принять её за преобладающий фактор, т.е. за причину, в связи *причина-следствие* согласно французской поговорке «ищите женщину» (фр. «Cherchez la femme»). Научный подход к определению её роли выполним с применением корреляционного анализа. Количественный анализ взаимовлияния

различных факторов позволяет заменить «очевидную» каузативную связь на одну из других видов связи: суммирующую или альтернативную. Обнаружение транзитивной связи требует анализа дополнительных (промежуточных) понятий.

Понятие *женщина* может принимать также одну из ролей в отношении *цель-средство*. Например, при формировании в 1960-м году первой группы по вычислительной технике в Ленинградском институте железнодорожного транспорта из студентов разных вузов с участием автора этих строк в первом варианте она состояла лишь из представителей мужского пола. Декан включил в неё трёх девушек для решения задачи «смягчения дисциплины в группе». Здесь женщина выступила в роли *средства* по отношению к частной цели дерева целей, изучаемого в системном анализе.

Заключение

Подробный разбор связей между понятиями и ролей, занимаемых в них понятиями, позволил предложить формализованный подход к установлению вида связей и роли в них понятий. Установление ролей пары сопоставляемых понятий в рамках родовидовой связи требует лингвистического анализа их определений с целью формирования совокупности существенных признаков. Предложен алгоритм установления ролей понятий, основанный на теоретико-множественных операциях над содержаниями понятий. Показано применение системных закономерностей иерархичности и целостности при установлении ролей понятий в агрегатных и инструментальных связях. Выявление вида функциональной связи требует определения степени влияния на исследуемое понятие помимо сопоставляемого понятия сопутствующих факторов. Эта задача решается с применением методов корреляционного анализа. Отмечается ситуативность роли понятия, меняющейся при его применении в рамках другого вида связи.

Исследования, выполненные по данной тематике, проводились при финансовой поддержке гранта РФФИ № 17-01-00139 в рамках бюджетной темы № 0073–2018–0003.

Литература

[1] Микони С.В. Дискретная математика для бакалавра: множества, отношения, функции, графы: Учебное пособие. СПб: «Лань». 2012.
[2] Анфилатов В.С., Емельянов А.Л., Кукушкин А.А. Системный анализ в управлении. М: Финансы и статистика. 2002.
[3] Микони С.В. О качестве онтологических моделей // Онтология проектирования. 2017. Т. 7, №3(25). С. 347-360.
[4] Семенов С.С. Оценка качества и технического уровня сложных систем. Практика применения экспертных оценок. М. ЛЕНАНД. 2015.
[5] Поспелов Д.А. Логико-лингвистические модели в системах управления. М.: Энергоиздат, 1981.
[6] Тарасенко, Ф.П. Прикладной системный анализ: учебное пособие. 2-е изд. перераб. и доп. М.: КНОРУС. 2017.
[7] Микони С.В. Общие диагностические базы знаний вычислительных систем. СПб: СПИИРАН, 1992. URL: http://www.mcd-svir.ru/books.html. (дата обращения: 06.04.2018).
[8] Теория систем и системный анализ в управлении организациями: справочник. Под редакцией В.Н. Волковой и А.А. Емельянова. М.: Финансы и статистика, 2009.
[9] Микони С.В. О классе, классификации и систематизации // Онтология проектирования. 2016. Т.6, №1(19). – С. 67-80.
[10] Mikoni S.V. Application of the Universal Decision Support System SVIR to Solving Urban Problems // DTGS 2016, CCIS 674. Springer International Publishing AG 2016. P. 1–14.

[11] Ackoff R.L. Differences That Make a Difference: An Annotated Glossary of Distinctions Important in Management. Devon. Triarchy Press, 2010.

A Formalized Approach to Establishing the Connection and Role of Concepts

S.V. Mikoni

Saint-Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences

The establishment of the type of connection in each pair of concepts and their role in this connection is in demand in the construction of ontological models for various purposes. In the absence of formal models, this problem is solved heuristically, or expertly. However, the practical knowledge of experts and their intuition do not guarantee correct decisions. Reducing the degree of subjectivism in solving the problem should be promoted by a higher degree of formalization of the connections between concepts.

The formalization of the connections and the role of the concepts to be compared in them is preceded by their exhaustive description and examples of application. Three groups of connections are considered: generic, aggregate and functional. For each of the groups it is proposed to use their own method of analysis. The logical-linguistic analysis used to analyze generic connection consists in extracting essential features from the definitions of the concepts compared and in set-theoretic operations on their contents and volumes. The method is implemented in the form of an algorithm for establishing the roles of concepts in generic connection.

To establish the role of concepts within the framework of aggregate and instrumental relationships, it is proposed to involve the regularities of system analysis. It shows the application of systemic regularities of hierarchy and integrity in establishing the roles of concepts in aggregate and instrumental connections. Varieties of functional connection differ in the degree of influence of the factors accompanying the concepts being compared. The degree of influence is proposed to be established by the methods of correlation analysis. The role of the concept is changeable. It changes with a change in the type of connection between concepts. The formalized approach outlined in the work is planned to be applied to the systematization of concepts used in the field of self-organizing systems.

Keywords: concept, relationship, relations, role, genus, species, part, aggregate, whole, generalization, aggregation, instrumental relations, influence relations

Программная реализация на базе платформы Apache Jena вопросно-ответной системы, использующей данные онтологий

А.В. Мочалова, В.А. Мочалов

Институт космофизических исследований и распространения радиоволн ДВО РАН

stark345@gmail.com, sensorlife@mail.ru

Аннотация

В настоящее время весьма актуальной задачей является разработка вопросно-ответных систем, позволяющих отвечать на вопросы пользователей, заданные на естественном языке по машиночитаемым текстам на естественном языке.
В работе описывается вопросно-ответная система, использующая данные из онтологии РуТез.
Предлагаются этапы решения задачи соотнесения частей текста с узлами онтологии. Всего выделяется 6 этапов: предварительная обработка текста; определение границ предложений; выделение границ синтаксем; определение возможных вариантов лемм для всех выделенных синтаксем; поиск в онтологии элементов, соответствующих начальным формам синтаксем; выбор из элементов онтологии, соответствующих синтаксемам.
Приводится описание архитектуры вопросно-ответной системы, основанной на использовании платформы Apache Jena, экспертной системы Drools и разработанного авторами семантического анализатора. Приводятся сквозные примеры работы системы.

Ключевые слова: вопросно-ответные системы, онтологии, автоматическая обработка текста, Drools, SPARQL, Apache Jena

1. Введение

1.1. Общая схема работы вопросно-ответной системы

На основе анализа существующих разработок вопросно-ответных систем, можно сделать вывод о том, что качественная система ответа на вопрос функционирует в соответствии с определенной схемой, представленной на рисунке 1 (подробный анализ вопросно-ответных систем проведен в работе [1]). На вход системе подается вопрос, сформулированный на естественном языке. Затем текст вопроса проходит автоматическую обработку, основные этапы которой следующие: предварительная обработка текста (включает в себя удаление лишних символов форматирования, исправление орфографических и пунктуационных ошибок, удаление лишних пробелов и символов переноса строк и т.п.); извлечение именованных сущностей; разбиение текста на предложения; токенизация (разбиение предложений на слова); морфологический, синтаксический и семантический анализ. Модуль автоматического анализа текста, как правило, использует различные структурированные лингвистические ресурсы: словари, базы данных, базы фактов, онтологии. В некоторых вопросно-ответных системах часть из вышеперечисленных этапов автоматической обработки текстов может быть пропущена или выполняться в упрощенном виде, а часть наоборот — являться основополагающими для

работы всей системы, как, например, семантический анализ в работе М.В. Мозгового [2]. Затем текст вопроса классифицируется в соответствии с принятой в данной системе классификацией. На базе результатов автоматической обработки текста вопроса и результатов классификации вопроса формируется запрос, который передается поисковой машине.

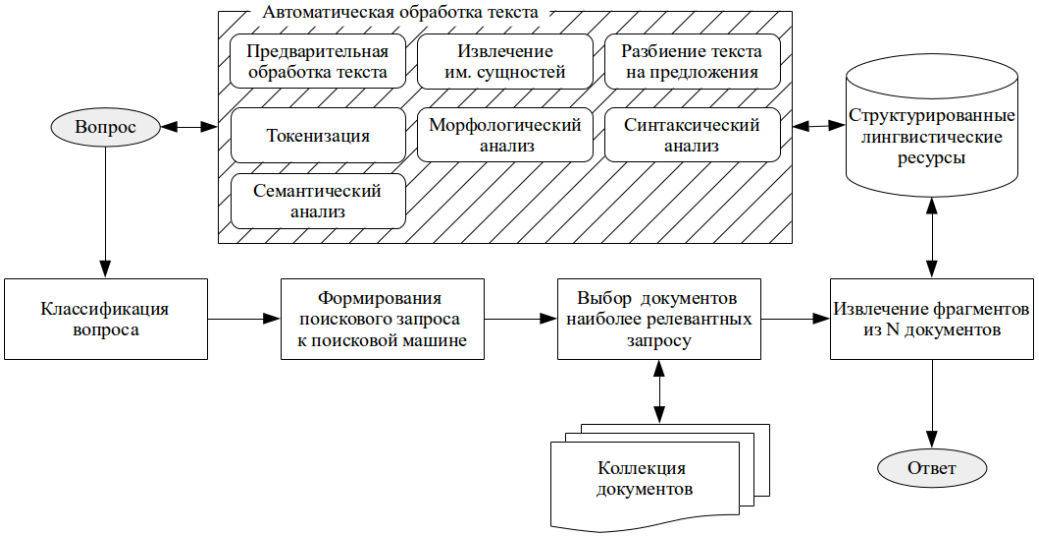

Рис. 1. Схема работы вопросно-ответной системы

Далее поисковая машина выбирает определенное количество документов (если поиск производится не по одному заданному документу), наиболее релевантных запросу. Выбор документов может производиться с помощью внешних поисковых систем, либо с помощью собственной поисковой машины, являющейся частью разрабатываемой системы. В некоторых случаях поиск документов производится в ограниченной специализированной коллекции документов, которой располагает система. Эффективный подход к организации поисковой системы предложен в работе [3], где предлагается архитектура субпоисковой системы, которая формирует собственную базу документов и собственный поисковый индекс, а для ускорения процесса сбора потенциально интересующих документов использует внешние поисковые системы (Google, Яндекс, Bing). На рисунке 2 представлена диаграмма соотношений множеств документов в такой субпоисковой системе (I — множество документов, доступных в сети интернет, W — множество документов, отобранных Интернет поисковой системой, S — множество документов, отобранных субпоисковой системой) [3].

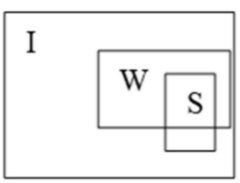

Рис. 2. Диаграмма соотношений множеств документов в субпоисковой системе

Текст каждого из выбранных документов, также как и текст вопроса, подвергается автоматической обработке. При этом алгоритмы машинной обработки текста вопроса могут отличаться от алгоритмов машинной обработки набора документов, выбранных поисковой

системой. Далее, посредством внутренних алгоритмов работы вопросно-ответной системы, происходит выбор конкретных фрагментов текстов из документов, переданных поисковой системой. Выбранные фрагменты текста представляются системой в качестве ответа. Наиболее продвинутые вопросно-ответные системы на этапе выбора фрагментов текста могут использовать данные из структурированных лингвистических ресурсов. Информация из этих лингвистических ресурсов может дополнять ответ/ответы системы.

В связи с тем, что результаты машинной обработки текста передаются модулям вопросно-ответной системы, от которых напрямую зависит ответ системы на вопрос, можно сделать вывод, что задача автоматической обработки текста является одной из важнейших задач, решаемых в рамках работы вопросно-ответной системы и от корректности работы модуля обработки текста напрямую зависит корректность работы всей системы.

1.2. Применение тезаурусов при разработке вопросно-ответных систем

Одно из пониманий тезауруса подразумевает словарь, с максимальной полнотой представляющий лексику языка во всей ее полноте с примерами употребления в текстах. Однако с точки зрения применения тезауруса в вопросно-ответной системе его следует понимать как информационно-поисковый тезаурус, как словарь общей или чаще специальной лексики, в котором в явном виде указаны семантические отношения между лексическими единицами (синонимия, антонимия, гипонимия, гиперонимия и т.п.). Многие прочие отношения часто объединяются в общий класс ассоциативных отношений. В отличие от толкового словаря, тезаурус позволяет выявлять смысл не только с помощью определения, но и посредством соотнесения слова с другими понятиями и их группами. Тезаурус — это терминологический ресурс, реализованный в виде словаря понятий и терминов со связями между ними. Основное его назначение — помощь при информационном поиске: на основе связей тезауруса происходит расширение запроса, навигация по связям тезауруса помогает четче сформулировать сам запрос.[4]

Качество работы вопросно-ответной системы напрямую зависит от качества и объема используемых тезаурусов. Использоваться они могут на разных уровнях реализации такой системы, например, при определении границ синтаксем, выделении именованных сущностей, в модуле, выполняющем семантический анализ текста, а также непосредственно в алгоритмах поиска ответа на вопрос в анализируемом тексте.

Следует отметить, что в широком понимании онтологии тезаурусы тоже являются онтологиями.

В этой работе описывается вопросно-ответная система, использующая данные из известного лингвистического ресурса — тезауруса RuTez [5], хранящего данные в структурированном виде. Описывается использование RuTez для выделения именованных сущностей в тексте, показывается как с помощью SPARQL-запросов и онтосемантического анализатора, используемого вопросно-ответной системой, формируется ответы на заданные пользователями вопросы.

2. Архитектура вопросно-ответной системы на базе платформы Apache Jena

В настоящее время весьма актуальной задачей является разработка вопросно-ответных систем, позволяющих отвечать на вопросы пользователей, заданные на естественном языке по машиночитаемым текстам на естественном языке. Ниже приводится описание архитектуры вопросно-ответной системы, основанной на базе платформы Apache Jena и использующей данные из онтологии [1].

На рисунке 3 приводится обобщенная архитектура вопросно-ответной системы [6], основанная на использовании семантического анализатора, построенного по математической модели, описанной в работе [7].

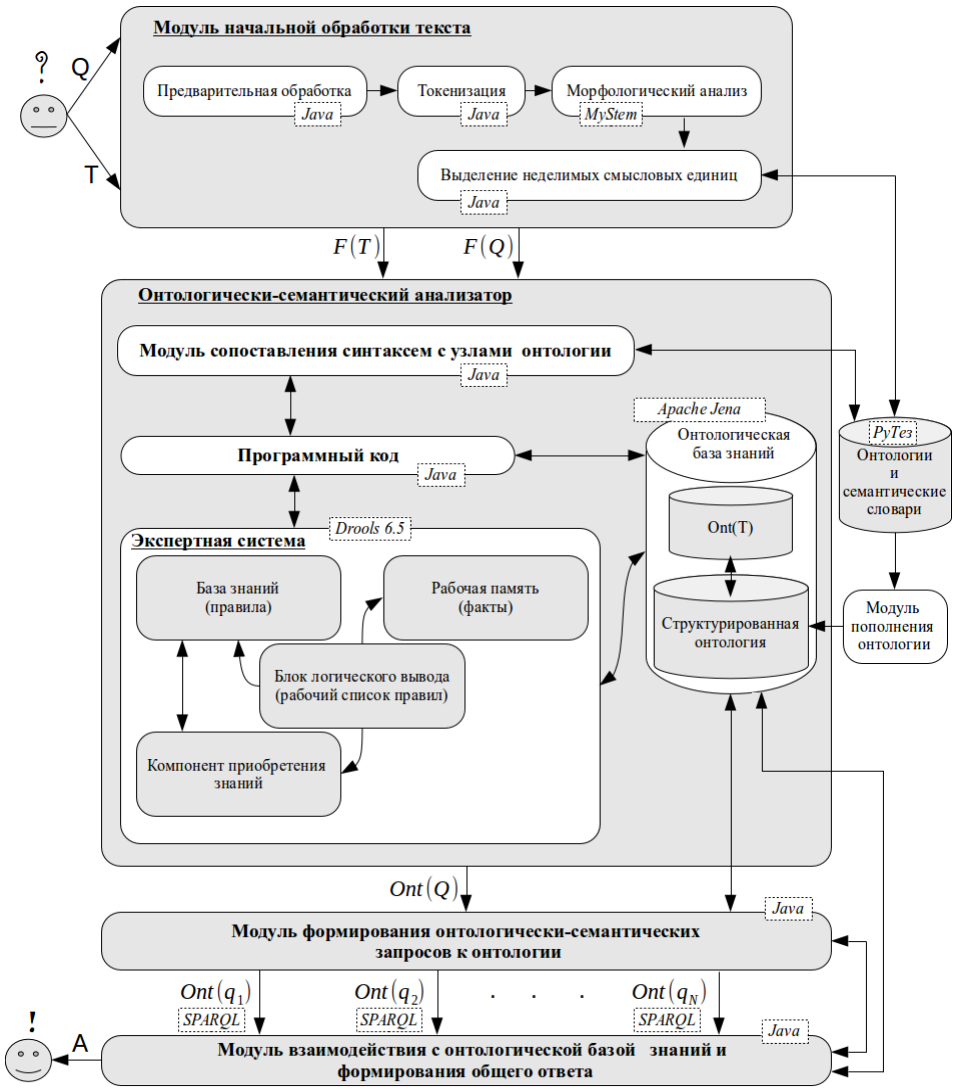

Рис. 3. Пример архитектуры вопросно-ответной системы

На вход вопросно-ответной системе подается вопрос Q на естественном языке и пользовательский текст T, выбранный пользователем для поиска ответа на вопрос Q. Текст T и вопрос Q поступают на вход модуля начальной обработки текста, в котором происходит выполнение следующих шагов: предварительная обработка, токенизация, морфологический анализ, выделение неделимых смысловых единиц. Результаты начальной обработки T и Q записываются в F(T) и F(Q) соответственно, после чего F(T) и F(Q) поступают на вход онтологически-семантическому анализатору, который на основе использования программного кода, экспертной системы, онтологической базы знаний выполняет следующие действия: сопоставление синтаксем с узлами структурированной онтологии; построение онтосемантических графов Ont(T) и Ont(Q), узлы которого ссылаются на элементы структурированной онтологии. Структурированная онтология формируется на базе загружаемых в систему онтологий и семантических словарей с помощью модуля пополнения онтологии. Далее Ont(Q) подается на вход модуля формирования онтологически-семантических запросов $Ont(q_i)$ к онтологии, которые

отправляются модулю взаимодействия с онтологией и формирования общего ответа. Этот модуль выполняет запросы к онтологической базе знаний и на базе полученных ответов формирует общий ответ A.

Пунктирной рамкой на рисунке 3 обведены названия языка программирования (Java), экспертной системы (Drools [8]) онтологии (РуТез [5]), семантической платформы (Apache Jena [9]) и языка запросов (SPARQL [10]), с помощью которых был программно реализован прототип такой системы.

Идея использования SPARQL-запросов при разработке вопросно-ответных систем не нова: пример такой системы описывается в работе [11].

3. Соотнесение частей текста с узлами онтологии

Необходимость определения соответствия частей текста и элементов онтологии возникает при решении целого ряда задач компьютерной лингвистики, связанных с автоматической обработкой текста (например, при реализации систем машинного перевода, автоматического аннотирования и реферирования, при разработке информационно-поисковых и вопросно-ответных систем, систем разметки корпусов текста и др.).

При решении задачи соотнесения частей текста с узлами онтологии можно выделить следующие этапы:

- (s1) Предварительная обработка текста;
- (s2) Определение границ предложений;
- (s3) Выделение границ синтаксем;
- (s4) Определение возможных вариантов лемм для всех выделенных синтаксем;
- (s5) Поиск в онтологии элементов, соответствующих леммам из (s4);
- (s6) Выбор из элементов онтологии, найденных в (s5), тех, которые соответствуют синтаксемам из (s3).

Первый этап — «Предварительная обработка текста» может включать такие действия по обработки естественно-языкового текста, представленного в электронном виде, как удаление символов форматирования текста, удаление лишних пробелов и переносов строк, исправление опечаток, правка всевозможных машинно-определяемых ошибок в написании оформлении текста.

Ниже приведены наиболее известные программные реализации, выполняющие некоторые задачи предварительной обработки текстов с указанием названия и вида лицензии для каждой из них или условий использования (приведено в скобках).

Проверка правописания:

- GNU Aspell (LGPL),
- Hunspell (GPL, LGPL, MPL),
- ОРФО Speller (Коммерческая),
- ОРФО Grammar Checker (Коммерческая).
- Проверка грамматики: LanguageTool (LGPL), Microsoft Word (Коммерческая).

Для решения второго этапа «Определение границ предложений» в сети Internet предлагается множество программных реализаций, выполняющих такую разбивку. Однако, в основе работы большинства таких программ лежит принцип определение конца предложения по терминальному знаку препинания (точка, вопросительный или восклицательный знак). Такой подход к решению задачи сегментации предложений привлекает своей простотой, но в реальной программной системе соотнесения частей текста с узлами онтологии использовать его нежелательно т.к. количество ошибочно найденных границ предложения при использовании описанного подхода, неоправданно велико.

В отечественной литературе проблема разбиения русскоязычного текста на предложения кратко освещается в работе [12]. В работе [13] предлагается метод автоматической сегментации русскоязычного текста на предложения на основе анализа контекста потенциальных границ предложений, при этом потенциальные границы

определяются либо посредством терминальных знаков, либо вообще с помощью всей пунктуация. При этом авторы не рассматривают предложения, не заканчивающиеся никаким знаком препинания.

Среди наиболее известных программных реализаций, выполняющих разбиение русскоязычного текста на предложения, можно выделить Aot (лицензия LGPL) — как часть графематического анализа и RCO (коммерческая лицензия).

Третий этап «Выделение границ синтаксем» — однозначно, самая сложная из задач, предшествующих непосредственно решению задачи соотнесения частей текста с узлами онтологии. Под синтаксемой будем понимать единицу текста, которая в работе [14] определяется как минимальная, далее неделимая семантико-синтаксическая единица русского языка, выступающая одновременно как носитель элементарного (категориально-семантического) смысла и как конструктивный элемент более сложных синтаксических построений. От того насколько корректно будут определены границы синтаксем в анализируемом тексте, напрямую зависит качество работы системы соотнесения частей этого текста с узлами онтологии.

Для примера рассмотрим предложение «В лесу у моря стоит замок» (см. рис. 4). Поставив задачей определить соответствие частей этого текста с элементами Wikidata — базы данных, которую также можно классифицировать как онтологию, столкнемся с трудностями, связанными с определением границ синтаксем. Например, в Wikidata присутствует как элемент «лес», имеющий несколько различных значений, так и элемент «В лесу», характеризуемого как «рассказ Бориса Александровича Лазаревского». Очевидно, что в зависимости от того как будут определены границы синтаксем в анализируемом предложении, напрямую зависит корректность соотнесения частей этого предложения с элементами онтологии.

После того, как в анализируемом тексте определены границы синтаксем, необходимо определить леммы (начальные формы) для всех синтаксем т.к. элементы онтологии обычно хранятся в начальной форме. Здесь начинается четвертый этап решения задачи соотнесения частей текста с элементами грамматический словарь Зализняка [15]. Для рассматриваемого в примере предложения «В лесу у моря стоит замок» с помощь словаря Зализняка для 3 синтаксем из 6 будут определены 2 леммы: слову «лесу» соответствуют леммы «леса» и «лес», слову «моря» — леммы «море» и «морить», слову «стоит» — леммы «стоить» и «стоять». Остальные синтаксемы анализируемого предложения употреблены в формах, совпадающих с их леммами (см. рис. 4).

Далее следует этап поиск в онтологии элементов, соответствующих леммам, найденным на предыдущем этапе. Для рассматриваемого в примере предложения для 7 лемм из 9 в онтологии РуТез будут найдены узлы с именами этих лемм. Ниже перечислены значения этих лемм (в соответствии с РуТез):

- Леса → {[рыболовная леса]};
- Лес → {[деловая древесина]; [лесной массив]; [лес (множество чего-н. поднятого)]};
- Море → {[водный объект]; [море (большое количество)]};
- Морить → {[травить отравой]; [морить (мучить, изнурять)]; [морение древесины]};
- Стоять → {[стоять (быть без движения)]; [стоять (бездействовать)]; [находиться, пребывать]; [стоять (сохраняться, не портиться)]; [стоять в вертикальном положении]};
- Стоить → {[подобать, надлежать, следовать]; [стоить, иметь цену]};
- Замок → {[замок для запирания]; [средневековый замок]}.

Для синтаксем «в» и «у» в онтологии РуТез не найдено элементов с названиями лемм этих синтаксем.

На завершающем этапе требуется из элементов онтологии, найденных для всех возможных лемм каждой синтаксемы, выбрать единственный. Для примера, предложения, рассматриваемого ранее, синтаксеме «лесу» должен быть поставлен в соответствие элемент онтологии «лесной массив», синтаксеме «моря» — элемент «водный объект», синтаксеме «стоит» — элемент «находиться, пребывать», и синтаксеме «замок» — элемент онтологии

«средневековый замок». На рисунке 4 выбранные элементы онтологии РуТез выделены серым цветом.

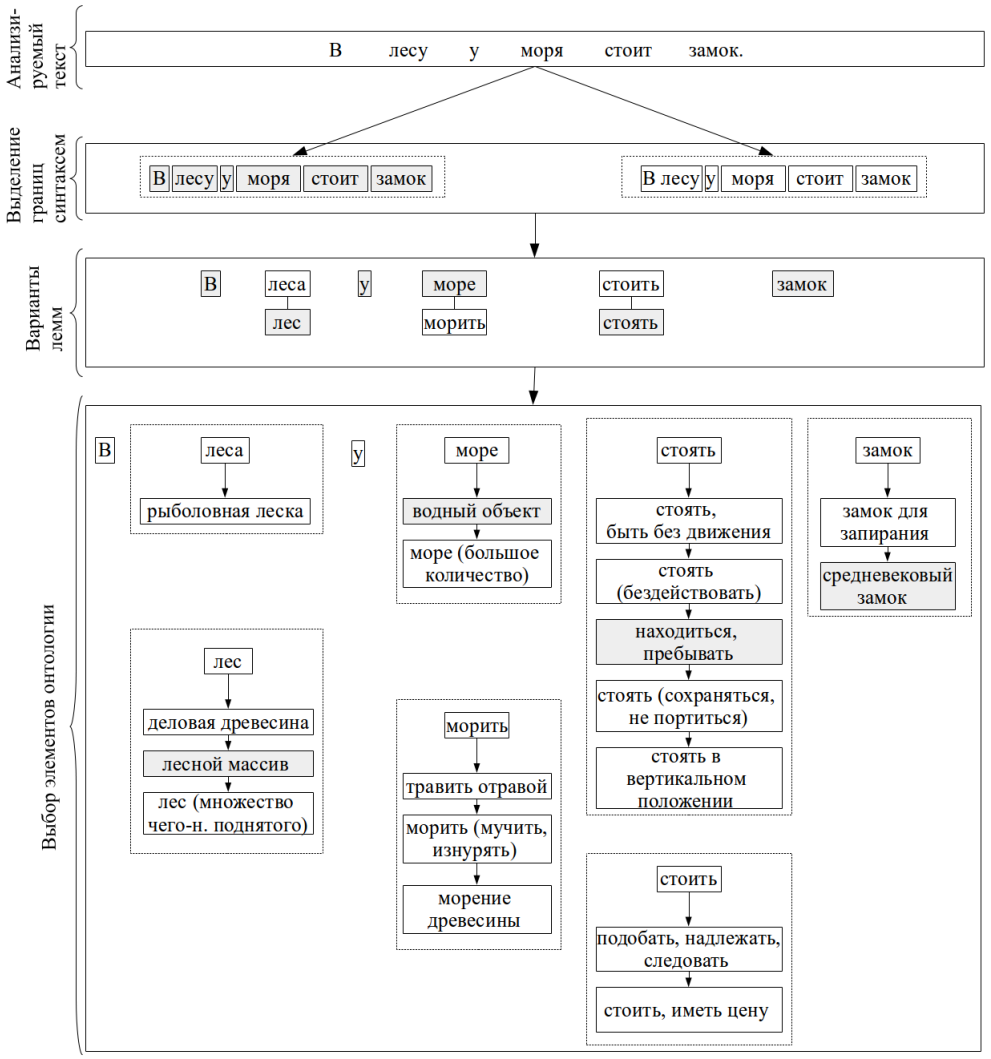

Рис. 4. Некоторые этапы решения задачи соотнесения частей текста с узлами онтологии

Краткий обзор методов и алгоритмов разрешения лексической многозначности приведен в работе [16]. Среди подходов к разрешению лексической многозначности выделяют методы, основанные на использовании внешних источников информации и методы, основанные на машинном обучении (обычно для этого используются семантически размеченные корпуса). Также применяются комбинации этих методов [17]. Автор работы [18] классифицирует методы разрешения лексической многозначности по типу используемых внешних источников информации:

- структурированные источники данных (машиночитаемые словари, тезаурусы, онтологии);
- неструктурированные источники данных в виде корпусов текстов делятся на:
  - неразмеченные корпуса;
  - синтаксически и/или семантически размеченные корпуса.

Одним из эффективных подходов к решению задачи соотнесения частей текста с узлами онтологии является использование правил, учитывающих контекст, в котором употреблена синтаксема, значение которой требуется определить, и информацию из онтологий — структурированных источников информации.

Будем предполагать, что этапы (s1)–(s6) уже выполнены и мы работаем с набором синтаксем, каждой из которых поставлено в соответствие множество элементов онтологии. Тогда задача разрешения лексической многозначности сводится к тому, чтобы из каждого соответствующего отдельной синтаксеме множества элементов онтологии выбрать один единственный, наилучшим образом отражающий лексическое значение рассматриваемой синтаксемы. Например, для предложения «В лесу у моря стоит замок», задача разрешения лексической многозначности сведется к выбору единственного верного значения из множества элементов РуТез (см. таблицу 1).

Таблица 1. Синтаксемы и соответствующие узлы онтологии РуТез

Название синтаксемы	**Узлы онтологии, из которых необходимо выбрать один соответствующий синтаксеме**	**Узел онтологии, соответствующий лексическому значению синтаксемы**
лесу	Рыболовная леска	Лесной массив
	Деловая древесина	
	Лесной массив	
	Лес (множество ч.-нибудь поднятого)	
море	Водный объект	Водный объект
	Море (большое количество)	
	Морить (травить отравой)	
	Морить (мучить, изнурять)	
	Морение древесины	
стоит	Стоять, быть без движения	Находиться, пребывать
	Стоять (бездействовать)	
	Находиться, пребывать	
	Стоять (сохраняться, не портиться)	
	Стоять в вертикальном положении	
	Подобать, надлежать, следовать	
	Стоить, иметь цену	
замок	Замок для запирания	Средневековый замок
	Средневековый замок	

При составлении правил, определяющих соответствие синтаксемы текста элементу онтологии, необходимо учитывать:

- Контекст синтаксемы (ближайший к синтаксеме текст имеет наибольшее значение: наиболее «важным» для анализа является текст предложения, в котором употребляется синтаксема, затем следует текст абзаца, содержащего это предложение, далее – раздел, содержащий упомянутое предложение, затем – раздел более высокого уровня (например, глава или параграф) и т. д., заканчивая всем анализируемым текстом).
- Семантическую близость синтаксемы, соотнесенную в процессе анализа с конкретным элементом онтологии, и синтаксем из контекста, учитывая «близость» контекста к анализируемой синтаксеме; для определения семантической близости можно использовать не только онтологию, но и ассоциативные словари.
- Тематику текста. Для определения тематики текста возможно либо попросить пользователя самого определить ее (например, предложив выбрать из списка), либо определить ее автоматически: в настоящее время существует множество алгоритмов для автоматической рубрикации текстов.

4. Примеры работы системы

4.1. Пример 1

Ниже приведен пример SPARQL-запроса, который формируется программно-реализованной вопросно-ответной системой для следующего вопроса «Какие существуют виды спорта?»:

SELECT DISTINCT ?x WHERE {?sub0 itfru:normalForm «спорт» .
?sub0 owl:sameAs ?samesub0 . ?x rdfs:subClassOf ?samesub0 .}

Приведенному SPARQL-запросу, адресованный онтологии РуТез, представленной в формате RDF, будет соответствовать ответ, состоящий из 133 элементов онтологии РуТез. Далее перечислены первые 10 из них: бег на длинную дистанцию, скелетон, прыжки в высоту, бобслейный спорт, баскетбол, легкоатлетический марафон, стендовая стрельба, горнолыжный супергигант, метание молота, хафпайп.

4.2. Пример 2

Исходный текст: Крупный, зеленый, добрый попугай съел кашу и яблоко, а воробей съел грушу

Найденные семантические отношения:

- ЧТО (съел#Г,изъяв,Г,мр,нс,прш,съедать,ед; кашу#С,С,но,каша,жр,вн,ед)
- ПРИЗНАК (попугай#С,С,попугай,мр,им,од,ед;
  Крупный#П,но,крупный,мр,им,вн,П,ед,полн)
- СПИСОК (зеленый#П,но,мр,им,вн,зеленый,П,ед,полн;
  добрый#П,но,мр,им,добрый,вн,П,ед,полн, ,#СИМВОЛ,- ,СИМВОЛ)
- ДЕЙСТВИЕ (съел#Г,изъяв,Г,мр,нс,прш,съедать,ед;
  воробей#С,С,мр,им,од,воробей,ед)
- ПРИЗНАК (попугай#С,С,попугай,мр,им,од,ед;
  добрый#П,но,мр,им,добрый,вн,П,ед,полн)
- СПИСОК (кашу#С,С,но,каша,жр,вн,ед; яблоко#С,С,но,им,яблоко,вн,ср,ед,
  и#СОЮЗ,СОЮЗ,и)
- ЧТО (съел#Г,изъяв,Г,мр,нс,прш,съедать,ед; грушу#С,С,но,груша,жр,вн,ед)
- ПРИЗНАК (попугай#С,С,попугай,мр,им,од,ед;
  зеленый#П,но,мр,им,вн,зеленый,П,ед,полн)
- СПИСОК (Крупный#П,но,крупный,мр,им,вн,П,ед,полн;
  зеленый#П,но,мр,им,вн,зеленый,П,ед,полн, ,#СИМВОЛ,- ,СИМВОЛ)
- ЧТО (съел#Г,изъяв,Г,мр,нс,прш,съедать,ед; яблоко#С,С,но,им,яблоко,вн,ср,ед)
- ДЕЙСТВИЕ (съел#Г,изъяв,Г,мр,нс,прш,съедать,ед; попугай#С,С,попугай,мр,им,од,ед)

Заданный вопрос: Какой фрукт съела большая птица?

Ответ: Яблоко

Сгенерированный SPARQL запрос:

SELECT DISTINCT ?x ?e WHERE { ?a itfru:normalForm "фрукт" . ?a owl:sameAs ?b . ?x rdfs:subClassOf ?b . ?x itfru:ns "NEW" . OPTIONAL { ?x itfru:ПРИЗНАК ?e } . ?x itfru:ЧТО_ИНВ ?x1 . ?x1 rdfs:subClassOf ?x1b . ?x1b owl:sameAs ?x1S . ?x1S itfru:normalForm "съедать" . ?x1 itfru:ДЕЙСТВИЕ ?x2 . ?x2 rdfs:subClassOf ?x2b . ?x2b owl:sameAs ?x2S . ?x2S itfru:normalForm "птица" . ?x2 itfru:ПРИЗНАК ?x3 . ?x3 rdfs:subClassOf ?x3b . ?x3b owl:sameAs ?x3S . ?x3S itfru:normalForm "большой" .}.

4.3. Пример 3

Исходный текст: Большое и красное яблоко

Найденные семантические отношения:

- ПРИЗНАК (яблоко#С,С,но,им,яблоко,вн,ср,ед;
  красное#П,красный,им,вн,П,ед,полн,ср)
- СПИСОК (Большое#П,им,большой,вн,П,ед,полн,ср;
  красное#П,красный,им,вн,П,ед,полн,ср, и#СОЮЗ,СОЮЗ,и)
- ПРИЗНАК (яблоко#С,С,но,им,яблоко,вн,ср,ед;
  Большое#П,им,большой,вн,П,ед,полн,ср)

Заданный вопрос: Какого размера яблоко ?

Ответ: большое

Сгенерированный SPARQL запрос:

SELECT DISTINCT ?x ?x1 WHERE { ?a itfru:normalForm "яблоко" . ?a owl:sameAs ?b . ?x rdfs:subClassOf ?b . ?x itfru:ns "NEW" . OPTIONAL { ?x itfru:ПРИЗНАК ?e } . ?x itfru:ПРИЗНАК ?x1 . ?x1 rdfs:subClassOf ?x1b . ?x1b owl:sameAs ?x1S . ?x1S itfru:normalForm "размер" . }

4.4. Пример 4

Исходный текст: Слоны обожают бананы, яблоки, морковь, свеклу

Найденные семантические отношения:

- СПИСОК (морковь#С,С,но,им,жр,вн,морковь,ед; свеклу#С,С,но,жр,вн,свекла,ед,
  ,#СИМВОЛ,- ,СИМВОЛ)
- СПИСОК (бананы#С,С,но,мн,банан,мр,им,вн; яблоки#С,С,но,мн,им,яблоко,вн,ср,
  ,#СИМВОЛ,- ,СИМВОЛ)
- СПИСОК (яблоки#С,С,но,мн,им,яблоко,вн,ср; морковь#С,С,но,им,жр,вн,морковь,ед,
  ,#СИМВОЛ,- ,СИМВОЛ)
- ДЕЙСТВИЕ (обожают#Г,изъяв,св,мн,буд,Г,пе,3л,обожать;
  бананы#С,С,но,мн,банан,мр,им,вн)
- ЧТО (обожают#Г,изъяв,св,мн,буд,Г,пе,3л,обожать; свеклу#С,С,но,жр,вн,свекла,ед)
- ЧТО (обожают#Г,изъяв,св,мн,буд,Г,пе,3л,обожать;
  бананы#С,С,но,мн,банан,мр,им,вн)
- ДЕЙСТВИЕ (обожают#Г,изъяв,св,мн,буд,Г,пе,3л,обожать;
  яблоки#С,С,но,мн,им,яблоко,вн,ср)
- ЧТО (обожают#Г,изъяв,св,мн,буд,Г,пе,3л,обожать;
  морковь#С,С,но,им,жр,вн,морковь,ед)
- ДЕЙСТВИЕ (обожают#Г,изъяв,св,мн,буд,Г,пе,3л,обожать;
  Слоны#С,С,мн,мр,им,слон,од)
- ДЕЙСТВИЕ (обожают#Г,изъяв,св,мн,буд,Г,пе,3л,обожать;
  свеклу#С,С,но,жр,вн,свекла,ед)
- ДЕЙСТВИЕ (обожают#Г,изъяв,св,мн,буд,Г,пе,3л,обожать;
  морковь#С,С,но,им,жр,вн,морковь,ед)
- ЧТО (обожают#Г,изъяв,св,мн,буд,Г,пе,3л,обожать;
  яблоки#С,С,но,мн,им,яблоко,вн,ср)

Заданный вопрос: Какие фрукты любят животные?

Ответ: банан, яблоко

Сгенерированный SPARQL запрос:

SELECT DISTINCT ?x ?e WHERE { ?a itfru:normalForm "фрукт" . ?a owl:sameAs ?b . ?x rdfs:subClassOf ?b . ?x itfru:ns "NEW" . OPTIONAL { ?x itfru:ПРИЗНАК ?e } . ?x itfru:ЧТО_ИНВ ?x1 . ?x1 rdfs:subClassOf ?x1b . ?x1b owl:sameAs ?x1S . ?x1S itfru:normalForm "любить" . ?x itfru:ДЕЙСТВИЕ_ИНВ ?x1 . ?x1 rdfs:subClassOf ?x1b . ?x1b owl:sameAs ?x1S . ?x1S itfru:normalForm "любить" . ?x1 itfru:ДЕЙСТВИЕ ?x2 . ?x2 rdfs:subClassOf ?x2b . ?x2b owl:sameAs ?x2S . ?x2S itfru:normalForm "животное" . }

5. Заключение

В настоящее время вопросно-ответная система, архитектура которой описывается в данной работе, находится в стадии разработки. В будущем планируется провести ряд работ, направленных на улучшение качества работы всех основных модулей системы (морфологический анализ, выделение границ именованных сущностей, соотнесение именованных сущностей с узлами онтологии, поиск семантических зависимостей, поиск ответа на вопрос пользователя). Также планируется провести полноценное тестирование системы, на больших объемах данных — например, на множестве статей с известных новостных сайтов. Подобный выбор тестовых данных весьма распространен среди разработчиков систем, нацеленных на извлечение информации из текстов. Например, в работе [19] авторы в качестве анализируемых корпусов текстов предлагают рассматривать статьи, опубликованные на известных новостных сайтах CNN и Daily Mail — этот набор данных стал стандартным для задач понимания текстов.

Литература

[1] Мочалова А.В. Семантический анализатор русскоязычного текста для вопросно-ответной системы: дис. ... канд. техн. наук: 05.13.18. Петрозаводск, 2017. 128 с.

[2] Мозговой М.В. Простая вопросно-ответная система на основе семантического анализатора русского языка // Вестник СПб университета. 2005. Сер. 10. Вып. 1. С. 116-122.

[3] Кулешов С.В. Вариант архитектуры субпоисковой системы для реализации функции аналитического мониторинга / С.В. Кулешов, С.Н. Михайлов // Труды СПИИРАН. 2013. № 8(31). С. 247-254.

[4] Захаров В.П., Мочалова А.В., Мочалов В.А. Вопросно-ответные системы. Некоторые проблемы автоматической обработки текста. Петрозаводск: ПИН, 2015.

[5] Лингвистическая онтология «Тезаурус РуТез» // URL: http://www.labinform.ru/pub/ruthes/index.htm (дата обращения 17.02.2018).

[6] Kuznetsov V.A. Ontological-semantic text analysis and the question answering system using data from ontology / Kuznetsov V.A., Mochalov V.A., Mochalova A.V. // ICACT Transactions on Advanced Communications Technology (TACT). 2015. Vol. 4, Issue 4. P. 651-658.

[7] Mochalova A.V., Mochalov V.A. Mathematical model of an ontological-semantic analyzer using basic ontological-semantic patterns // Lecture Notes in Artificial Intelligence, Proceedings of 15th Mexican International Conference on Artificial Intelligence. 2016. P. 53–66.

[8] Экспертная система Drools // URL: https://www.drools.org (дата обращения 17.02.2018).

[9] Apache Jena // URL: https://jena.apache.org/ (дата обращения 17.02.2018).

[10] SPARQL Query Language for RDF // URL: https://www.w3.org/TR/rdf-sparql-query/ (дата обращения 17.02.2018).

[11] Богуславский И.М. и др. Семантический анализ и ответы на вопросы: система в стадии разработки / И.М. Богуславский, В.Г. Диконов, Л.Л. Иомдин, А.В. Лазурский и др. // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Москва, 27–30 мая 2015 г.). Вып. 14 (21): В 2 т. М.: Изд-во РГГУ, 2015. Т. 1. С. 62 – 79.

[12] Автоматическая обработка текста // URL: http://www.aot.ru. (дата обращения 10.03.2018).

[13] Урюпина О. Автоматическое разбиение текста на предложения для русского языка // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог». Вып. 7 (14). М.: РГГУ, 2008.

[14]Золотова Г.А. Синтаксический словарь: Репертуар элементарных единиц русского синтаксиса. М.: Наука, 1988.
[15]Зализняк А.А. Грамматический словарь русского языка. Словоизменение. М.: Русский язык, 1977.
[16]Каушинис Т.В. и др. Обзор методов и алгоритмов разрешения лексической многозначности: Введение / Каушинис Т.В., Кириллов А.Н., Коржицкий Н.И., Крижановский А.А., Пилинович А.В., Сихонина И.А., Спиркова А.М., Старкова В.Г., Степкина Т.В., Ткач С.С., Чиркова Ю.В., Чухарев А.Л., Шорец Д.С., Янкевич Д.Ю., Ярышкина Е.А. // Труды КарНЦ РАН. № 10. Сер. Математическое моделирование и информационные технологии. 2015. С. 69-98
[17]Лукашевич Н. В. Тезаурусы в задачах информационного поиска. М.: МГУ, 2011. 495 с.
[18]Navigli R. Word sense disambiguation: A survey. // ACM Computing Surveys (CSUR). 2009. Vol. 41, № 2.
[19]Hermann K. M. and etc. Teaching machines to read and comprehend / Hermann K. M., Kocisky T., Grefenstette E., Espeholt L., Kay W., Suleyman M., and Blunsom P. // Advances in Neural Information Processing Systems. 2015. P. 1684–1692.

Software Implementation of Question-Answering System that Uses Ontology Data on the Basis of Apache Jena Framework

A.V. Mochalova, V.A. Mochalov

Institution Institute of Cosmophysical Research and Radio Wave Propagation FEB RAS

The development of question-answer systems that allow users to answer questions asked in natural language on machine-readable texts is very urgent task.
The paper describes a question-answer system that uses data from RuTez ontology. The paper describes the stages of solving the problem of mapping parts of text with ontology nodes. There are 6 stages in total: preliminary processing of the text; definition of the boundaries of sentences; allocation of syntaxemes boundaries; determination of possible lemmas variants for all allocated syntaxemes; search in the ontology for elements corresponding to initial forms of the syntaxemes; selection among the ontology elements corresponding to the syntaxemes.
The description of the question-answer system architecture based on Apache Jena, Drools expert system and semantic analyzer developed by authors is provided. End-to-end examples of the system operation are given.

Keywords: question-answer systems, ontologies, automatic text processing, Drools, SPARQL, Apache Jena

Методологические проблемы создания Компьютерной антологии русского рассказа как языкового ресурса для исследования языка и стиля русской художественной прозы в эпоху революционных перемен (первой трети XX века)

Г.Я. Мартыненко, Т.Ю. Шерстинова, А.Г. Мельник, Т.И. Попова

Санкт-Петербургский государственный университет

g.martynenko@spbu.ru, t.sherstinova@spbu.ru,
st064458@student.spbu.ru, tipopova13@gmail.com

Аннотация

В контексте социальных потрясений первой трети XX в. рассматриваются революционные изменения в русском языке и художественной литературе на материале русского рассказа. Материалом для такого рода исследований является Компьютерная антология русского рассказа, разрабатываемая на кафедре математической лингвистики Санкт-Петербургского университета.
Рассмотрены теоретические и прикладные предпосылки данной масштабной задачи. В основе проекта лежат идеи русской формальной школы и, прежде всего, системные идеи Ю.Н. Тынянова. Привлечены критерии, обеспечивающие представительность авторов и их произведений на основе объективных формальных процедур, в частности, размера статьи в Краткой литературной энциклопедии, посвященной конкретному беллетристу. В дальнейшем предполагается расширение хронологических рамок антологии с охватом всего XX в., а также конца XIX и начала XXI столетия.

Ключевые слова: русский язык, русская литература, рассказ, революционная эпоха, компьютерная антология, первая треть XX века

«В русской литературе рассказ традиционно был сильным жанром. Пожалуй, лишь американская литература приближается в этом отношении к нашей»

Юрий Нагибин. О рассказе [22]

1. Введение

История России XX в. драматична. Драматичны и исторические судьбы русского языка и русской литературы этой эпохи. Особенно богата переломными событиями первая треть столетия, вместившая в себя три социальные революции: революцию 1905 года, Февральскую и Октябрьскую революции 1917 года и три войны (русско-японскую, первую мировую и гражданскую).

Упомянутые события привели к многократной перестройке русского языка и его стилевых парадигм. Эти изменения носили преимущественно стихийный характер, но был и опыт вмешательства государства и активной части общества в процесс формирования языковой политики и построения новой революционной эстетики, дабы противопоставить волю революционных масс достижениям прежних эпох. Имеется в виду реформа русской графики и орфографии, эстетические эксперименты революционных направлений в искусстве, например, Пролеткульта.

Часть языковых сдвигов заметна «невооруженным глазом», другие языковые изменения имеют латентный характер. Однако для осознания масштабности и тех и других трансформаций языка необходимо применение строгих количественных методов, анализ представительного объема языкового материала на разных лингвистических уровнях и сравнение разных хронологических срезов в динамическом аспекте. Только в результате такого анализа можно будет с уверенностью сказать, в какой степени одна из самых драматических эпох русской истории повлияла на трансформацию русского языка, какие языковые уровни были затронуты в первую очередь, в чем конкретно выразилось это изменение и какова реальная суть произошедших преобразований.

2. Теоретические предпосылки

Теоретическими предпосылками создания Компьютерной антологии являются три источника: теория совокупности выдающегося русского статистика начала XX века А.А. Чупрова, статистические представления русского писателя и ученого А. Белого о массовом создании словарей русских писателей и системные представления выдающегося представителя русской формальной школы Ю.Н. Тынянова.

1) А.А. Чупров, основываясь на теории сообществ, выдвинул идею статистической совокупности, логика которой основана на собирательных понятиях [30]. Эта теория исходит из идеи сообщества Густава Рюмелина и идеи ценоза Карла Мёбиуса. Впоследствии логика Чупрова была применена к концепции корпуса [21].

2) Андрей Белый достиг важных качественных результатов, основываясь на массовом исследовании русской поэзии и открыв ряд принципиальных закономерностей в истории ритмики русского стиха. Впечатляющих результатов он достиг именно на акцентном уровне, так как только на нем можно было обеспечить значительную массовость наблюдения при состоянии скромных технологических средств начала прошлого века. На лексическом уровне А. Белому это не удалось, хотя он и поставил важную задачу «производственного» построения словарей русских писателей.

3) Ю.Н. Тынянов выдвинул концепции синхронических и диахронических литературно-художественных систем. Под синхроническими системами он понимал совокупность произведений данной литературной эпохи, а под диахроническими — последовательность сменяющих друг друга синхронических систем. При этом автор сетует на то, что усилия большей части литературоведов устремлены на изучение произведений выдающихся писателей, тогда как периферия литературы и даже ее «центр» остаются за бортом исследовательского интереса. Это означает, что для Тынянова крайне важным был вопрос максимальной представленности авторов в той или иной системе литературы и представительности корпуса текстов этих авторов. Любой текст Тынянов рассматривал как литературный факт, который должен приниматься во внимание независимо от масштабов дарования автора и его роли в литературном процессе.

Проблема системного анализа литературы тесно переплетается с проблемой возрождения и сохранения литературного наследия, существенная часть которого до недавнего времени была вычеркнута из памяти народа, ведь профессиональная текстовая рефлексия предполагает обращение к информационным ресурсам, несоизмеримым с теми, с помощью которых удовлетворяются интересы массового читателя. Для сообщества исследователей современной формации необходим доступ ко всей литературной продукции, созданной в ту или иную историко-литературную эпоху. Наш подход позволяет заполнить лакуны и области представленности литературной периферии.

Следует обратить особое внимание на некоторые успешные практические шаги в изучении русского новеллистического наследия. Пальму первенства следует отдать замечательной антологии писателя Юрия Нагибина [22], который в теперь уже далеком 1987 г., то есть еще при советской власти, но уже в период назревавших перемен, на страницах «Книжного обозрения» опубликовал свою антологию, отражающую развитие

советской новеллистики за семьдесят лет, истекших со дня Великой Октябрьской революции. В предисловии к антологии он пишет: «Мы чужды каких-либо научных, литературоведческих претензий и потому не стремимся охватить как можно больше имен, но по возможности — как можно больше хороших рассказов разных писателей. Таким образом, у нас нет намерения создать всеобъемлющую антологию, да это и не по плечу одному составителю. Быть может, эта несовершенная попытка подвигнет наших специалистов на создание научной антологии советского рассказа в целом» [16]. Предлагаемая нами антология является откликом на призыв выдающегося писателя. Это первая попытка создания всеобъемлющей научной антологии русского рассказа на объективной основе, восходящей к идеям Тынянова.

Остановимся на еще одном важном пункте. Интерес к жанру рассказа в писательской среде, возможно, и среди читателей в течение XX столетия, был непостоянным. В значительной степени он был связан с социальной активностью различных слоев общества в России на различных этапах социально-политического развития — с его подъемами и спадами, взрывами и апатией. Ведь, как отмечает Юрий Нагибин, «в жизни страны бывают разные периоды, которые можно сравнить с боевыми действиями во время войны: затишье, отступление, бои местного значения, наступление… А какое же наступление без атаки? На острие атаки должен находиться рассказ» [22, с. 3]. Все эти размашистые колебания отражаются на мощности информационного потока жанра короткой строки. Эти волны позволяют выявить с помощью рассказа всплески и спады социальной динамики. То есть рассказ выступает в роли диагноста социальных процессов. Эти два вектора должны постоянно находиться в активной зоне исследовательского интереса.

3. Технологические предпосылки

К настоящему времени накоплен значительный опыт в осмыслении специфики работы с большими массивами данных в рамках компьютерной, квантитативной и корпусной лингвистики. Особенно значительные успехи характерны для последнего времени и связаны они с развитием усилий в области обработки очень больших массивов данных — так называемых *big data*. Эти усилия влились в традиционные статистические проблемы формирования и анализа малых и больших, в традиционном смысле, выборок.

Для традиционной статистики малая выборка — это несколько десятков единиц, а большая — несколько сотен. Однако в подходе, основанном на big data, технические средства и программное обеспечение позволяют работать с практически неограниченными объемами данных. Да, собственно, и сам человек в наши дни живет среди этих данных и сам стал частицей киберпространства. Любопытно, что такие данные нашли применение не только при решении актуальных современных задач гуманитаристики, но и стали внедряться в способы решения проблем традиционного языкознания и литературоведения при изучении стиля, сюжета, жанра, эмоциональной напряженности текста, траекторий нарратива и пр. [35, 37]. При этом возникают проблемы соотношения малых, больших и сверхбольших выборок.

Проблема эта не проста, так как согласно устоявшимся статистическим представлениям чрезмерно большие выборки, формируясь на основе практически бесконечной генеральной совокупности, весьма неоднородной, стирают различия между неоднородными фрагментами. То, что при этом выявляются какие-то сверхобщие закономерности, находится под большим вопросом: разве можно найти какие-то закономерности в конгломерате, напоминающем хаос? Такое исследовательское поведение с точки зрения классической статистики является абсурдным. Тем не менее, данный подход имеет некоторое оправдание с точки зрения языковеда, поскольку при сверхбольших выборках вычерпываются ресурсы языка, а лингвиста волнует, прежде всего, это. Но если, лингвист — не только лингвист, но и словесник, то есть ориентирован на изучение жанрово-стилевой дифференциации литературного языка, то он не может

смириться с таким подходом. Большие числа ему тоже нужны, но только с учетом фактора неоднородности коммуникативного поведения человека. Литературоведа интересует, прежде всего, индивидуальное речевое поведение и поведение отдельных групп, разделяющих осознанно или инстинктивно свою принадлежность к группе.

При формировании Компьютерной антологии нужно иметь в виду, что такая работа предполагает крен в сторону литературоведения и той части лингвистики, которая использует лингвистические методы для решения проблем литературоведческих. Такое литературоведение ориентировано, прежде всего, в сторону той части поэтики, которая восходит к методам русской формальной школы XX в. [14, 28, 32]. Данный подход предполагает, прежде всего, ориентацию на исследование отдельных жанров и только затем — обращение ко всему жанровому разнообразию при учете самобытности каждого. В нашей концепции акцент сделан на жанре рассказа.

В пользу такого выбора говорит то, что рассказ выполняет функцию «разведчика» и жанра быстрого реагирования, отвечая на требования и вызовы эпохи и даже порой предвидя их. Следует также учитывать, что рассказ принадлежит к малым формам прозы, а это позволяет вовлечь в орбиту исследования большое количество прозаиков и их произведений. Важной особенностью рассказа является то, что он остро реагирует именно на современность, вбирая в себя характерные черты действительности и особенности языка современников. Причем этот жанр отражает все стороны реальной жизни, все ее богатство и огромное разнообразие индивидуальных стилевых систем.

При формировании антологии предполагается обеспечить достаточно полную представленность авторов с соблюдением принципа учета масштабов их роли в литературе на том или ином этапе эволюции жанра. В каждом периоде столетия должны себя обозначить и представители старшего поколения (архаисты), и молодые писатели. Часть из них следует в форватере прежних эпох, а другая проповедует новые тенденции (это новаторы). Необходимо также учитывать масштабы дарования каждого писателя и степень его влияния на коллег по писательскому цеху. Однако в любом случае нужно иметь в виду, что в пределах данного жанра художественная литература конкретного периода представляет собой монолитную целостность, спаянную «духом эпохи», единством общности социального, литературного и языкового существования.

Формирование списка авторов конкретной эпохи осуществляется на основе существующих библиографий, обзоров, энциклопедий, словарей писателей. Таких источников довольно много. Учитываются также интернет-коллекции и существующие корпуса. После составления списка авторов каталогизируются их произведения. При формировании выборки произведений будет использоваться критерий размера статьи, посвященной данному автору в Краткой литературной энциклопедии в 8 т. [18] и в Биографическом словаре русских писателей (1889–2007), а также объем списка литературы, относящейся к творчеству данного автора.

В работе используется широкий спектр филологической и исторической литературы, а также литературных обзоров и критических источников. Такие тексты можно разделить на две группы: лингвистические и литературоведческие. В пределах каждой группы рассматриваются труды, написанные в исследуемый период или «по свежим следам» [4-8, 11-13, 15, 17, 19, 23, 24, 27, 34, 36], а также публикации, подготовленные в более поздние годы вплоть до нашего времени [1-3, 10, 26]. При этом предпочтение отдается нами первой группе, так как она представляет собой непосредственную реакцию на процессы, протекающие на глазах у исследователя или критика, в том числе тех, которые принимают личное участие в этих процессах.

4. Заключение

В статье поставлена задача построения Компьютерной антологии русского рассказа первой трети XX века. Рассмотрены теоретические предпосылки и прикладные аспекты

этой масштабной задачи. В основе проекта лежат идеи русской формальной школы и прежде всего системные идеи Ю.Н. Тынянова, обеспечивающие представительность авторов и их произведений на основе объективных формальных критериев.

В дальнейшем предполагается расширение хронологических рамок антологии с охватом всего XX века а, возможно, и начала XXI столетия.

Работа выполнена при поддержке Российского фонда фундаментальных исследований, грант № 17-29-09173 офи_м «Русский язык на рубеже радикальных исторических перемен: исследование языка и стиля предреволюционной, революционной и постреволюционной художественной прозы методами математической и компьютерной лингвистики (на материале русского рассказа)».

Литература

[1] Аверин Б.В., Нитраур Э. (ред.) Русская литература XX века: Исследования американских ученых. СПб: Петрориф, 1993.
[2] Аверин Б.В. От Толстого до Набокова. Из истории русской литературы. Издательство имени Н. И. Новикова, Gallina Scripsit, 2014.
[3] Аверин Б.В., Козакевич А. Великая Война 1914–1918 гг. в поэзии и прозе. Антология / Сост., вступ. статья Б. В. Аверина. М.: Научно-образовательное культурологическое общество, 2014.
[4] Белый Андрей. Мастерство Гоголя: Исследование. М.-Л.: ГИХЛ, 1934.
[5] Верховской П.В. Письменная деловая речь. Словарь, синтаксис, стиль. Разбор бюрократических шаблонов и нарушений грамматики в языке документов. М.: Техника управления, 1930.
[6] Винокур Г.О. Речевая практика футуристов // Винокур Г. Культура языка. М.: Изд-во «Федерация», 1929. С. 304 – 318.
[7] Винокур Г.О. Язык НЭПа // Винокур Г. Культура языка. М.: Изд-во «Федерация», 1929. С. 115 – 139.
[8] Винокур Г.О. О революционной фразеологии (Один из вопросов языковой политики). // ЛЕФ. 1923. № 2. С. 104 – 118.
[9] Горнфельд А. Новые словечки и старые слова. М.: Колос, 1922.
[10] Гречнев В.Я. Русский рассказ конца XIX–XX века. Л.: Наука, 1979.
[11] Гринберг И. Широкое дыхание рассказа // Нева. 1968. № 8. С. 161 – 162.
[12] Гринберг И. Энергия рассказа // Нева. 1978. № 1. С. 191 – 192.
[13] Гусев Вл. «Левша» и стилевые тенденции в советской прозе // В мире Лескова: Сб. статей. М.: Советский писатель, 1983. С. 91 – 192.
[14] Жирмунский В.М. Вопросы теории литературы. Статьи 1916–1926. Л.: Academia, 1928.
[15] Карцевский С.И. Язык, война и революция. Берлин: Русское универсальное издательство, 1923.
[16] Книжное обозрение. № 7, 13 февраля 1987 г. С. 3.
[17] Крамов И. Достоинства рассказа // Дружба народов. 1977. № 8. С. 249 – 266.
[18] Краткая литературная энциклопедия. М.: Сов. Энцикл., 1962—1978. Т. 1–9.
[19] Мазон А. Лексика войны и революции в России (1914–1918). Введение. Аббревиация // Политическая лингвистика. 2013. № 1 (43). С. 203 – 210.
[20] Мартыненко Г.Я. Введение в теорию числовой гармонии текста. СПб: Изд-во Санкт-Петерб. ун-та, 2009.
[21] Мартыненко Г.Я. Основы стилеметрии. Л.: Изд-во ЛГУ, 1988.
[22] Нагибин Ю.М. (сост.). Антология русского советского рассказа. Предисловие Ю. Нагибина. Библиотечка «Книжного обозрения». М.: Книжное обозрение, 1987.
[23] Новиков Л.А. Стилистика орнаментальной прозы Андрея Белого. М.: Наука, 1990.

[24] Поливанов Е.Д. Революция и литературные языки Союза ССР // За марксистское языкознание. М.: Федерация, 1931. С. 73 – 94.
[25] Русские писатели 1800–1917. Биографический словарь. М.: Большая российская энциклопедия, 2007. Т. 1–5. 1989–2007.
[26] Русский советский рассказ: очерки истории жанра / под ред. В.А. Ковалева, Академия наук СССР, Институт русской литературы, 1970.
[27] Селищев А. Язык революционной эпохи. Из наблюдений над русским языком (1917–1926). М.: Работник просвещения, 1928.
[28] Томашевский Б. Теория литературы: Поэтика. М.–Л.: ГИЗ, 1928.
[29] Тынянов Ю.Н. Архаиста и новаторы. Л.: Прибой, 1929.
[30] Чуковский К.И. От Чехова до наших дней. Литературные портреты. Характеристики. СПб-М.: Издание Т-ва Вольф, 1910.
[31] Чупров А.А. Очерки по теории статистики. СПб: М. и С. Сабашниковы, 1910.
[32] Шкловский Б.В. Теория прозы. М.: Федерация, 1929.
[33] Шор Р. Язык и общество. М.: Работник просвещения, 1926.
[34] Эйхенбаум Б. О прозе. О поэзии. Сб. статей. Л.: Художественная литература, 1986.
[35] CLiC Dickens project. URL: http://clic.bham.ac.uk (дата обращения: 23.04.2018).
[36] Jakobson R. Vliv revoluce na ruský jazyk. (Влияние революции на русский язык). Praha, 1921.
[37] Reagan A.J., Mitchell L., Kiley D., Danforth C.M., Dodds P.S. The emotional arcs of stories are dominated by six basic shapes. arXiv: 1606.07772 [cs.CL] https://arxiv.org/pdf/1606.07772.pdf (дата обращения: 23.04.2018).

Methodological Issues Related with the Compilation of Digital Anthology of Russian Short Stories (the First Third of the 20th Century)

G.Y. Martynenko, T.Y. Sherstinova, A.G. Melnik, T.I. Popova

Saint-Petersburg State University

In the context of social upheavals of the first third of the 20th century, the revolutionary changes in the Russian language and fiction are to be studied on the basis of Russian short stories. The material for this kind of research is Digital Anthology of Russian short stories, which is currently being developed at the Department of Mathematical Linguistics of the St. Petersburg University. The theoretical basis and applications of this large-scale task are considered. The project is based on the ideas of the Russian formal school and, first of all, on the systemic ideas by Yury N. Tynyanov. Special criteria are developed to ensure the representativeness of authors and their works on the basis of objective formal procedures (e.g., the size of the article in the Concise Literary Encyclopedia devoted to a specific writer). In the future, an extension of the chronological framework of the Anthology with coverage of the entire 20th century, as well as the end of the nineteenth and the beginning of the 21st century, is proposed.

Keywords: Russian language, Russian literature, short story, revolutionary era, computer anthology, first third of the 20th century

Автоматизированный синтез когнитивной модели на основе анализа больших данных и глубокого обучения

А.Н. Райков

Институт проблем управления РАН, Институт философии РАН

alexander.n.raikov@gmail.com

Аннотация

Когнитивное моделирование позволяет учитывать одновременно денотативные (формализованные, вещные) и сигнификативные (когнитивные, мыслительные, эмоциональные) семантики. Такое моделирование позволяет уникальным образом, достаточно быстро и всесторонне отразить особенности возникшей проблемы, поддержать принятие правильного решения. Для обеспечения целостности когнитивной модели, учета наиболее полного объема информации, ускорения моделирования и роста его адекватности в настоящей работе предлагается проводить автоматическую верификацию (проверку) когнитивной модели на основе анализа Больших Данных. Также рассматривается возможность осуществлять автоматизированный синтез когнитивных моделей. В работе применяются методы глубокого обучения, теории категорий, а также авторский конвергентный подход к управлению и поддержке решений. Экспериментально показаны достаточно высокие показатели качества глубокого обучения и точности автоматического распознавания элементов когнитивных моделей (93%), что может служить весомой гарантией плодотворности предлагаемых решений. Предлагаются направления перспективных исследований, в частности, охватывающие методы квантовых семантик.

Ключевые слова: большие данные, глубокое обучение, квантовая семантика, когнитивное моделирование, конвергентный подход

1. Введение

Как правило, когнитивная модель неотчуждаема от построившей ее группы людей (экспертов, команды, менеджеров). Такая модель позволяет команде быстро, уникальным образом и всесторонне отразить особенности неожиданно возникшей или давно созревшей проблемы, удобна для поддержки политических решений, оценки целесообразности запуска инновационных проектов, формирования идей и замыслов.

Когнитивное моделирование помогает ответить на вопросы типа «Почему?», «Что будет, если …?» и «Что надо сделать, чтобы ...?» при большой неопределенности ситуации. Если для одного типа вопросов необходимо решение прямой задачи, то для более сложных случаев — обратной. Прямая задача решается обычно синхронным суммированием значений импульсов в узлах когнитивного графа. Для решения обратной задачи, как показано в работе [1], удобно применять эволюционные вычисления, например, генетический алгоритм. Он помогает найти несколько адекватных локальных решений, из которых лидер команды выбирает наиболее подходящее с учетом его личной осведомленности и интересов.

Когнитивное моделирование характеризуется оперативностью. Например, модель можно построить за время проведения совещания, в том числе сетевого, за 2–3 часа. Вместе с тем это моделирование сопровождается риском упущения отдельных параметров

ситуации (факторов и их взаимосвязей), и, следовательно, риском получения неадекватной модели и решения. Эти риски порождаются возможным отсутствием у участников моделирования всей необходимой информации, высокими требованиями к скорости принятия решения, предвзятостью отдельных участников принятия решения и пр.

В работе [2] для учета наиболее полного объема информации, а также одновременно ускорения моделирования и роста адекватности моделей, предложено проводить автоматическую верификацию (проверку) когнитивной модели на специально подобранном массиве документов с применением методов анализа больших данных. Показана плодотворность подхода, в том числе для очистки и формирования больших данных под определённую тематику.

Вместе с тем всегда остается открытым вопрос учета когнитивных семантик, поскольку они характеризуются принципиальной неформализуемостью и латентностью. Вопрос оперативности также нуждается в дополнительных решениях, что, по-видимому, лежит в плоскости автоматизации синтеза когнитивных моделей, в том числе, с учетом когнитивных семантик. В настоящей работе рассматривается подход к получению ответов на эти вопросы на основе применения методов глубокого обучения [3], а в перспективе — квантовых семантик [4].

2. Когнитивная поддержка решения

Когнитивное моделирование является частью группового процесса обсуждения и принятия решения возникшей проблемы. Весь процесс принятия решения может быть проиллюстрирован рисунком 1. Подразумевается наличие модератора-аналитика и экспертов (участников). Первый формирует запрос в виде открытой анкеты участникам, которые отвечают на вопросы, дают комментарии и формируют балльные оценки по заданным шкалам. Ответы используются для построения дерева целей и собственно когнитивной модели. На когнитивной модели решается прямая и обратная задачи, выбирается подходящий сценарий и план действий.

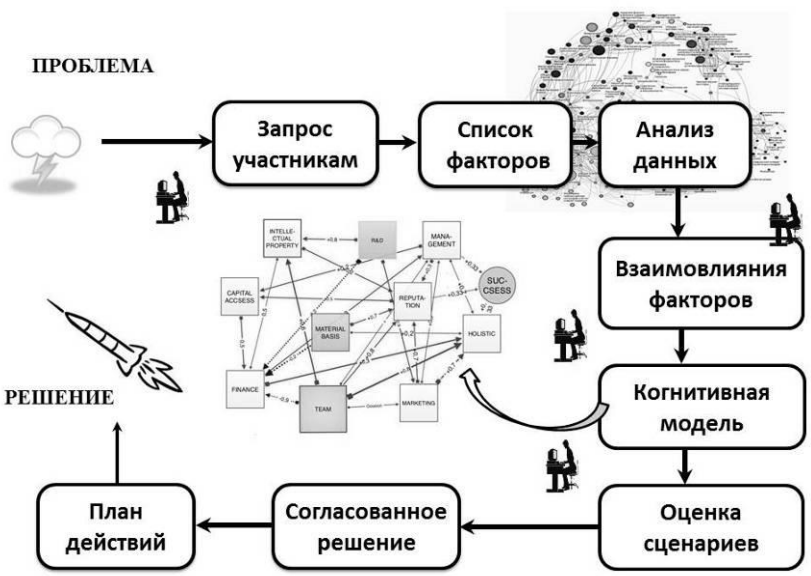

Рис. 1. Процесс принятия решения с когнитивным моделированием

Участникам может предоставляться дополнительная информация, справочные данные или даже результат другого моделирования, например, с помощью эконометрического подхода: модели общего равновесия, гравитационной модели и пр. Для анализа участникам может также подаваться некоторый поток данных, который необходимо читать и осмысливать. Если ситуация требует срочных и неотложных мер, а этот поток достаточно интенсивный, то для его качественной обработки и учета множества нюансов ситуаций у экспертов может не хватить времени.

В настоящей работе выявление факторов из потока данных для построения адекватной когнитивной модели предложено осуществлять на основе использования методов глубокого обучения. Обоснование решения проведено в следующем порядке:

- определены особенности сигнификативных семантик для обоснования возможности их косвенного автоматизированного учета при семантической интерпретации когнитивной модели;
- обоснован выбор метода глубокого обучения нейронной сети, создан тестовый корпус релевантных документов и макет программной среды для ее обучения;
- экспертно построена когнитивная модель для поддержки решения определенной проблемы (например, авария на железной дороге, приоритизация экспортной деятельности, развитие туризма в мегаполисе);
- разработан алгоритм по выделению из текстов потока документов (сообщений) факторов и взаимосвязей между ними для построения когнитивной модели;
- проведено обучение нейронной сети под факторы когнитивной модели на основе анализа некоторой части выбранного корпуса документов;
- проведена апробация возможности распознавания факторов, то есть синтеза фрагментов когнитивной модели, на основе другой части корпуса документов.

В основу реализации положены следующие подходы и методы: когнитивная и рефлексивная психология, сетевые технологии экспертизы, семиотика, искусственный интеллект, когнитивное моделирование, глубокое обучение, генетические алгоритмы, решение обратных задач на неметрических пространствах, управляемая термодинамика, конвергентный подход, квантовая семантика, теория категорий.

3. Когнитивная семантика

Большинство семантических школ исходят из того, что в содержании текстов изучается и описывается только то, что представлено текстом или другими формализованно-лингвистическими конструкциями. Например, значение любого элемента когнитивной модели может определяться через его отображение на подмножества больших данных. При этом слабо учитывается ментальность, исследования которой опираются на различные подходы к изучению мышления, вплоть до квантово-механических [4]. Это уже пространство сигнификативных (когнитивных) семантик.

Для лучшего автоматизированного учета при когнитивном моделировании неформализуемого ментального дискурса в настоящей работе использованы понятия теории категорий и монады. Теория категорий помогает изучать свойства отношений объектов без учета их структур. Ментальные же интерпретации имеют очень сложную и даже пока непознанную, а может и непознаваемую конструкцию, вот почему для ее исследования использована именно теория категорий. На рисунке 2 показана иллюстрация семантик текстовых (знаковых) элементов когнитивной модели.

На рисунке 2 когнитивной семантике соответствует категория C, в которой:

- под объектом A понимается замкнутый мыслительный феномен, соответствующий фрагменту когнитивной модели. Все мыслительные феномены, включаемые в дискурс, формируемый коммуникативной ситуацией, образуют класс объектов;

– множество морфизмов $Hom_C(A, B)$ формируется для каждой пары объектов A и B. Для пары объектов может быть сопоставлено множество отношений, характеризующих различие понимания участниками проблемного события и слов в текстах;
– для пары морфизмов $f \in Hom_C(A, B)$ и $g \in Hom_C(A, C)$ определяется композиция $g \circ f \in Hom_C(A, C)$. Вместе с тем такие конструкции существуют не всегда, например, возможны и нетранзитивные отношения между объектами;
– выполняется аксиома ассоциативности для морфизмов;
– для объекта A может быть задан тождественный морфизм $id_A \in Hom_C(A, A)$ с целью обеспечения замкнутости мыслительного феномена на себя.

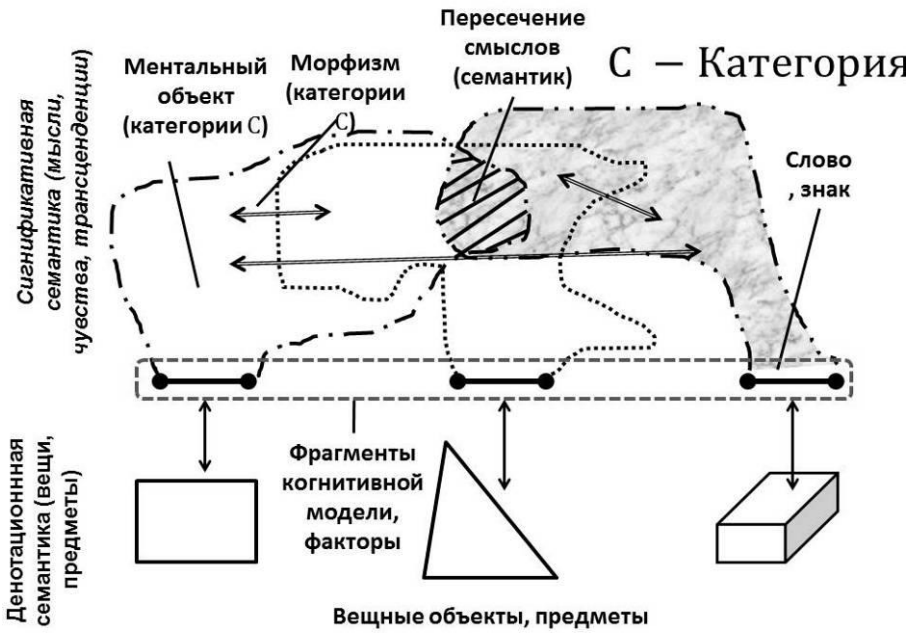

Рис. 2. Денотативные и сигнификативные (когнитивные) семантики

Опыт использования теории категорий в областях, близких к рассматриваемой в настоящей статье, уже имеется. Например, в [5] исследуются дистрибутивные категориальные модели языка, показывается ограниченность использования механизма статистической оценки векторных пространств, в том числе с квантовыми вычислительными приложениями.

С целью ускорения сборки коллективной мысли, создания необходимых условий для инкапсуляции (схватывания) целостного явления потребовалось использовать понятие *монады* $\mathcal{E}$. Однако для ускорения построения когнитивной модели свойств обычной монады недостаточно, поскольку требуется обеспечение необходимых условий для конвергенции (устойчивой сходимости) коммуникационного процесса коллективного принятия решения.

Для обеспечения конвергентности в настоящей работе введено понятие «Конвергентной монады», для чего к аксиомам классической монады добавлены следующие условия, которые являются необходимыми (но недостаточными) для обеспечения сходимости процессов коллективного согласования решений:

– D: **Set** → **Set**, где число элементов в системе множеств **Set** бесконечно, а графики отображений объектов замкнуты (замкнутость);

– $\mathcal{B}$ — непустое конечное подпокрытие монады $\mathcal{E}$ (бикомпактность);
– каждой точке монады $e \in \mathcal{E}$ может быть сопоставлена некоторая окрестность (всякое открытое множество, содержащее эту точку), такая, что для каждых двух точек всегда существует их непересекающиеся окрестности (хаусдорфова отделимость).

Такая, конвергентная, структуризация интерпретационного множества данных помогла обеспечить корректный выбор способа обучения нейронной сети, а также автоматического синтеза факторов модели под новую проблемную ситуацию.

4. Глубокое обучение и автоматизированный синтез факторов

При выборе типа модели нейронной сети рассматривались методы: Долгой краткосрочной памяти (LSTM), рекуррентный, сверточный. Акцент сделан на LSTM. Это искусственная нейронная сеть, содержащая модули, являющиеся рекуррентными модулями сети, способными запоминать значения, как на короткие, так и на длинные промежутки времени. Ключом к данной возможности является то, что LSTM-модуль не использует функцию активации внутри своих компонентов. Таким образом, хранимое значение не размывается во времени, и градиент или штраф не исчезает при использовании метода обратного распространения ошибки во времени при тренировке сети.

В отличие от традиционных рекуррентных нейросетей LSTM-сеть приспособлена к обучению на задачах классификации, обработки и прогнозирования временных рядов в случаях, когда допускается разделение событий временными лагами с неопределённой продолжительностью и границами. Для векторизации данных в настоящей работе использована модель Doc2Vec (Distributed memory model, Paragraph vectors [6]). Преимуществом этой модели является то, что она может хорошо работать для задач, в которых недостаточно помеченных данных.

В настоящей работе приняты следующие стадии создания нейронной сети:
− выбор проблемы и построение под нее когнитивной модели;
− подбор тестового корпуса релевантных для когнитивной модели документов;
− распределение подмножеств документов по факторам модели;
− токенизация, лемматизация и фильтрация текстов документов;
− тренировка модели Doc2Vec и ее конвертирование;
− обучение LSTM-модели;
− распознавание факторов под другой набор документов, порождаемых новой проблемой.

Исследование основывалось на наборе факторов для когнитивной модели, характеризующей туристскую деятельность в городе Москве. Всего в модели определено 19 факторов, включая целевой, например, факторы: «Рост доходности от туристского потока», «Рост числа туристов», «Рост времени пребывания туристов» и др.

Тестовый корпус релевантных документов построен на основе доступных в Интернет массивов данных. В результате исследования различных источников взяты данные с сетевого ресурса https://webhose.io/. Это набор новостей с классификацией тематических выборок. Размер корпуса документов, на базе которого обучена модель: 16227 текстовых документов. Для обучения нейронной сети сформирована выборка из 4868 документов. Сравнивались также результаты обучения на 10, 100, 1000 документов. При этом для обучения нейронной сети бралась 70% часть выборки, а 30% — использовалась для тестирования, проверки аккуратности обучения. На рисунке 3 показаны два распределения вероятностей интерпретирующих документов по трем факторам.

В ходе тестирования выявлено, что обученной нейронной сетью документы из тестирующего пакета данных идентифицируются по заданным трем факторам (категориями) с вероятностью 93%. Это определяет достаточно высокое качество

распознавания факторов и позволяет утверждать возможность и целесообразность предварительного автоматического синтеза новой когнитивной модели на основе потока данных под новую проблемную ситуацию.

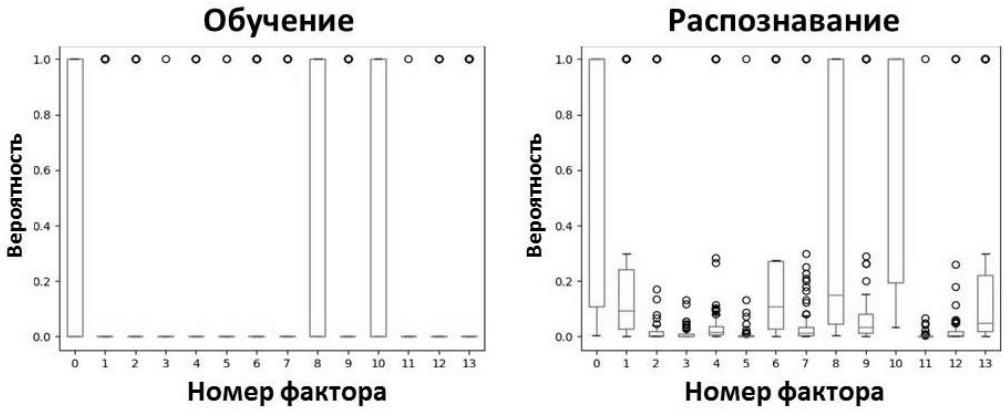

Рис. 3. Распределения вероятностей интерпретирующих документов по трем факторам

Общая схема, отражающая порядок глубокого обучения и распознавания факторов для автоматизации построения когнитивной модели, приведен на рисунке 4.

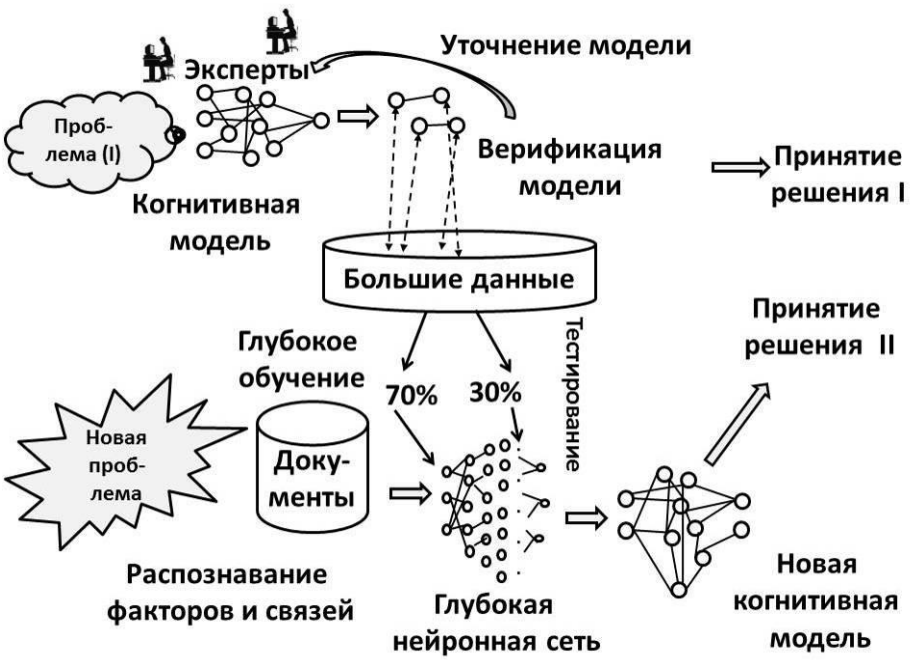

Рис. 4. Общая схема «Обучение — Распознавание»

5. Планируемые перспективные исследования

Основы классический подхода к моделированию, который сложился в прошлом веке, используют идеи теории искусственного интеллекта. Акцент в таком подходе пока в основном делался на формализованном представлении знаний об объекте управления и способах управления на уровне логико-лингвистичечких моделей, использовании дедуктивного и индуктивного вывода для построения многошаговых решений с выбором из множества альтернатив.

Вместе с тем текущие реалии, развитие цифровой экономики и методов искусственного интеллекта порождают вызовы, подвергающие сомнению возможности классического подхода. Такими вызовами становятся:

- потребность учета некаузального субъективного фактора при осуществлении коллективного обсуждения проблем и принятия решения;
- необходимость формирования безальтернативного решения проблем без применения традиционного многокритериального выбора с акцентом на привлечение методов решения обратных задач на концептуальных (когнитивных) пространствах;
- задачи анализа уступают место задачам синтеза. Причем, если первые носят дивергентный, расходящийся, характер, то вторые должны носить характер конвергентный (сходящийся);
- потребность в использовании схем и технологий поддержки когнитивных и бессознательных аспектов в процессах мышления и самоорганизации групп людей.

Таким образом, дальнейшее развитие затронутой в настоящей работе тематики связано с постановкой и решением широкого спектра междисциплинарных проблем в полисубъективной среде.

Учитывая потребность более полного учета явно неподдающейся формализации сигнификативной (когнитивной, ментальной, мыслительной) семантики, в настоящей работе сделаны сравнительные оценки отдельных параметров объемных характеристик мозга человека. Так, если число нейронов в мозге, как известно, порядка 10^{11}, то число атомов, из которых состоят эти нейроны и их окружение — порядка 10^{26}. Допускается гипотеза о том, что мыслительная деятельность связана с наличием более сложных структур, чем нейроны, а именно — микротрубками и когерентными скоплениями, воздействием полей (электромагнитных, гравитационных, сильных и слабых), сцепленностью (запутанность [7], квантовая нелокальность, entanglement) с атомарными элементами вещества из внешнего, в том числе — дальнего, окружения [4].

При квантово-механических интерпретациях сигнификативных семантик атомы предопределят возможность долговременного хранения знаний, а фотонные структуры и полевые эффекты позволят учесть в процессах мышления удаленные на большие расстояния элементы различной природы. Тогда операционное обеспечение процессов мышления и принятия решений можно будет представить в виде поведения составных квантовых систем. Например, сигнификативной семантике когнитивной модели может быть сопоставлена составная квантовая система в состоянии с полным спином, равным нулю. Отклонение от такого состояния будет характеризовать возможность снижения или повышения устойчивости и целенаправленности коллективного процесса обсуждения проблемы и принятия решения.

Таким образом, мыслительная деятельность человека может представляться не только логикой и динамикой взаимосвязи нейронов. Ее можно также ассоциировать с функционированием особым образом настроенной антенной решетки, состоящей из вибраторов, волновых щелей, излучателей и пр. Известно также, что квантовое волновое поле синтезирует понятия электромагнитного поля и вероятностное поле квантовой механики. Оно является наиболее фундаментальным обобщением теории поля. Возможно, при сигнификативной семантической интерпретации знаков когнитивной модели стоит учитывать корпускулярно-волновой характер процессов мышления. Возможно, такая

ассоциация поможет учесть и впоследствии определить сигнификативный аспект семантики через оценку воздействия внешних полей на нейронные процессы и мышление.

Такая ассоциация уже позволяет сделать вывод, что для учета сигнификативной семантики знаков (высказываний, слов, факторов, связей в когнитивной модели) необходимо использовать интерпретирующие множества мощностью (кардинальное число) на десятки порядков (!) выше, чем мощность множеств, отражающих денотативные семантики. Таким образом, увеличение мощности логически представленного знания за счет его интерпретации с применением больших данных или нейронных сетей всегда будет «погружено» в «пространство незнания» (некаузального знания), мощность которого много выше мощности множества любого логически представленного знания.

6. Заключение

В настоящей работе показана возможность ускоренного формирования когнитивной модели под новую проблемную ситуацию на основе автоматического анализа потока данных об этой ситуации. Для этого предварительно необходимо накапливать когнитивные модели с семантической интерпретацией факторов, которую целесообразно осуществлять с применением метода глубокого обучения. Эксперименты показывают, что результат автоматического распознавания факторов приближается к 93%.

В основу построения алгоритма положен авторский конвергентный подход с обращением к теории категорий и моноидальной структуризацией данных. Такой подход создает необходимые условия для ускоренной сходимости процесса глубокого обучения и коллективного принятия решения на когнитивной модели.

Достаточные условия достигаются за счет применения методов решения обратных задач и расширения семантических интерпретаций на сигнификативные семантики. Для более полного учета сигнификативных (когнитивных, мыслительных, эмоциональных, медитативных) семантик, по всей видимости, требуется апелляция к квантово-механическим аналогиям, порождающими особый нелокальный квантово-семантический ресурс.

Работа выполнена при поддержке Российского научного фонда, грант № 17-18-01326; Российского фонда фундаментальных исследований, грант № 15-29-07112.

Литература

[1] Raikov A.N., Panfilov S.A. Convergent Decision Support System with Genetic Algorithms and Cognitive Simulation // Proceedings of the IFAC Conference on Manufacturing Modelling, Management and Control, MIM’2013, Saint Petersburg, Russia, June 19-21, 2013. P. 1142-1147. DOI: 10.3182/20130619-3-RU-3018.00404.

[2] Raikov A.N., Avdeeva Z., Ermakov A. Big Data Refining on the Base of Cognitive Modeling // Proceedings of the 1st IFAC Conference on Cyber-Physical&Human-Systems, Florianopolis, Brazil. 7-9 December, 2016. P. 147-152. DOI: 10.1016/j.ifacol.2016.12.205.

[3] Гудфеллоу Я., Бенджио И., Курвилль А. Глубокое обучение. Пер. с англ. А.А.Синкина. М.: ДМК Пресс, 2017.

[4] Atmanspacher H. Quantum approaches to brain and mind. An overview with representative examples. The Blackwell Companion to Consciousness. Ed. Susan Schneider and Max Velmans. John Wiley & Sons Ltd. 2017. P. 298-313. DOI: 10.1002/9781119132363.ch21.

[5] Marsden D. Ambiguity and Incomplete Information in Categorical Models of Language. University of Oxford. R. Duncan and C. Heunen (Eds.). Quantum Physics and Logic (QPL) 2016. EPTCS 236, 2017. P. 95–107. DOI: 10.4204/eptcs.236.7.

[6] Le Q., Mikolov T. Distributed Representations of Sentences and Documents // Proceedings of the 31 st International Conference on Machine Learning, Beijing, China, 2014. JMLR:

W&CP volume 32. https://cs.stanford.edu/~quocle/paragraph_vector.pdf (просмотр 06.02.2018)
[7] Баргатин И.В., Гришанин Б.А., Задков В.Н. Запутанные квантовые состояния атомных систем // Успехи физических наук. 2001. Т. 171, № 6. С. 625 – 647.

Automated Synthesis of Cognitive Model on the Base of Big Data Analysis and Deep Learning

A.N. Raikov

Institute of Control Sciences RAS, Institute of Philosophy RAS

The cognitive models are created by team's participants or experts from different subjects' fields. They cannot quickly describe the problem situation with a high level of quality and integrity. Cognitive modeling takes into account denotative (formalized, material) and significative (cognitive, intellectual) semantics simultaneously. Ehe latter cannot be formalized. Cognitive modeling is a unique way to comprehensively reflect the features of the problem that has arisen. To ensure the integrity of the cognitive model, taking into account the fullest amount of information, as well as speeding up modeling and increasing the adequacy of models, this paper address the issue of performing automatic verification of cognitive models on the base of analyzing Big Data. It is also considered the possibility to perform automated synthesis of cognitive models. In the work the methods of deep learning, category theory and quantum semantics ideas are applied, as well as author's convergent approach to management and group decision making support. Experimentally the high level of quality of deep learning and accuracy of automatic recognition of elements of cognitive models are shown (93%). It can serve as a weighty guarantee of fruitfulness of offered solutions.

Keywords: big data, deep learning, quantum semantics, cognitive modeling, convergent approach

Корпус русских локальных документов и актов CorRIDA: цели формирования, состав, структура

С.А. Белов[1], О.В. Блинова[1], В.Б. Гулида[1], В.И. Зубов[1],
Е.Ю. Ларионова[2], П.С. Толстикова[3]

[1] Санкт-Петербургский государственный университет,
[2] Европейский Университет в Санкт-Петербурге,
[3] Институт лингвистических исследований РАН

s.a.belov@spbu.ru, o.blinova@spbu.ru, v-gulida@yandex.ru,
vladzubov21@gmail.com, e.j.larionova@gmail.com,
hokori.chan@gmail.com

Аннотация

В статье описывается начальный этап создания лингвистически размеченного корпуса русских локальных документов и актов CorRIDA. В повседневной жизни носители русского языка всё чаще сталкиваются с необходимостью читать и подписывать различные официальные документы. Обычно это так называемые локальные документы, например, «Договоры на оказание платных услуг» или «Информированные согласия». Однако язык локальных документов исследован недостаточно и практически не рассматривался с применением корпусных методов. Существующие корпусы русского языка пока не предоставляют возможностей для систематического анализа языка документа. Это связано в том числе с проблемами жанровой классификации и разметки нехудожественных текстов. Поэтому формирование корпуса локальных документов является актуальной задачей.
CorRIDA насчитывает 1,5 млн. слов, охватывает тексты, адресованные широким категориям пользователей (клиентам), принадлежащие трём социально значимым доменам (здравоохранение, образование, культура), и содержит в том числе разметку по типам текстов. Целью формирования корпуса является, во-первых, описание локальных документов разных типов через выделение и сравнение их языковых черт, во-вторых, оценка официально-деловых текстов с точки зрения их языковой сложности, удобства для восприятия и понимания «простым носителем» русского языка.

Ключевые слова: корпус русских локальных документов, официально-деловые тексты, типы текстов, социально-значимые домены (здравоохранение, образование, культура).

1. Русский официальный документ как объект изучения

В русистике существует традиция описания языка официально-деловых текстов. Прежде всего это работы в рамках стилистики, практической стилистики, «документальной лингвистики», которая в последнее время стала называться «документной лингвистикой», см., например, [1–5].

Компьютерной лингвистикой детальное описание деловых текстов рассматривалось преимущественно в контексте прикладных задач: для успешного решения вопросов технической и деловой коммуникации [6], в том числе — коммуникации человек-машина [7]. Решается задача автоматической кластеризации текстов, в частности, через отнесение к

одному из функциональных стилей, «регистров» или жанров, см., например, [8, 9], обзор см. в [10].

Насколько нам известно, специализированных корпусов русских официально-деловых текстов пока не существует. При этом электронные коллекции документов (репозитории оцифрованных и не оцифрованных текстов на русском языке, относящихся к разным историческим периодам, в том числе — к современности) относительно многочисленны. Упомянем, в частности, собрания нормативных документов на сайте РОМИП (документы Законодательства РФ, Москвы и Санкт-Петербурга, см. [11]), «Полное собрание законов Российской империи» РНБ [12], библиотеку нормативно-правовых актов СССР [13] и др.

Официально-деловые тексты вошли в состав некоторых корпусов русского языка. Так, в «Машинном фонде русского языка» был запланирован текстовой блок «деловая проза» [14]. В Основном корпусе НКРЯ [15] содержатся «нехудожественные тексты», которые можно выбирать, во-первых, по «сфере функционирования», — представлена, в том числе, «официально-деловая» сфера; во-вторых, по «типу текста» (представлены «деловые документы», «законодательные документы», «правовые документы», «судебные документы», «нотариальные документы», внутри каждого типа доступен поиск по жанрам); в-третьих, по «тематике текста» (представлены, в частности, тематические кластеры «администрация и управление» и «право»). Тексты официально-деловой сферы составляют 3,2% от объёма Основного корпуса, см. [16].

Юридические документы, всего 441 текст, в основном — кодексы, входят в Открытый корпус (OpenCorpora), [17]. Некоторое количество деловых текстов (из-за отсутствия жанровой разметки трудно сказать, какое) попало и в состав Russian Business Corpus, одного из корпусов С. Шарова на сайте Университета Лидса [18], и, видимо, в другие русскоязычные веб-корпусы.

Между тем, работ, в которых свойства русских официально-деловых текстов исследовались бы корпусными методами или хотя бы на корпусном материале, крайне мало, см., например, [19]. Это может объясняться, в частности, широчайшим жанровым разнообразием нехудожественных текстов, различиями в применяемых разными авторами типологиях жанров, отсутствием релевантной метаразметки внутри корпусов, а также недостаточной представленностью текстов отдельных жанров в их составе.

Именно в силу «несистематической представленности» текстов официально-делового стиля в существующих ресурсах планируется создание Корпуса официально-деловых текстов русского языка, куда будут включены законы Российской империи, СССР и РФ, императорские указы и постановления советского правительства, см. [20, с. 227].

Формируемый нами Корпус русских локальных документов и актов CorRIDA включает малоисследованную категорию текстов — так называмые **локальные документы** (Internal Documents). Они издаются в конкретной организации или на предприятии администрацией и касаются деятельности только этого предприятия или организации. Для включения в корпус выбраны **документы, адресованные пользователю (клиенту)**: пациенту в поликлинике, абитуриенту в университете и т. д. По-видимому, прежде всего с такими официальными текстами мы (носители русского языка) периодически сталкиваемся: например, читаем и подписываем «Согласия на обработку персональных данных», «Информированные добровольные согласия на медицинское вмешательство», или «Договоры об оказании платных дополнительных услуг».

2. Цели формирования корпуса CorRIDA

Создание корпуса CorRIDA (Corpus of Russian Internal Documents and Acts, Корпус русских локальных документов и актов) производится в рамках исследования, посвящённого функционированию официальных документов в социальных доменах здравоохранения, культуры и образования, подробнее см. [21]. Исследование имеет две магистральные линии, которые можно условно назвать «перцептивной» и

«дескриптивной». В рамках «перцептивного» направления производится анкетирование и интервьюирование носителей русского языка, направленное на выявление доступности официальных документов для восприятия и понимания представителями разных социальных групп.

«Перцептивная» часть исследования начата раньше «дескриптивной». Первоначальным её этапом стало анкетирование ограниченной выборки респондентов и проведение после анкетирования полуструктурированных интервью, см. [21]. Анкеты содержат перечень вопросов к обширным выдержкам из трёх текстов, находящихся в открытом доступе на сайтах учреждений: одного медицинского учреждения (клиники), одного образовательного учреждения (университета) и одного учреждения культуры (музея). Это тексты «Информированного согласия на проведение эндодонтического лечения», «Правил приема в Федеральное государственное бюджетное образовательное учреждение высшего образования» и «Правил поведения» для посетителей музея. Для массового анкетирования мы создали электронные формы анкет.

Перечень вопросов к каждому из трёх перечисленных документов (разных по уровню сложности для восприятия и понимания) направлен, среди прочего, на получение общей оценки текста. Данные массового опроса только предстоит обработать, здесь можно привести некоторые примеры ответов. Так, после выдержки из «Правил приема» следует просьба «*Опишите, пожалуйста, Ваше первое впечатление о тексте. Удобен ли он для чтения? Понятен ли?*».

Большинство респондентов отвечает, что текст в общем понятен (хотя понимание от многих требует усилий), но неудобен для чтения, перегружен, длинен и т. д., ср.: (1) «*слишком громоздко, но понятно*», (2) «*Неудобен, но в целом понятен*», (3) «*Понятен, но перегружен повторами*», (4) «*Все понятно, если читать медленно, но будет ли кто-то это делать? Очень долго, сложно и нудно*», (5) «*неудобен, так как длинные синтаксические конструкции; отчасти понятен, если заставить себя вчитаться. Описания "документов установленного образца" невозможно просто даже дочитать до конца*» и др. Часть респондентов уточняет, что текст будет понятен не всем читателям: (6) «*Нет, не удобен. Понятен более-менее, если уметь выделить в тексте главное. Если русский у человека второй язык (поступающие из союзных республик), то вообще ничего не будет понятно благодаря весёлому согласованию слов в предложениях*», (7) «*Мне да, абитуриенту - вряд ли <понятно> также, как мне*». Некоторые респонденты указывают, что не стали дочитывать фрагмент «Правил приема»: (8) «*многабукафниасилил))) очень длинный, внимание теряется на первых же пунктах*», (9) «*Бросила читать после первой трети, ненужные подробности отвлекают от сути*».

В рамках «дескриптивного» направления исследования планируется многоаспектное лингвистическое описание текстов документов, выполненное корпусными методами. В частности, мы планируем выяснить, насколько сложен для чтения текст официального документа (точнее, определённые типы текстов), опираясь на языковые свойства собранных в корпусе CorRIDA текстов. Для этого мы будем пользоваться и традиционными методами (включающими использование т. наз. «readability formulas», основанными на данных о средней длине предложения или средней длине слова, о частотах слов и пр.), и относительно более новыми методиками оценки языковой сложности (об этом см., в частности, [22, 23]).

Запланировав «дескриптивную» часть исследования, мы столкнулись с нехваткой текстовых ресурсов для её реализации (об официально-деловых текстах в составе русских корпусов см. п. 1 выше). Это привело к мысли о необходимости создания лингвистически размеченного корпуса, содержащего интересующие нас тексты локальных документов.

Общепринятой классификации жанров для нехудожественных текстов не существует, см., например, [8]. Вводятся классификации по разным основаниям, однако разработка жанровой таксономии, основанной на лингвистических свойствах текстов — сложная проблема, которую только предстоит решить, об этом см. [24].

При продумывании состава будущего корпуса мы решили совместить лингвистический и юридический взгляд на документ. Юридическая теория различает, прежде всего, **сделки** (правовые действия, которые порождают конкретные правовые последствия в виде обязательств самих участников сделок) и **правовые акты** (выражают волю уполномоченного лица устанавливать в одностороннем порядке предписания другим лицам). Сделки делятся на **односторонние сделки** (например, завещания) и **договоры** (для которых характерно встречное волеизъявление двух или более лиц). Правовые акты делятся на: **нормативные** (содержат нормы права — общие правила поведения, адресованные неопределенному кругу лиц и рассчитанные на неоднократное применение) и **ненормативные** (содержат конкретные предписания, адресованные конкретным лицам), см. [25].

При формировании корпуса нас интересовали тексты документов, которые можно отнести к категориям **односторонних сделок**, **договоров** и **нормативных правовых актов**. Различение этих категорий значимо и с точки зрения описания композиции и языкового содержания документа. Проиллюстрируем это утверждение одним примером: в текстах односторонних сделок («Согласий на обработку персональных данных») употребительны местоимения 1 л. и глаголы 1 л. ед. ч. («*я … подтверждаю своё согласие*», «*моих персональных данных*», «*предоставляю Оператору право*», «*согласие дано мной*» и т. д.), а в текстах договоров (например, договоров об оказании платных услуг) и разнообразных правил (например, правил поведения пациента) личные и притяжательные местоимения практически не встречаются, контрагенты или лица, чьи права и обязанности оговариваются в документе, поименованы стандартным образом (например, «*Пациент*» и «*Исполнитель*»).

Таким образом, включению в корпус CorRIDA подлежали локальные документы трёх перечисленных категорий (односторонних сделок, договоров и нормативных правовых актов), находящиеся в открытом доступе на сайтах государственных учреждений здравоохранения, образования и культуры (поликлиник, больниц, школ, университетов, музеев, театров и др.).

В конечном счете, нас интересует оценка официально-деловых текстов с точки зрения их языковой сложности, удобства для восприятия и понимания «простым носителем» русского языка (именно в этой точке смыкаются «дескриптивное» и «перцептивное» направления исследования), поэтому в корпус включались только **локальные документы, потенциально адресованные широкому пользователю (клиенту):** пациенту или посетителю в больнице, ученику или родителю в школе, зрителю в театре. К примеру, выбирая между «Кодексом этики и служебного поведения медицинского работника» и «Правилами поведения пациента», мы предпочитали последний тип документа, поскольку он адресован более обширной категории граждан, а не представителям одного профессионального сообщества или коллективу сотрудников конкретного учреждения.

3. Состав корпуса CorRIDA

В корпус вошли тексты односторонних сделок, договоров и нормативных правовых актов. Нас интересовали только **документы, выпущенные государственными учреждениями и адресованные широким категориям граждан**. В результате анализа содержания сайтов учреждений мы выбрали типы документов, которые размещаются на сайтах наиболее регулярно.

В категории так называемых «односторонних сделок» это, прежде всего, «Согласие на обработку персональных данных», а также «Информированное добровольное согласие», встречающееся в доменах здравоохранения. Среди договоров это «Договор об оказании платных услуг». В категории нормативных правовых актов это «Правила поведения (пациента, обучающегося, посетителя)», «Правила оказания платных услуг» и некоторые другие документы, различающиеся по доменам (например, «Правила госпитализации»,

«Правила проведения вступительных испытаний» или «Правила возврата театральных билетов»). Таким образом, мы получили 6 базовых разновидностей локальных документов, отвечающих нашим требованиям. Было решено назвать каждую такую разновидность «типом текста».

Таким образом, в корпус CorRIDA вошли документы, относящиеся к трём социально значимым доменам (образование, здравоохранение, культура), в каждом домене собраны тексты пяти типов. Шестой тип («Информированное согласие») представлен не во всех доменах, поэтому было решено объединить такие тексты с текстами «Согласий на обработку персональных данных».

Поиск и скачивание текстов выполнялось вручную. Такой порядок действий позволил существенно сократить количество времени и усилий, направленных на их предварительную обработку (о стандартных шагах по обработке текстов, собранных с помощью краулеров, см., например, [26]).

Опыт показал, что для нахождения документов удобно пользоваться поисковыми запросами с применением аббревиатур, принятых для обозначения государственных учреждений (ГБУЗ «государственное бюджетное учреждение здравоохранения», ГБОУ «государственное бюджетное образовательное учреждение», ГБУК «государственное бюджетное учреждение культуры», см. также аббревиатуры с префиксами «федеральное», «республиканское», «областное» типа ФГБУЗ, РГБУЗ, ОГБУЗ и др.).

В результате поиска в Интернете с применением кратких названий типов текста и аббревиатур, т.е. при помощи запросов типа «Правила поведения * ГБУЗ» нам удалось собрать текстовую коллекцию размером в 1,5 млн. слов. Состав коллекции описан в таблице 1.

Таблица 1. Состав корпуса в цифрах

домен	тип	кол-во текстов	кол-во слов	среднее значение (медиана)	мин. длина текста в словах	макс. длина текста в словах
Здравоохранение	1	77	110790	1439 (1223)	49	4859
Здравоохранение	2	135	107265	794 (547)	34	3863
Здравоохранение	3	59	101787	1725 (1630)	100	4991
Здравоохранение	4	70	100044	1429 (1362)	439	2766
Здравоохранение	5	153	50337	331.2 (313.5)	27	787
Здравоохранение	6	99	49567	500.7 (362.0)	60	9219
Образование	1	38	105905	2787 (2948)	255	5513
Образование	2	50	104513	2090.3 (1023)	202	13035
Образование	3	51	100432	1969 (1888)	219	4498
Образование	4	73	102318	1402 (1354)	345	3003
Образование	5	258	100973	391.2 (367)	60	1349
Культура	1	106	100861	951.5 (832)	118	3220
Культура	2	62	103312	1666.3 (1487.5)	315	5912
Культура	3	65	100292	1543 (1591)	91	4404
Культура	4	83	100012	1205 (1122)	261	3087
Культура	5	179	45162	252.3 (233)	31	1747
ВСЕГО		**1558**	**1483570**			

Типы текстов обозначены индексами: «1» — Правила поведения (Правила внутреннего распорядка и поведения пациентов, обучающихся и др.); «2» — Порядок (правила) госпитализации, диспансеризации, организации вызова врача, приёма в школу, колледж, университет, Положение о порядке перевода, восстановления, отчисления обучающихся и др.; «3» — Положение об оказании платных услуг (Порядок оказания платных услуг), «4» — Договор на оказание платных услуг (Договор об оказании платных услуг), «5» — Согласие на обработку персональных данных пациента, законного представителя, ребёнка,

родителя и др., «6» — Информированное добровольное согласие (на медицинское вмешательство и др.).

Устойчивые характеристики типов документов, связанные с объёмом в словах и, соответственно, косвенно влияющие на сложность восприятия соответствующих текстов, ещё предстоит описать. Уже сейчас, основываясь на таблице 2, можно заключить, что наиболее протяженными в словах типами текстов в корпусе являются тип «1» в домене «Образование» («Правила внутреннего распорядка обучающихся», «Правила поведения учащихся» и т. п.) и тип «3» в том же домене («Порядок оказания платных образовательных услуг»).

Таблица 2. Ранжирование по убыванию медианных значений длины текстов в словах

домен	тип текста	медиана
Образование	1 (Правила поведения)	2948
Образование	3 (Положение об оказании платных услуг)	1888
Здравоохранение	3 (Положение об оказании платных услуг)	1630
Культура	3 (Положение об оказании платных услуг)	1591
Культура	2 (Порядок (правила) сдачи театральных билетов и др.)	1487
Здравоохранение	4 (Договор на оказание платных услуг)	1362
Образование	4 (Договор на оказание платных услуг)	1354
Здравоохранение	1 (Правила поведения)	1223
Культура	4 (Договор на оказание платных услуг)	1122
Образование	2 (Порядок (правила) отчисления обучающихся и др.)	1023
Культура	1 (Правила поведения)	832
Здравоохранение	2 (Порядок (правила) госпитализации и др.)	547
Образование	5 (Согласие на обработку персональных данных)	367
Здравоохранение	6 (Информированное добровольное согласие)	362
Здравоохранение	5 (Согласие на обработку персональных данных)	313
Культура	5 (Согласие на обработку персональных данных)	233

4. Структура корпуса CorRIDA

Корпус будет состоять из трёх подкорпусов по доменам (домены «здравоохранение», «образование», «культура») и шести подкорпусов по типам текстов. Так как коллекция, которая станет основой корпуса, достаточно однородна по составу, представляется, что минимального набора метаданных для описания текстов в её составе достаточно. Каждый документ внутри коллекции сопровождается следующим набором сведений:

- название домена;
- название учреждения, с сайта которого получен документ;
- название документа;
- стандартное название типа документа;
- условный номер документа в субколлекции;
- источник электронной версии (адрес сайта в Интернете, с которого документ был скачан);
- дата скачивания документа.

Пользователь корпуса будет иметь доступ к информации о домене, типе документа, названии документа.

Документы скачивались вручную, поэтому предварительная обработка текстов для включения в корпус была несложной и заключалась, прежде всего, в удалении подряд идущих пробелов и табуляций, удалении пустых строк, удалении строк, содержащих только пробелы или табуляции, удалении пробелов в начале строк, замене парных кавычек на прямые и т. п. Дедупликация на уровне целых документов не требуется, так как тексты, вошедшие в корпус, оценивались экспертами (дубликаты в корпус не включались). Проверка на наличие нечетких дубликатов на уровне компонентов документа не

планируется, так как нам интересны, в том числе, повторяющиеся элементы, способные показывать строгость соблюдения в конкретном тексте определенного шаблона.

Вопросом, потребовавшим решения, стал вопрос об анонимизации данных. В корпусной практике анонимизации подвергаются различные персональные данные, в первую очередь — фамилии, адреса, телефонные номера и т. д.

В нашем корпусе собраны документы, размещённые в Интернете в открытом доступе. Тем не менее, поскольку в дальнейшем мы, во-первых, планируем сделать корпус общедоступным, предоставив желающим возможность его скачивания, во-вторых, намереваемся оценивать документы с точки зрения их языковых качеств, мы приняли решение выполнить анонимизацию некоторых личных данных.

Существует два основных способа анонимизации:

1. Удаление данных [27, p. 62].

2. Замена данных, например, замена имён псевдонимами, буквенными или цифровыми кодами. О случаях, когда замены необходимы или нежелательны, см. [28, p. 14–19]. В частности, правила замен для фамилий таковы: псевдонимы должны быть фонетически и просодически похожи на оригинальные имена, содержать одинаковое число слогов, должны начинаться с тех же букв, что и оригинальные имена.

Можно с уверенностью утверждать, что в документе имена собственные не обыгрываются ни фонетически, ни ритмически, ни в ходе языковой игры, поэтому отображать их первоначальный облик в псевдонимах не обязательно. Соответственно, выбран способ анонимизации, при котором названия государственных учреждений, почтовые и электронные адреса, телефоны и пр. частично преобразуются, при этом количество слов в тексте сохраняется неизменным. В результате получаем наименования и фамилии типа: «*ГБУЗ NК "Nская ЦРБ"*», «*КГАУЗ "Nская стоматологическая поликлиника"*», «*портал государственных услуг N-ского карая* (sic) *(https://uslugi00.ru/)*», «*записаться на прием можно лично у секретаря главного врача либо по телефону: 000-00-00*», «*Nкину Константину Владимировичу*», «*Nкиной Ирине Николаевне*» и пр. Таким образом, мы легко можем определить место замены и общее значение слова, аббревиатуры, алфавитно-цифрового комплекса и т. п., подвергшегося преобразованию.

5. Заключение

В повседневной жизни носители русского языка всё чаще сталкиваются с необходимостью читать и подписывать различные официальные документы. Обычно это так называемые локальные документы (Internal Documents). Мы встречаемся с ними, обращаясь за медицинской помощью, оформляя Шенгенскую визу или возвращая в кассу театральные билеты. Между тем, язык таких документов исследован недостаточно и практически не рассматривался с применением корпусных методов.

Существующие корпусы русского языка пока не предоставляют возможностей для систематического анализа языка документа. Это связано, в частности, с проблемами жанровой классификации и разметки нехудожественных текстов. Поэтому мы решили сформировать лингвистически размеченный Корпус русских локальных документов и актов CorRIDA. Этот корпус позволит выполнить описание локальных документов разных жанров через выделение и сравнение их языковых черт.

Исследование выполняется в рамках НИР по анализу соблюдения норм современного русского литературного языка при его использовании в качестве государственного в деятельности организаций культуры, здравоохранения и образования, включённой в План мероприятий НИИ Проблем государственного языка СПбГУ во исполнение Комплекса мер, направленных на совершенствование государственной политики в области развития, защиты и поддержки русского языка на 2016-2020 гг. Совета по русскому языку при Правительстве РФ.

Литература

[1] Рахманин Л.В. Стилистика деловой речи и редактирование служебных документов. Учебное пособие. М.: Высшая школа, 1988.

[2] Шварцкопф Б.С. Официально-деловой язык // Культура русской речи и эффективность общения. М., 1996. С. 270 – 281.

[3] Дюженко Г.А. Документальная лингвистика. М., 1975.

[4] Янковая В.Ф. Документная лингвистика. М.: Издательский центр «Академия», 2011.

[5] Муравьёва М.Н. Документная лингвистика. М.: Издательство «ТЕРМИКА», 2016.

[6] Герд А.С. Предмет и основные направления прикладной лингвистики // Прикладное языкознание. СПб., 1996. URL: http://project.phil.spbu.ru/lib/data/ru/heard/prikling.html (дата обращения: 27.01.2018).

[7] Ершов А.П. К методологии построения диалоговых систем. Феномен деловой прозы. Новосибирск, 1979.

[8] Пиперски А.Ч. Жанровая классификация в Генеральном интернет-корпусе русского языка // Современные проблемы науки и образования. 2013. № 4. URL: https://www.science-education.ru/ru/article/view?id=9762 (дата обращения: 27.01.2018).

[9] Дубовик А.Р. Автоматическое определение стилистической принадлежности текстов по их статистическим параметрам // Компьютерная лингвистика и вычислительные онтологии. Выпуск 1 (Труды XX Международной объединенной научной конференции «Интернет и современное общество, IMS-2017, Санкт-Петербург, 21 - 23 июня 2017 г. Сборник научных статей). СПб: Университет ИТМО, 2017. С. 29 – 45. URL: http://openbooks.ifmo.ru/ru/file/6502/6502.pdf (дата обращения: 27.01.2018).

[10] Conrad S. Register variation // Biber D., Reppen R. (eds.) The Cambridge handbook of English corpus linguistics. Cambridge University Press, 2015. P. 309 – 329.

[11] РОМИП (Российский семинар по оценке методов информационного поиска). URL: http://romip.ru/index.html (дата обращения: 27.01.2018).

[12] Полное собрание законов Российской империи. URL: http://www.nlr.ru/e-res/law_r/about.html (дата обращения: 27.01.2018).

[13] Библиотека нормативно-правовых актов СССР. URL: http://www.libussr.ru/ (дата обращения: 27.01.2018).

[14] Леонтьева Н.Н. Об информационной системе словарей Машинного фонда русского языка // Машинный фонд русского языка: идеи и суждения. М.: Наука, 1986. С. 109 – 125.

[15] Национальный корпус русского языка. URL: http://www.ruscorpora.ru/ (дата обращения: 27.01.2018).

[16] Статистика корпуса. URL: http://www.ruscorpora.ru/corpora-stat.html (дата обращения: 27.01.2018).

[17] «Открытый корпус» (OpenCorpora). URL: http://opencorpora.org/ (дата обращения: 27.01.2018).

[18] Russian Business Corpus. URL: http://corpus.leeds.ac.uk/ruscorpora.html (дата обращения: 27.01.2018).

[19] Буторина Е.П. Категория официальности в современном русском языке. Автореф. дисс. … докт. филол. наук. М., 2016.

[20] Крылов С.А., Фролова О.Е. О корпусе официально-деловых текстов русского языка // Труды международной конференции «Корпусная лингвистика-2017». СПб, 2017. С. 226 – 230.

[21] Гулида В.Б. Социолингвистическая проблематика официальных документов // Социо- и психолингвистические исследования. Вып. 4. 2016. С. 112–125. URL: https://elibrary.ru/item.asp?id=28381522 (дата обращения: 27.01.2018).

[22]Hancke J., Vajjala S., Meurers D. Readability classification for German using lexical, syntactic, and morphological features // Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012). Mumbai, India, 2012. P. 1063–1080.
[23]Crossley S., Dufty D., McCarthy Ph., McNamara D. Toward a new readability: A mixed model approach // Proceedings of the 29th annual conference of the Cognitive Science Society. Nashville, Tennessee, USA, 2007. P. 197–202.
[24]Кибрик А.А. Анализ дискурса в когнитивной перспективе. Дис. … д-ра филол. наук. М.: Ин-т языкознания РАН. 2003. URL: http://iling-ran.ru/kibrik/DA_cognitive_perspective@Diss_2003.pdf (дата обращения: 27.01.2018).
[25]Бошно С.В. Развитие признаков нормативного правового акта в современной правотворческой практике // Журнал российского права. 2004. №2. С. 95 – 106. URL: https://elibrary.ru/item.asp?id=26350658 (дата обращения: 27.01.2018).
[26]Barbaresi A. Ad hoc and general-purpose corpus construction from web sources. Linguistics. ENS Lyon, 2015. URL: https://tel.archives-ouvertes.fr/tel-01167309/document (дата обращения: 27.01.2018).
[27]McEnery T., Hardie A. Corpus Linguistics: Method, Theory and Practice. Cambridge University Press, 2012.
[28]Hasund K. Protecting the innocent: the issue of informants' anonymity in the COLT corpus // Explorations in Corpus Linguistics. Amsterdam: Rodopi, 1998. P. 13 – 28.

Corpus of Russian Internal Documents and Acts CorRIDA: Goals, Composition and Structure

S.A. Belov [1], O.V. Blinova [1], V.B. Gulida [1], V.Yu. Zubov [1],
E.Yu. Larionova [2], P.S. Tolstikova [3]

[1] St. Petersburg State University, [2] European University at Saint-Petersburg,
[3] Institute for Linguistic Studies, Russian Academy of Sciences

The existing Russian corpora do not yet provide opportunities for a systematic analysis of the language of official documents. There are few such texts in existing corpora. Moreover, there are the problems of genre classification and markup of non-fiction (incl. official, legal) texts.
The paper describes the initial creation stage of the corpus of Russian Internal Documents and Acts «CorRIDA». In everyday life, Russian speakers are increasingly faced with the need to read and sign various official documents. Usually these are so-called «internal documents», for example, Contracts or Informed Consents. However, the language of such documents has not been examined with the use of corpus methodology.
The corpus contains 1.5 million words, includes documents belonging to three socially significant domains (health, education, culture) and will allow the description of internal documents of various types.

Keywords: Legal Corpora, Official Texts, Corpus of Russian Internal Documents, Socially Important Domains

К вопросу о репрезентации данных о сочетаемости в электронных лексикографических ресурсах

М.В. Хохлова[1], А.М. Попов[2]

[1] Санкт-Петербургский государственный университет, [2] ООО Инфо-Кьюбс

m.khokhlova@spbu.ru, hedgeonline@gmail.com

Аннотация

В статье дается обзор существующих систем, представляющих информацию о сочетаемости. К ним относятся разнообразные словари, а также специализированные базы данных и другие ресурсы. Также затрагиваются вопросы, связанные с реализацией проекта по созданию интегрированной базы данных, которая содержит автоматически извлеченные коллокации и дополнительную информацию. В системе будут представлены примеры, полученные как на основе корпусов при помощи автоматических методов, так и на материале словарей русского языка. Ресурс может быть использован в разнообразных задачах прикладной лингвистики, в том числе связанных с автоматической обработкой текстов.

Ключевые слова: сочетаемость, коллокации, словари, база данных, корпусы текстов, статистика

1. Введение

Сведения о сочетаемости лексических единиц традиционно представлены в словарях и справочниках. При этом с появлением больших корпусов текстов и методов их обработки открылись новые возможности для репрезентации данных подобного рода. Прежде всего имеются в виду количественные подходы, которые хотя и стали использоваться применительно к русскоязычному материалу довольно давно (см., например, первые исследования [1]), однако получили новый импульс в связи с развитием информационных технологий. Таким образом, появилась возможность описать сочетаемость с точки зрения ее двусторонней природы, имеющей как языковой, так и вероятностный характер.

2. Обзор проектов

2.1. Словарные источники

Информация о лексической и синтаксической сочетаемости различных слов обычно описана в словарях (толковых, специализированных и др.), реже в языковых грамматиках. Для русского языка можно назвать целый ряд толковых словарей, в которых сочетаемость отражена довольно подробно [2-4 и др.]. Тем не менее, можно отметить, что не существует единой концепции описания данных в разных источниках. В толковых словарях информация, указывающая на ограниченную сочетаемость, представлена несколькими способами: знак ромба обозначает устойчивые словосочетания и фразеологизмы в словаре [3] (в нем приводятся данные о 13 тыс. подобных сочетаний при общем объеме словника более 80 тыс. единиц), в то время как в [4] для первых используется он же, а вторые вводятся при помощи знака тильды. Сочетаемость может не выделяться специальным образом, а приводиться в цитатах или речениях. Сами словарные статьи также имеют

разную структуру. Так, для слова «надежда» словари [3, 5] перечисляют только следующие устойчивые словосочетания «возлагать надежду», «питать надежду», «подает надежду» и «льстит себя надеждой». Однако эти примеры не включают другие словосочетания, которые тоже являются свойственными этой лексеме (например, «оправдывать надежды» или «вселять надежду»), поэтому специализированные словари могут служить источником дополнительной информации. Для русского языка такие словари существуют и являются великолепными справочными ресурсами, представляющими сочетаемость [6, 7]. Необходимо отметить, что словарь [4] является в некоторой степени уникальным, так как словарные статьи содержат особый справочный отдел, в котором приводятся сведения из других словарных источников. В основном они касаются первой фиксации слова или изменений, произошедших в его произношении или написании. Также существует уникальный лексикографический проект под руководством Ю.Д. Апресяна «Активный словарь русского языка» [8], который включает обширную информацию о сочетаемости, которая выделяется отдельно в словарных статьях. Материал отлично структурирован и включает сведения о синтаксических актантах, коллокациях и конструкциях. Тем не менее, словари по-разному отражают сочетаемость и покрывают примеры, поэтому так важно рассмотрение отличных друг от друга источников.

2.2. Электронные ресурсы и базы данных

Методы, применяемые во многих задачах прикладной лингвистики, можно разделить на две большие категории: использующие правила и реализующие статистические алгоритмы. Применительно к задаче автоматического определения сочетаемости можно сказать, что системы, основанные на первом подходе, появились раньше. В некоторой степени они напоминают словари, представленные в электронном виде. Информация о сочетаемости может быть получена в них для определенных моделей и с отсылкой к корпусу. Второй подход реализован в значительно меньшем количестве систем и подразумевает автоматическое извлечение сочетаемостной информации статистическими методами из данных большого объема. Далее выделенные словосочетания могут сопровождаться количественными характеристиками, позволяющими оценить степень устойчивости (или воспроизводимости) конструкций.

В основном работы, посвященные описанию сочетаемости и ее последующему представлению в специализированных базах данных, на протяжении долгого времени затрагивали англоязычный материал. Так, был разработан словарь «The Pattern Dictionary of English Verbs», который базируется на методологии Corpus Pattern Analysis, предложенной П. Хэнксом [9] и включает семантико-синтаксические шаблоны глагольного управления с иллюстрациями (словосочетаниями и предложениями). Уникальным проектом является ресурс FrameNet, созданный Ч. Филлмором [10]. В нем представлена информация о валентности и о семантических ролях для английского языка Общее число примеров, эксплицирующих употребление лексических единиц, превышает 200 тыс. предложений. Проекты, выполненные в русле этого подхода, существуют для испанского, китайского, корейского, немецкого, французского, шведского и японского языков, а также для бразильского варианта португальского языка. Также можно назвать проект “Collocations” Ланкастерского университета, который посвящен определению сочетаемости у пар синонимов на материале корпуса BNC [11]. Эта же функция реализована в ряде корпусов текстов английского языка, разработанных М. Дэвисом и доступных на его портале. Пользователь имеет возможность искать слова, находящиеся в одном и том же контекстном окружении с ключевым словом (функция “compare”), так и проверять сочетаемость у близких по значению единиц [12].

Для английского языка существует лексическая база данных DANTE [13], в которой описаны свыше 40 тыс. наиболее частотных слов. Каждое значение лексической единицы проиллюстрировано автоматически собранными цитатами из двухмиллиардного корпуса.

Статистический механизм лежит в основе подхода к представлению сочетаемости в системе Sketch Engine [14, 15]. В ней можно получить информацию о контекстном окружении лексем для многих языков в виде таблиц, моделирующих сочетаемость и соответствующих заранее определенных синтаксическим моделям. Автоматически выделенные словосочетания сопровождаются количественными оценками, указывающими на силу синтагматической связи. Также существует многоязычный проект SkELL (Sketch Engine for Language Learning) [16], в котором представлены заранее отобранные цитаты, репрезентирующие словоупотребление. В рамках данного проекта представлены корпусы для учебных целей, в которых представлены «чистые» тексты и наиболее удачные примеры.

Что касается других языков, то отдельно можно выделить систему DWDS (Das Wortauskunftssystem zur deutschen Sprache in Geschichte und Gegenwart) для немецкого языка, которая объединяет словарь и базу данных [17]. Система IDS, разработанная в Институте немецкого языка, предоставляет возможность просматривать более чем 220 тыс. коллокационных профилей [18]. Выдаваемые коллокации сопровождаются статистической оценкой при помощи коэффициента логарифмического правдоподобия. Также для каждого примера приводится наиболее характерная синтаксическая модель словосочетания. Для словенского языка была разработана база данных коллокаций [19]. В ней представлены более 44 тыс. словосочетаний и 150 тыс. примеров.

На сегодняшний день также существует ряд проектов, ориентированных на материал русского языка и направленных на статистическое исследование лексической и синтаксической сочетаемости на основе корпусов текстов. К ним относятся, например, база данных «Лексикограф» [20], база сочетаемости FrameBank, в которой представлены лексические конструкции [21], и словари, созданные на основе НКРЯ [22]. Среди последних можно назвать словарь глагольной сочетаемости непредметных имен русского языка и словарь русской идиоматики. Первый проект основывается на понятии лексических функций, введенном в работе [23], и описывает более 10 тыс. словосочетаний следующих моделей: 1) существительное + глагол; 2) глагол + существительное; 3) глагол + прилагательное + существительное. Есть возможность проводить поиск по значению, синтаксическому отношению, фазовому значению и оценке. Тем не менее, в словаре отсутствует количественная характеристика силы данной сочетаемости, которая была бы весьма актуальна для исследователей.

Проект CoSyCo ориентирован на создание базы данных синтаксических конструкций, в которой на данный момент представлены именные и глагольные словосочетания [24]. Поиск единиц осуществляется в корпусах разных функциональных стилей: художественном, научном и публицистическом. Каждое опорное слово (по которому производится поиск) снабжается списком синтагматических партнеров с их частотами.

3. База данных

3.1 Вводные замечания

На основании анализа имеющихся систем, можно сказать, что существует необходимость в интегрированном ресурсе, который бы давал доступ к информации о сочетаемости, полученной при помощи разнообразных квантитативных методов на материале корпусов текстов, а также сопровождал бы словосочетания ссылками на традиционные лексикографические ресурсы. Таким образом, речь идет о совмещении двух упомянутых ранее подходов.

После проведенного анализа толковых и специализированных словарей русского языка был сделан вывод о том, что необходимо также ввести единый формат представления информации о сочетаемости, так как словарные статьи в разных источниках имеют разную

структуру. Например, в едином формате могут быть указаны части речи входящих в словосочетание единиц, а также информация о том, в каких словарях они встретились.

3.2. Структура базы данных

На данном этапе работы была произведена морфологическая разметка текстов, включающая неоднозначность (см. пример разбора для словосочетания «по словам министра»). Так, единица «министра» анализируется как существительное в родительном или винительном падежах. Также мы решили рассматривать многокомпонентные единицы как единое целое (например, «по словам»). Это даст возможность пользователю в качестве выходных данных рассматривать одну единицу, не разделенную на слова, что важно для целого ряда задач.

<token start="7608" end="7610" value="по" lemma="ПО" tag="W,Prep"/>
<token start="7608" end="7610" value="по" lemma="по" tag="Prefix"/>
<token start="7611" end="7617" value="словам" lemma="СЛОВО" tag="W,Noun,dat,neu,pl,in,NP"/>
<token start="7608" end="7617" value="по словам" lemma="по словам" tag="Prnt"/>
<token start="7618" end="7626" value="министра" lemma="МИНИСТР" tag="W,Noun,gen,mas,sg,an,NP"/>
<token start="7618" end="7626" value="министра" lemma="МИНИСТР" tag="W,Noun,acc,mas,sg,an,NP"/>

В ходе работы над проектом коллокации нами извлекались в два этапа. На первой стадии нами были разработаны правила, которые описывали русскоязычные словосочетания и были применены к извлечению данных. Следующие синтаксические модели представлены в базе данных: 1) ADJ+N; 2) N+N; 3) V+N; 4) V+Prep+N. Второй этап включал такие статистические метрики как MI, t-score, log-likelihood и другие меры ассоциации [25]. Статистические данные использовались для оценки извлеченных словосочетаний.

Как уже указывалось, лексикографические ресурсы предоставляют полезную информацию о сочетаемости. Поэтому при создании системы мы планируем использовать данные из ряда словарей [2-6]. Будет использована информация, иллюстрирующая контексты слов (например, речения, цитаты, указания на ограниченную сочетаемость). Такие данные также позволят оценить и верифицировать коллокации в базе данных.

В качестве материала при разработке базы данных были привлечены разнообразные тексты. Были использованы новостная коллекция, специальные тексты (техника) и беллетристика. Также были выполнены эксперименты на доступных русскоязычных данных следующих корпусов:: ruTenTen, Aranea Russicum Maximum и НКРЯ.

Для нашего исследования нами была разработана специализированная база данных MySQL для хранения пар слов и их корреляционных значений согласно разным коллокационным мерам, содержащая три основных таблицы: таблицу слов; таблицу коллокаций; таблицу метрик.

Таблица слов состоит из троек, каждая тройка занимает свой ряд и хранит следующую информацию: уникальное слово (в нашем случае словарная форма, т.е. лемма), тег части речи и их совместная частота (сколько раз данная пара «слово-часть речи» встретилась в рассматриваемом корпусе).

Таблица коллокаций содержит частотную информацию о каждой паре ряда из таблицы слов, включая информацию о линейном порядке (это означает, что сочетание двух слов с противоположным порядком будет рассмотрено как два разных сочетания и независимыми частотами).

Ниже показана схема базы данных (см. рисунок 1).

Также имеется доступ ко всей необходимой количественной информации (размер корпуса, частота биграммы и независимые частоты двух отдельных слов), которая необходима для вычисления мер связанности, хранящихся в третьей таблице. Она

содержит значения статистических мер для каждой коллокации (ряда в таблице коллокаций), словосочетание может иметь иного значений метрик, вычисленных заранее и хранящихся в базе данных для исследовательских целей. Каждая метрика снабжена уникальным значимым строковым ID (обычно название метрики).

Рис. 1. Пример схемы базы данных

Простой SQL запрос предоставляет средства для извлечения необходимой информации с требуемыми ограничениями. Например, пользователь может ограничиться определенными частями речи и/или типами метрик.

Таблицы слов и коллокаций заполняются заранее созданной нами программой, записывая выходной результат обработки корпуса прямо в базу. Таблица мер, с другой стороны, может быть заполнена позже и даже увеличена, когда требуется ввести новую коллокационную метрику в базу. Также возможно вычислить некоторые метрики онлайн при помощи формулы, инкорпорированной в SQL запрос select.

4. Заключение

Следующим шагом могут быть методы машинного обучения применительно к задаче автоматического определения сочетаемости на материале языковых данных.

В статье мы попытались проследить основные принципы, лежащие в основе базы данных. Главная проблема заключается в необходимости интегрированного ресурса, который включит данные из словарей и корпусов текстов, снабженные достаточным количеством примеров для каждого слова. База данных использует MySQL и уже доступна пользователям на сайте в сети Интернет. Она содержит три таблицы, каждая из которых хранит информацию о словах, коллокациях и статистических мерах соответственно. Ресурс включит примеры из ряда корпусов (как специализированных, так и общих), статистическую оценку выделенных коллокаций при помощи мер ассоциации, ссылки на другие ресурсы, такие как словари и корпусы текстов. Разрабатываемая база данных может быть использована при обучении русскому как иностранному, для создания приложений, связанных с автоматической обработкой текстов, и при составлении словарей. В качестве будущей перспективы мы планируем добавить оценку пользователей для выделенных коллокаций.

Статья подготовлена в рамках работы по гранту Президента Российской Федерации для государственной поддержки молодых российских ученых № МК-2513.2018.6 «Исследование методов автоматического извлечения лексических конструкций на основе машинного обучения».

Литература

[1] Марков А.А. Об одном применении статистического метода // Известия Императорской Академии Наук. Серия VI. 1916. Т. 10, № 4.

[2] Словарь современного русского литературного языка: В 17 т. / Под ред. В.И. Чернышёва. М., Л.: Изд-во АН СССР, 1948—1965.
[3] Словарь русского языка: В 4-х т. / АН СССР, Ин-т рус. яз.; Под ред. А.П. Евгеньевой. 2-е изд., испр. и доп. М.: Русский язык, 1981—1984.
[4] Большой академический словарь русского языка: В 30 т. / Под ред. К.С. Горбачевича. СПб: Изд-во «Наука», 2004.
[5] Большой толковый словарь русского языка: А-Я / РАН. Ин-т лингв. исслед.; Сост., гл. ред. канд. филол. наук С.А. Кузнецов. СПб: Норинт, 1998.
[6] Борисова Е.Г. Слово в тексте: Словарь коллокаций (устойчивых сочетаний) рус. яз. с англо-рус. слов. ключевых слов. М.: Филология, 1995.
[7] Словарь сочетаемости слов русского языка / Под ред. П.Н. Денисова, В.В. Морковкина. – 3-е изд., испр. М., 2002.
[8] Активный словарь русского языка. Т. 1. А—Б / Отв. ред. акад. Ю.Д. Апресян. М.: Языки славянской культуры, 2014.
[9] Hanks P. Mapping meaning onto use: a Pattern Dictionary of English Verbs. AACL, Utah 2008.
[10] FrameNet. URL: https://framenet.icsi.berkeley.edu/fndrupal (дата обращения: 25.05.2018).
[11] Getting Collocations. URL: https://www.lancaster.ac.uk/fss/courses/ling/corpus/blue/l08_3.htm (дата обращения: 25.05.2018).
[12] Corpora created by Mark Davies. URL: https://corpus.byu.edu (дата обращения: 25.05.2018).
[13] Database of Analyzed Texts of English. URL: http://www.webdante.com/index.html (дата обращения: 25.05.2018).
[14] Sketch Engine. URL: http://www.sketchengine.co.uk (дата обращения: 25.05.2018).
[15] Kilgarriff A., Baisa V., Bušta J,, Jakubíček M., Kovář V., Michelfeit J., Rychlý P., Suchomel V. The Sketch Engine: ten years on. Lexicography. 1. 2014. P. 7-36.
[16] SkELL – examples and collocations for learners of English. URL: https://www.sketchengine.eu/skell/ (дата обращения: 25.05.2018).
[17] Das Wortauskunftssystem zur deutschen Sprache in Geschichte und Gegenwart. URL: https://www.dwds.de/ (дата обращения: 25.05.2018).
[18] Kookkurrenzdatenbank CCDB. URL: http://corpora.ids-mannheim.de/ccdb/ (дата обращения: 25.05.2018).
[19] Slovene Lexical Database. URL: http://eng.slovenscina.eu/spletni-slovar/leksikalna-baza (дата обращения: 25.05.2018).
[20] Проект «Лексикограф». URL: http://lexicograph.ruslang.ru (дата обращения: 25.05.2018).
[21] FrameBank. URL: http://framebank.ru/ (дата обращения: 25.05.2018).
[22] Словари, созданные на основе Национального корпуса русского языка. URL: http://dict.ruslang.ru/ (дата обращения: 25.05.2018).
[23] Бирюк О.Л., Гусев В.Ю., Калинина Е.Ю. Словарь глагольной сочетаемости непредметных имен русского языка. http://dict.ruslang.ru (дата обращения: 25.05.2018).
[24] Kopotev M., Escoter L., Kormacheva D., Pierce M., Pivovarova L., Yangarber R. CoCoCo: Online Extraction of Russian Multiword Expressions // Proceedings of the 5th Workshop on Balto-Slavic Natural Language Processing (10–11 September 2015, Hissar, Bulgaria). Sofia: INCOMA Ltd, 2015. P. 43–45.
[25] Manning Ch., Schütze H. Foundations of Statistical Natural Language Processing. Massachusetts: MIT Press, 1999.

On the Representation of Collocability in Online Lexicographic Resources

M. Khokhlova[1], A. Popov[2]

[1] Saint-Petersburg State University, [2] Info-Qubes

The paper gives an overview of the existing systems that represent collocability. Among these one can name various dictionaries and specialized databases and other resources. The authors also pay attention to the issues related to the project on building an integrated database that includes automatically extracted collocations and additional information. The system will comprise both automatically extracted examples and ones from Russian dictionaries. The present tool can be used in a wide range of tasks of applied linguistics, e.g. natural language processing.

Keywords: collocability, collocations, dictionaries, database, text corpora, statistics

Оригинальные метафорические употребления цветообозначений (корпусный анализ)

А.В. Чекменева

Санкт-Петербургский государственный университет

anyasidr1@gmail.com

Аннотация

В данной работе автор исследует проблему употребления метафорических цветообозначений в русском языке на основе корпусных исследований.
Из всей лексико-семантической группы прилагательных, обозначающих цвет, выделяются 12 базовых цветов. Исследуется корреляция абстрактных существительных с прилагательными-цветообозначениями. Определяется коннотация для каждого базового цвета.

Ключевые слова: корпусы текстов, частотность лексических единиц, поэтический корпус НКРЯ, прилагательные обозначающие цвета, теория базовых обозначений цвета Берлина и Кэя, синестетическая метафора

1. Введение

Проблема метафоры является чрезвычайно популярной и актуальной темой в современном языкознании. Метафора прочно укоренилась в современном языке, являясь не только тропом, средством усиления образности языка и выразительности речи, но и способом передачи новых понятий и явлений. При описании какого-либо объекта не в последнюю очередь указывается его цвет, причем обозначение цвета может обладать и прямым значением («черный стол», «белая ваза»), и метафорическим, когда предмет не может иметь цвет по каким-либо причинам, но наделяется им.

Цветообозначения являются важнейшей частью культуры любого народа. Они «отображают особенности категоризации и концептуализации цветового пространства того или иного этноса» [1, с.74–76]. То есть, у человека формируется свойственное его культуре представление о цветовой картине мира, что непременно отображается в языке. Вследствие чего у каждого народа появляются свои уникальные примеры употребления цветообозначений в речи, формирование которых зависит от культурных традиций народа, окружающего его мира и особенностей менталитета.

В данной работе будут исследоваться метафоры, которые являются сочетанием прилагательного, имеющего семантический признак «цвет», с абстрактным существительным. Также будет проведено сравнение частотности выбранных нами далее прилагательных и полученных в результате поиска метафор, будут выявлены наиболее редкие и уникальные метафоры.

Иоганн Вольфганг фон Гете в труде «К теории цвета» (нем. Zur Farbenlehre) [2] утверждал, что цвет имеет способность воздействовать на человека, и выделял при этом два способа возможного воздействия: физиологическое (на физическую оболочку человека) и психическое (на духовный мир человека).

По нашему предположению, из вышеприведенного утверждения Гете о том, что цвет может воздействовать на духовный мир человека, может следовать то, что человек будет склонен описывать тот или иной свой опыт с точки зрения этого цвета, наделяя абстрактные понятия свойством цвета, используя по отношению к ним цветообозначения.

Также Гете разделяет цвета на положительные (красный, оранжевый и желтый), отрицательные (синий, фиолетовый) и нейтральные (зеленый).

В дальнейшем мы попробуем найти подтверждение этой классификации, проследив, действительно ли обладают эти цвета подобной коннотацией.

Бесспорно, существуют такие цветообозначения, которые обладают очень прочно с ними связанными коннотациями. Они получили название «этноэйдем». А.И. Белов отмечал, что «в традиционной русской культуре к цветовым этнойдемам следует прежде всего относить такие цветообозначения, как «красный», «белый», «черный» и «голубой», отчасти «синий» [3, с.49–58]. Итак, мы попытаемся определить, какие цвета являются базовыми, чтобы на их примере рассмотреть проблему метафорических цветообозначений. Прилагательные, являющиеся цветообозначениями, образуют лексико-семантическую группу (ЛСГ). Мы же стараемся выделить ядро этой ЛСГ.

В структуру данной ЛСГ входят:

- монолексемные имена прилагательные («синий», «белый», «красный», «желтый», «зеленый»);
- сложные имена прилагательные, которые являются названием цвета с уточнением его оттенка или интенсивности («ярко-желтый», «розово-красный», «темно-синий», «светло-зеленый»);
- сложные цветообозначения со структурой «сущ. цвет + имя сущ. в им. падеже» («цвет хаки», «цвет коралл»);
- сложные цветообозначения со структурой «сущ. цвет + имя прил. + имя сущ. в им. падеже» («цвет мокрый асфальт», «цвет морской волны») [4].

Чтобы слово принадлежало ядру ЛСГ, необходимо, чтобы оно было базовым понятием. Чтобы данное условие выполнялось, необходимы следующие условия:

- слово должно быть не сложным и непроизводным;
- значение слова должно быть более широким по отношению к слову, выражающему похожее значение;
- слово должно обладать широкой сочетаемостью.

Ученые-физики выделили семь основных спектральных хроматических цветов: красный, оранжевый, желтый, зеленый, голубой, синий, фиолетовый. К ним прибавляются два ахроматических цвета: белый и черный.

По теории базовых цветовых терминов Берлина и Кэя, у всех языков мира существует общий базовый набор слов, обозначающих цвета. В английском языке авторы выделяют 11 базовых прилагательных (красный, оранжевый, желтый, зеленый, синий, фиолетовый, розовый, коричневый, серый, чёрный, и белый), в русском таких прилагательных 12 (ко всем вышеперечисленным в английском языке цветам добавляется голубой).

По нашему предположению, самыми частотными будут являться ахроматические цвета (белый и черный), так как они обладают самой сильной коннотацией «хорошо» — «плохо», что доказывает огромное количество устоявшихся выражений («делить мир на черное и белое», «черный четверг», «белый свет»).

Собрав воедино все теории и утверждения вышеприведенных исследователей, остановимся на двенадцати базовых цветах. Эти 12 цветов будут объектом нашего исследования.

2. Эксперимент

2.1. Постановка задачи

Итак, мы постараемся:

- проверить, какой коннотацией обладают выбранные нами цвета;
- собрать статистику употребления данных цветов;
- проверить наше предположение о частотности ахроматических цветов;

- выявить уникальные, наиболее редкие метафорические цветообозначения.

Среди всех метафор будут выделены синестетические метафоры. Синестетическая метафора — это «перенос наименования на основе сходства ощущения, при котором и исходное, и производное значения слова являются сенсорными» [5, с. 20]. В настоящее время исследователи выделяют несколько видов синестетической метафоры:

- слухо-зрительная;
- зрительно-слуховая;
- слухо-(вкусо)-обонятельная синестезия;
- слухо-вкусовая и зрительно-вкусовая;
- слухо-тактильная;
- графемно-цветовая;
- хроместезия;
- кинестетико-слуховая;
- акустико-тактильная;
- вибрационные, температурные и гравитационные синестетические метафоры.

Нас будут интересовать все метафоры, включающие в себя зрительную перцепцию, то есть какой-либо цвет, называющий что-либо из других любых областей восприятия.

2.2. Инструменты

В данной работе все предположения мы будем доказывать эмпирическим путем, а именно, на примере корпусного исследования.

Под корпусом текстов понимается массив языковых данных, обладающий следующими свойствами:

- репрезентативность (корпус должен отражать все свойства исследуемой области, в достаточном объеме и пропорционально отображать явления, происходящие в языке);
- полнота;
- экономичность;
- наличие металингвистической информации;
- компьютерная поддержка.

Корпуса позволяют применять статистические методы в решениях лингвистических задач, могут быть использованы в качестве экспериментальной базы для проверки гипотез. То есть, с помощью поиска в корпусе можно проиллюстрировать исследование примерами, установить частотность исследуемых единиц языка и подтвердить выдвинутые предположения.

Поиск в корпусе осуществляется с помощью корпусного менеджера. Корпусный менеджер является поисковой системой в данном корпусе, которая обеспечивает поиск данных в корпусе, позволяет получить статистическую информацию и предоставляет результаты пользователю [6].

Наше исследование проводится с помощью поиска в Национальном корпусе русского языка (в дальнейшем — НКРЯ), в поэтическом подкорпусе. Объектом нашего исследования является употребление метафорических цветообозначений, поэтому был взят поэтический подкорпус, так как именно художественная речь обладает большей, по сравнению с другими вариантами языка, образностью и выразительностью, ей в большей степени присущи метафорические обороты. Поэтический подкорпус обеспечен морфологической, семантической и специальной стиховедческой разметками. Тексты (прозаические и поэтические), собранные в поэтическом подкорпусе НКРЯ, были написаны в XVIII–XXI вв., что обеспечивает широкий охват нашего исследования.

Итак, исследование проводилось с помощью поэтического подкорпуса Национального корпуса русского языка.

В качестве инструментов исследования были применены корпусный менеджер НКРЯ и электронная таблица MS Excel 2010.

С помощью корпусного менеджера можно производить поиск, задавая грамматические, семантические и дополнительные (например, чтобы слово было написано с заглавной буквы или перед точкой) признаки, или просто ввести слово. Можно производить поиск точных форм, задать подкорпус определенного автора, искать только среди произведений, написанных лицами мужского пола и т.д.

3. Методика исследования

3.1. Шаг первый

В поэтическом подкорпусе НКРЯ осуществляется поиск с заданием следующих условий: первое слово должно обладать грамматическим признаком «прилагательное» и семантическим признаком «цвет», а второе слово, согласующееся с ним, должно обладать грамматическим признаком «существительное» и семантическим признаком «непредметные». Рассматриваются сочетания именно с непредметными (абстрактными) существительными, так как предполагается, что цвет — это качество, свойство, присущее видимым объектам, если же оно присваивается абстрактным предметам, то можно предположить, что наблюдается метафорическое употребление.

Необходимо заметить, что в число непредметных существительных попадают существительные с семантическим признаком «свет» («окраска, колорит, желтизна, прозелень») и «природное явление» («зарница, вьюга, зной»). Нами предполагается, что такие существительные действительно могут иметь цвет. Проведем эксперимент и проверим, действительно ли это так.

В точности повторяем вышеизложенный пункт 1, за исключением того, что существительным присваиваются только признаки «свет» и «природное явление».

По нашему запросу найдено 3228 вхождений. Настраиваем выдачу результатов в формате KWIC. Полученные результаты копируем в электронную таблицу Excel.

На рисунке 1 видно, что из общего количества вхождений слов, которые вряд ли могут иметь собственный цвет, оказывается очень мало (примерно 2%).

Названия строк	Сумма по полю 2
свет	9
зной	9
холод	8
жар	7
дождь	7
ветер	6
гром	5
тьма	5
непогодь	3
метель	2
вихрь	1
погода	1
порыв	1
пурга	1
разлив	1
распутица	1
сквозняк	1
сумерки	1
закат	1
шквал	1
Общий итог	**71**

Рис. 1. Демонстрация слов со значениями «свет» и «природное явление»

Следовательно, словами с признаками «свет» и «природные явления» можно пренебречь.

3.2. Шаг второй

Результаты, полученные по нашему запросу, копируем в электронную таблицу Excel. Далее анализируется полученный материал. Поскольку в НКРЯ присутствует семантическая неоднозначность, может оказаться так, что не все полученные в результате поиска примеры будут соответствовать необходимым параметрам. Например, встречаются такие сочетания, как «красным горя» (деепричастие имеет форму родительного падежа слова «горе»), «белая горячка», «ультра-фиолетовый луч», форма именительного падежа слова «сера» совпадает с кратким прилагательным женского рода «сера», «желтая лихорадка», «черная оспа» и т.д. Поэтому выявлять нужные сочетания необходимо вручную.

Всего из изначально полученных 12582 вхождений подходящими для нашего исследования остались 5956 вхождений (из всех цветообозначений, полученных по нашему запросу, остались лишь 12 базовых цветов). После ручной обработки данных были оставлены 1076 сочетаний.

3.3. Шаг третий

Частотность употребления всех цветообозначений распределилась следующим образом (рис. 2).

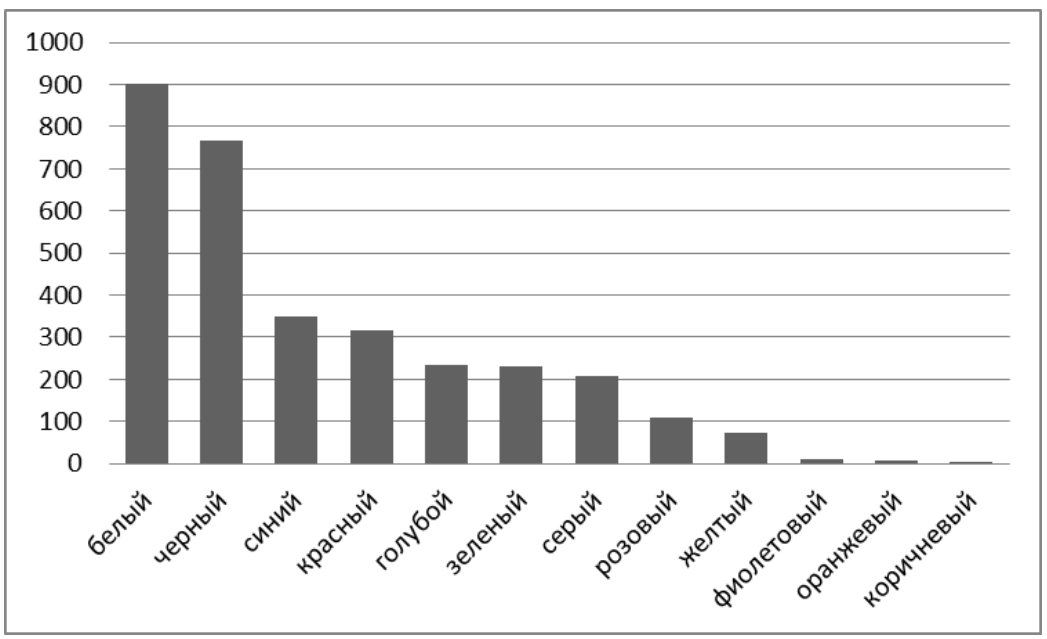

Рис. 2. Частотность употребления всех цветообозначений

Как мы видим, наше предположение о том, что наиболее частотными будут ахроматические цвета, подтвердилось. Наименее частотными оказались цветообозначения «фиолетовый», «оранжевый» и «коричневый».

Для каждого базового цвета были выявлены 10 самых частотных сочетаний, что будет продемонстрировано на диаграммах ниже (рис. 3-12).

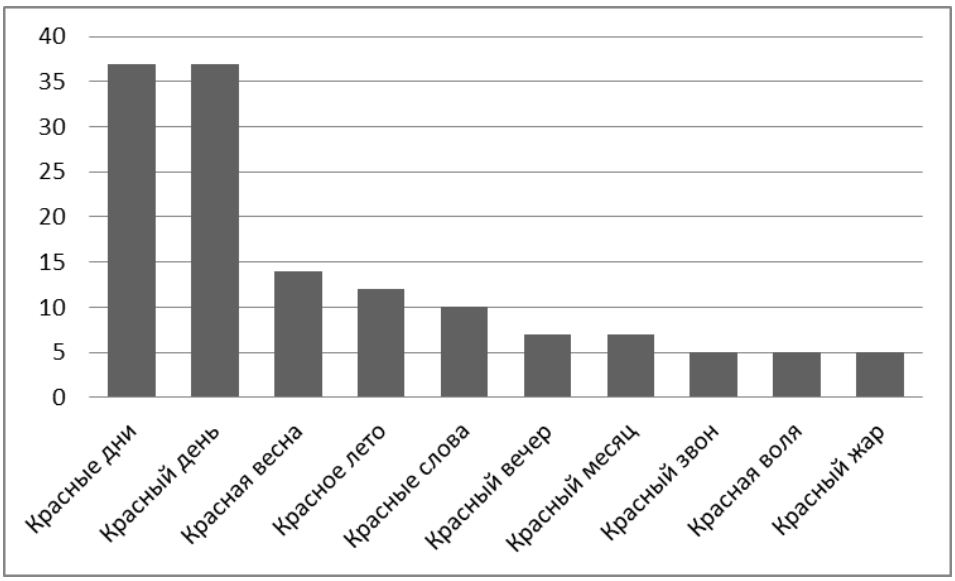

Рис. 3. Красный цвет

Как видно на диаграмме, самыми частотными оказались метафорические сочетания, в которых существительное имеет значения «время суток» и «время года». Очень частотными оказались сочетания, связанные с военной тематикой, например «красный лозунг», «красный парад», «красная присяга», «красный Октябрь», «красная борьба». Менее частотными оказались сочетания с негативным значением («красный ад», «красная гибель», «красная ярость»).

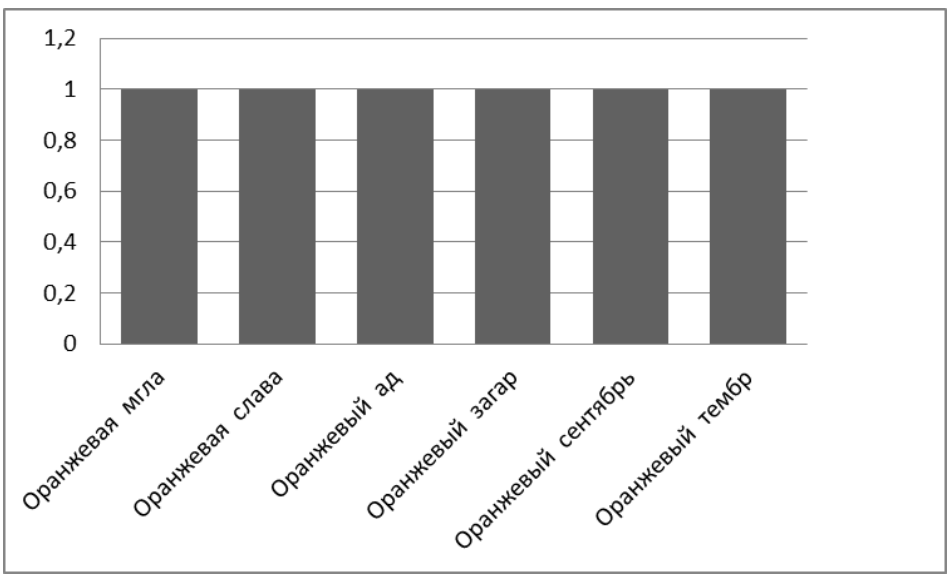

Рис. 4. Оранжевый цвет

Из 6 полученных сочетаний только одно содержит существительное, обладающее негативной коннотацией – «ад».

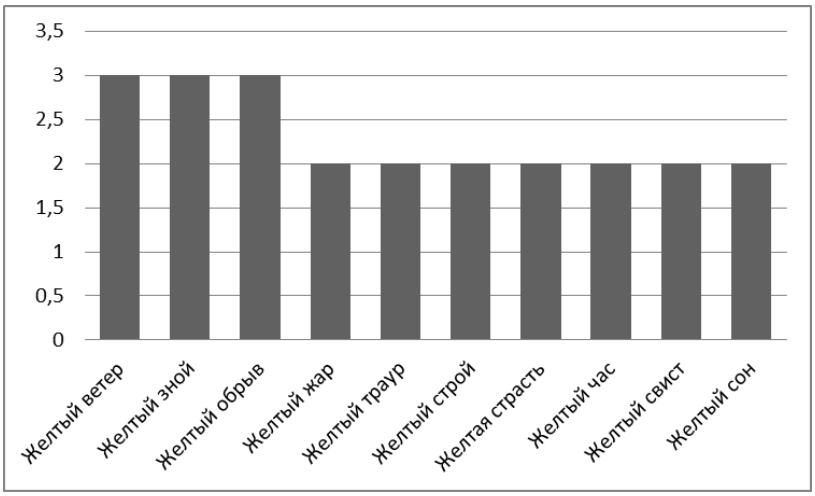

Рис. 5. Желтый цвет

Из 10 существительных, входящих в сочетания со словом «желтый», только одно обладает негативной коннотацией — «траур». Остальные являются нейтральными. Менее частотными, но многочисленными оказались слова с негативной коннотацией: «желтый тлен», «желтый сплин», «желтая скука», «желтая зависть», «желтая ложь». Если слово «красный» встречается в сочетаниях с негативным оттенком, то, в большинстве случаев, эти сочетания обозначают сильные, бурные эмоции (например, ярость или гнев). Что нельзя сказать о слове «желтый».

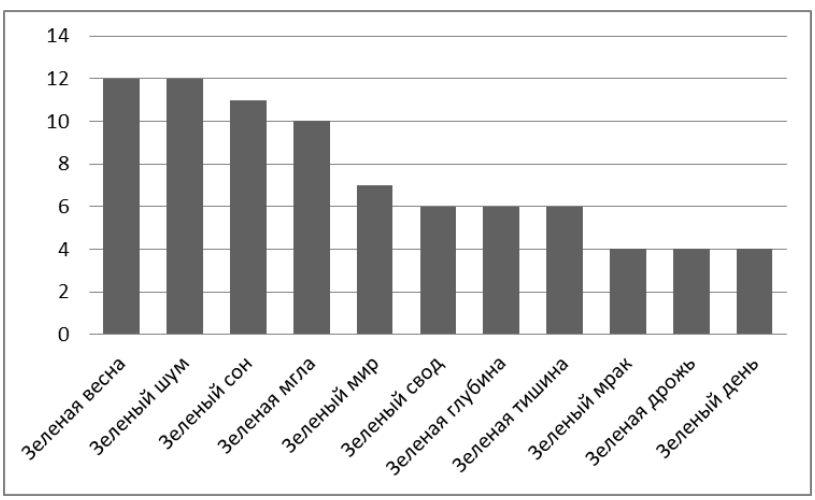

Рис. 6. Зеленый цвет

В сочетании со словом «зеленый» самыми частотными оказались слова, имеющие отношение к природе. Кроме того, частотными оказались сочетания с существительными, обозначающими месяца (апрель, июнь, июль, май). В основном, слово «зеленый» встречается в сочетаниях, имеющих положительное значение («зеленые надежд флаги», «зеленое празднество», «зеленая юность»).

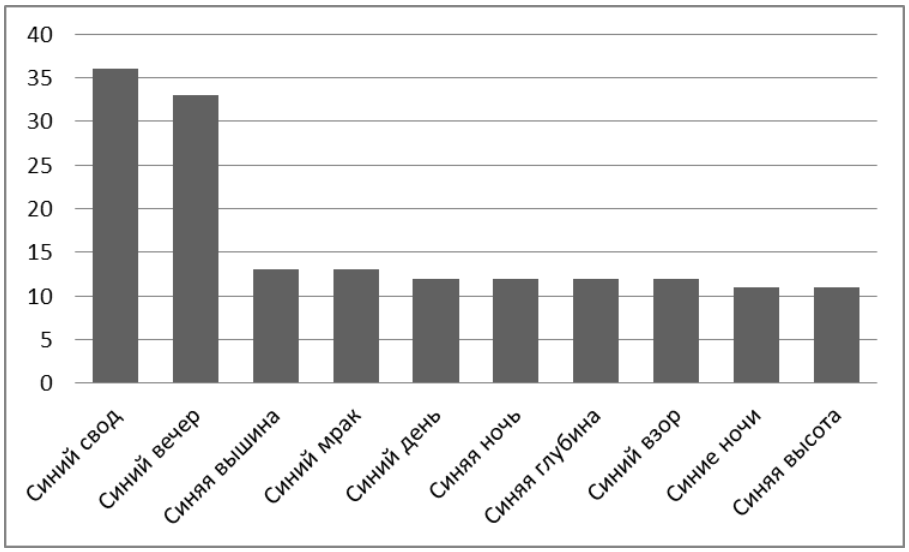

Рис. 7. Синий цвет

Синий цвет оказался частотным в сочетаниях со словами, обозначающими время суток и размер (высота, глубина) какого-либо объекта. В основном, сочетания со словом «синий» имеют оттенки «задумчивость», «загадочность», «меланхоличность» («синяя безбрежность», «синяя бездонность», «синяя загадка», «синее счастье», «синяя вечность»).

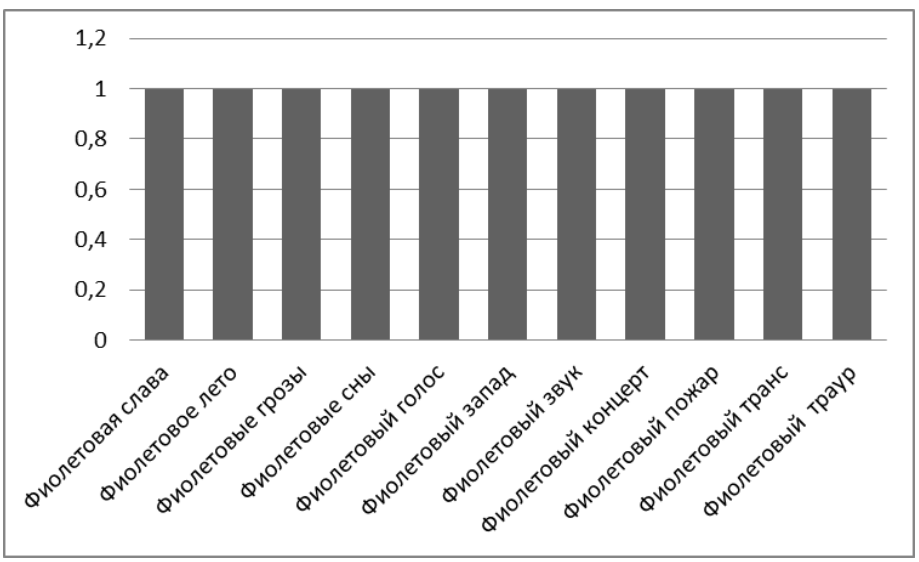

Рис. 8. Фиолетовый цвет

Из всех существительных, входящих в метафорические сочетания со словом «фиолетовый», негативный оттенок имеет только одно — «траур». 2 существительных из 11 в сочетании со словом «фиолетовый» составляют синестетические метафоры: «фиолетовый голос» и «фиолетовый звук».

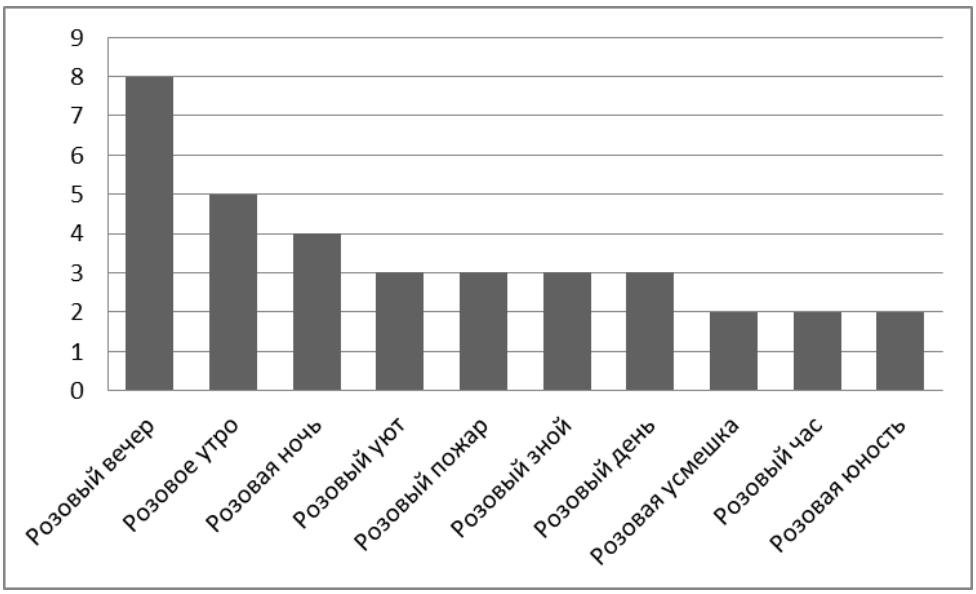

Рис. 9. Розовый цвет

Три самых частотных сочетания со словом «розовый» обозначают время суток. Большое количество существительных в сочетании с цветообозначением «розовый» образуют метафоры с различными положительными оттенками значения («розовая свежесть», «розовая любовь», «розовые мечты», «розовое детство»).

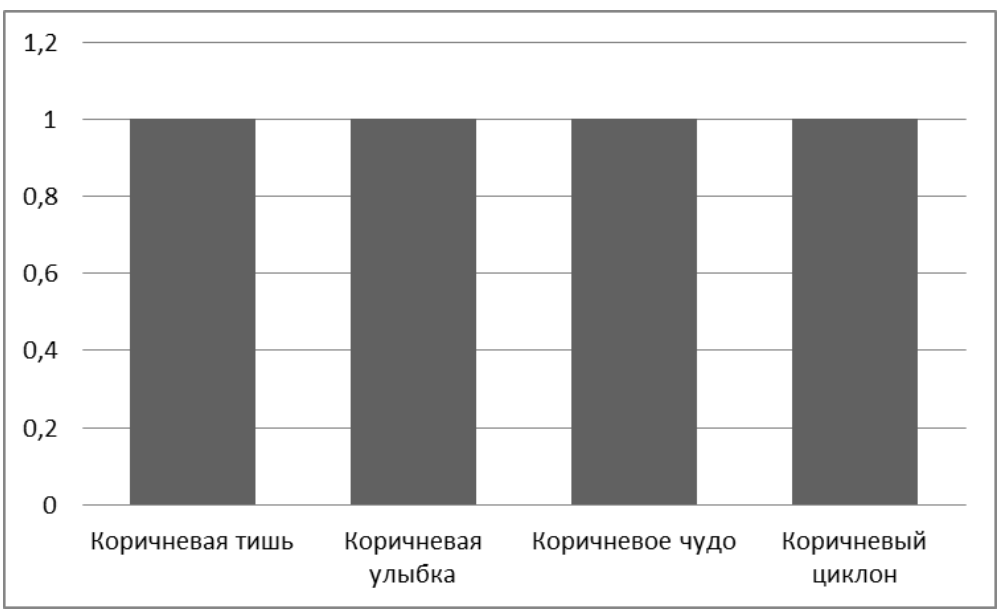

Рис. 10. Коричневый цвет

Цветообозначение «коричневый» оказалось наименее частотным.

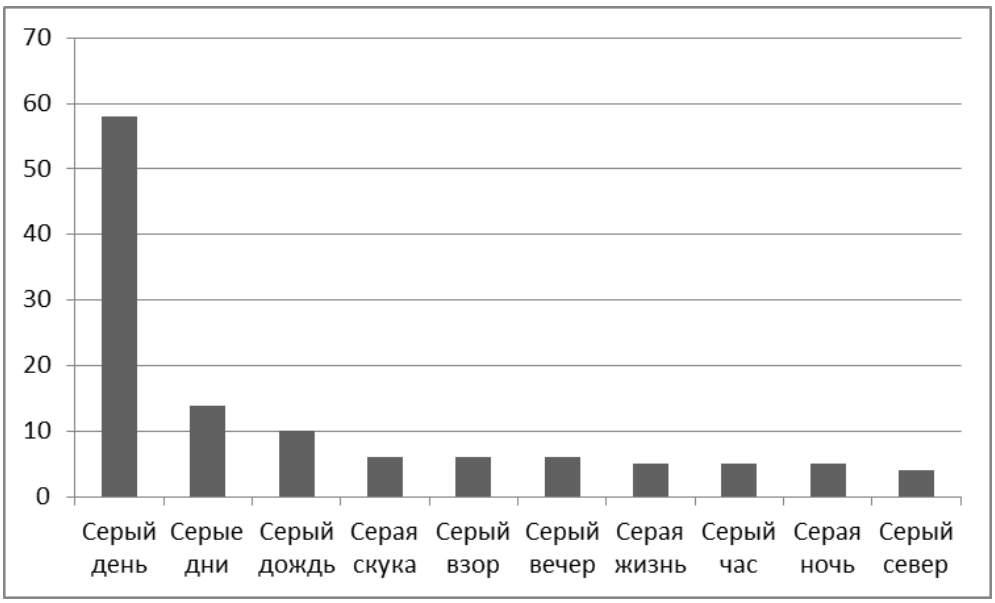

Рис. 11. Серый цвет

Самыми частотными сочетаниями со словом «серый» оказались, опять же, обозначения времени суток или отрезка времени. Негативные сочетания оказались весьма многочисленными («серая тоска», «серые смерти», «серый террор», «серая сирость»).

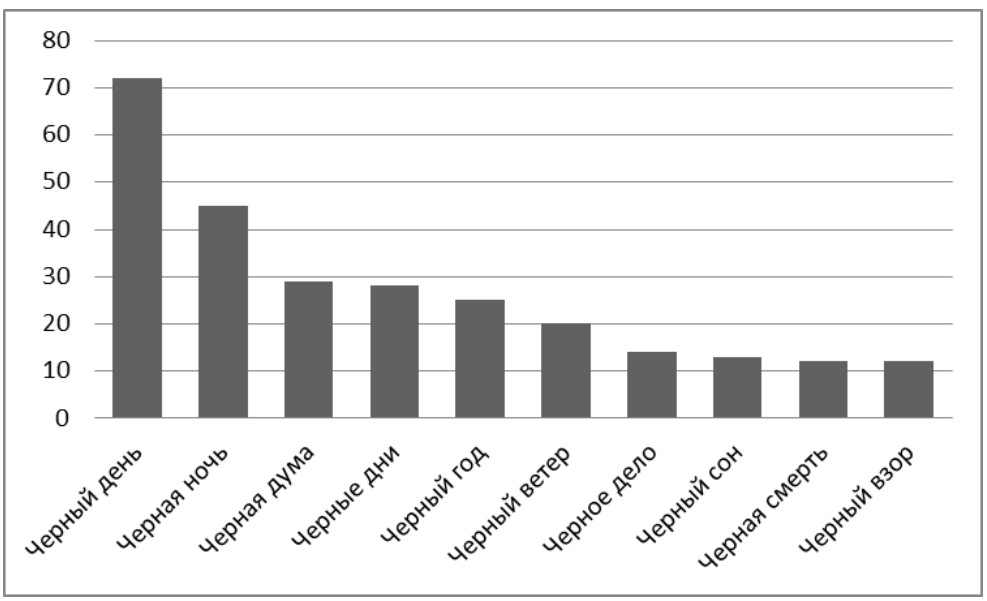

Рис. 12. Черный цвет

Сразу же после 10 самых частотных сочетаний со словом «черный» идут сочетания с сильным негативным значением («черная зависть», «черная тоска», «черный ад», «черное горе», «черная беда»).

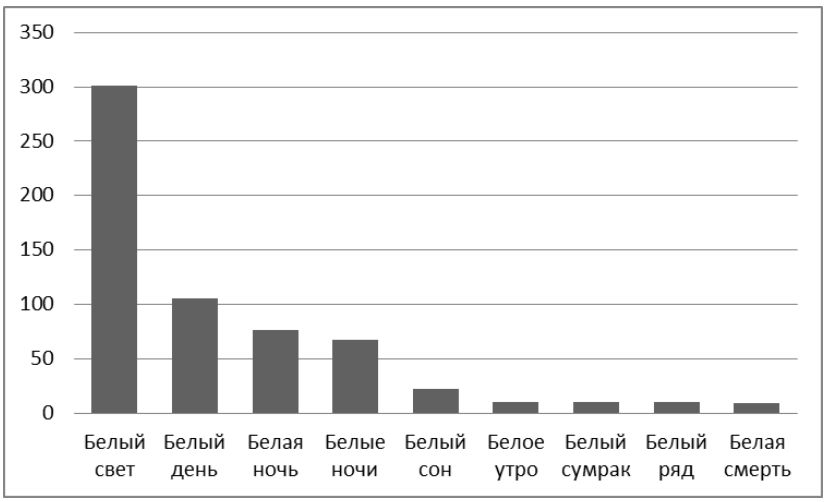

Рис. 13. Белый цвет

Самым частотным среди всех сочетаний со всеми цветами оказалось сочетание «белый свет» (рис. 13). Белый цвет является отсутствием цвета, поэтому и сочетания, содержащие цветообозначение «белый», являются нейтральными (в большинстве случаев). Также близким к слову «белый» по значению является слово «светлый». В этих случаях сочетания приобретают преимущественно положительное значение.

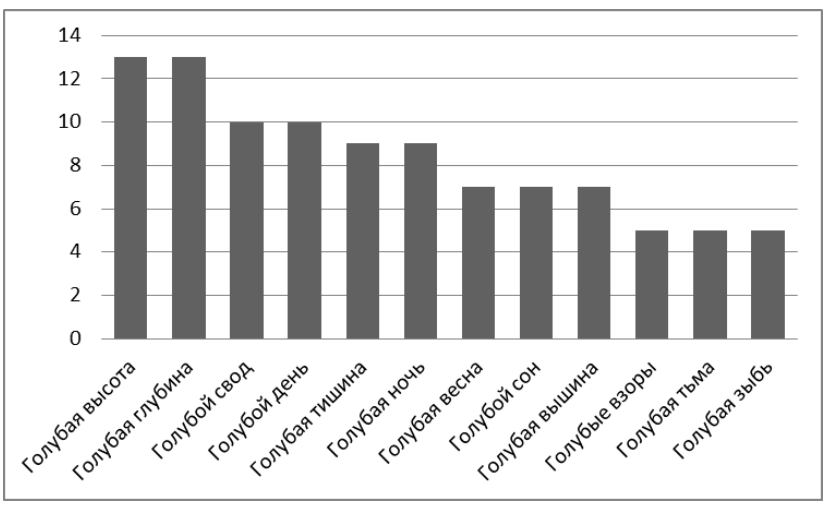

Рис. 14. Голубой цвет

Цветообозначение «голубой» (рис. 14) имеет подобное с цветообозначением «синий» значение — размер (высота, глубина) какого-либо объекта. Также слово «голубой» часто встречается в сочетаниях, иносказательно называющих небо, воду и холод («голубое пространство волн», «голубой загар небес», «голубые прорывы робы неба», «голубая стужа»).

Далее на диаграммах (рис. 15-23) будет продемонстрировано процентное соотношение следующих коннотаций для цветообозначений: положительная, отрицательная, нейтральная и «время года, время суток». Цвета «коричневый», «оранжевый» и

«фиолетовый» не будут проанализированы, так как количества их словоупотреблений недостаточно, чтобы сделать какой-либо вывод. Почти у всех цветов большее количество процентов имеет нейтральная коннотация. Поэтому те цвета, у которых негативное значение значительно превышает в процентном соотношении положительное значение, будем считать цветами, обладающими негативной коннотацией (и наоборот). В том случае, если процентное соотношение примерно одинаковое, будем считать, что цвет обладает нейтральной коннотацией.

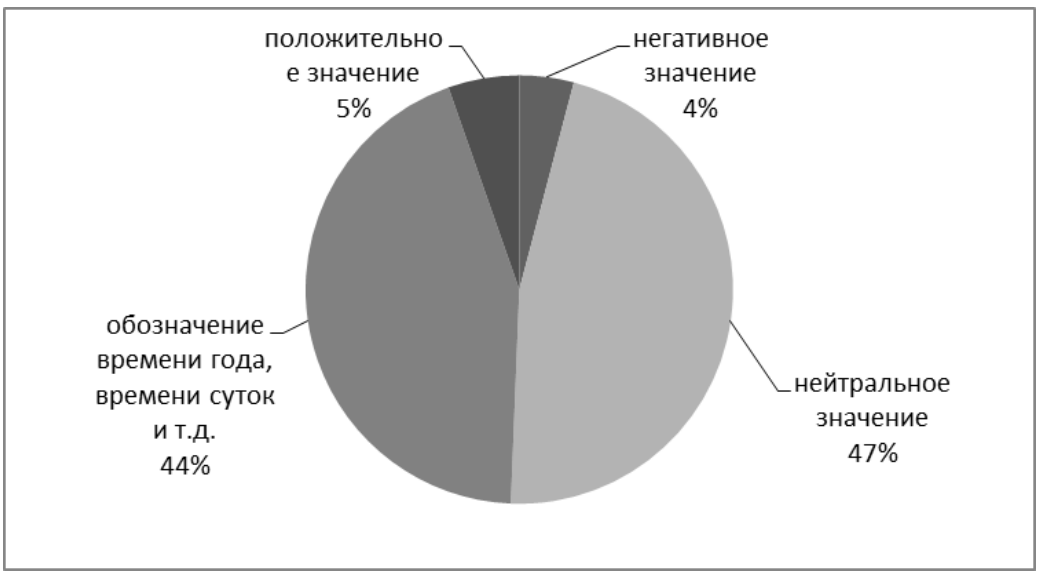

Рис. 15. Коннотации для красного цвета

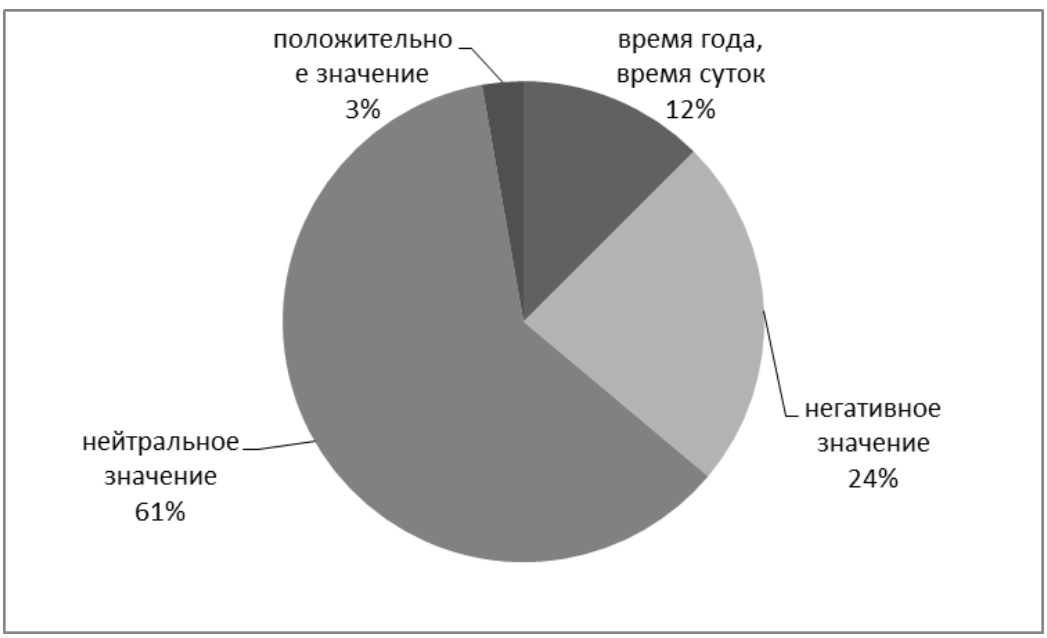

Рис. 16. Коннотации для желтого цвета

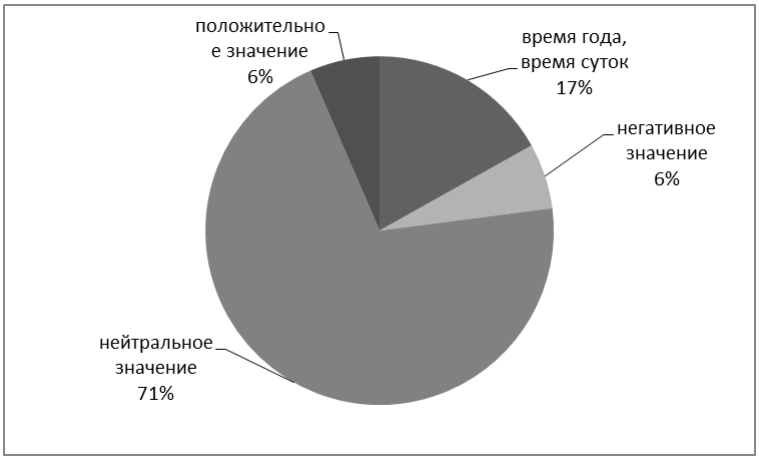

Рис. 17. Коннотации для зеленого цвета

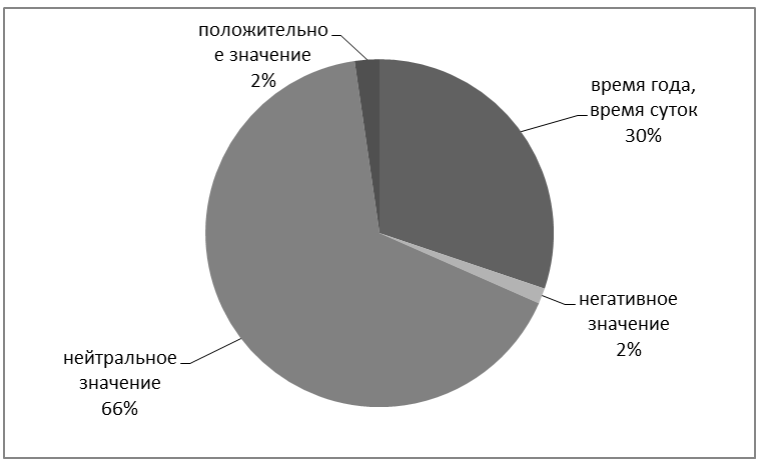

Рис. 18. Коннотации для синего цвета

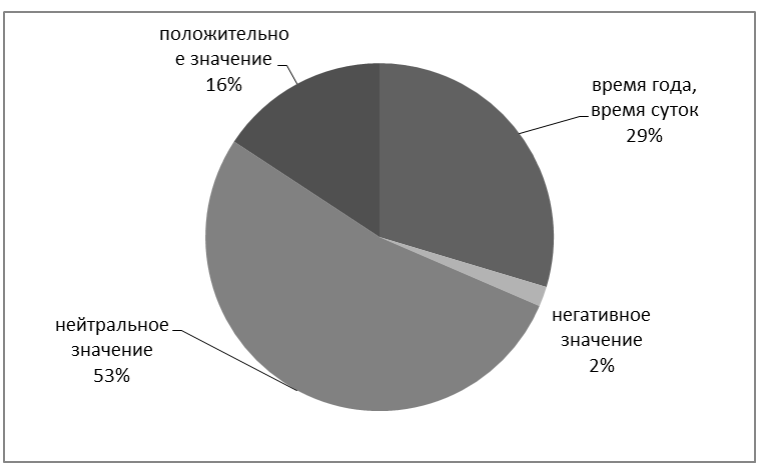

Рис. 19. Коннотации для розового цвета

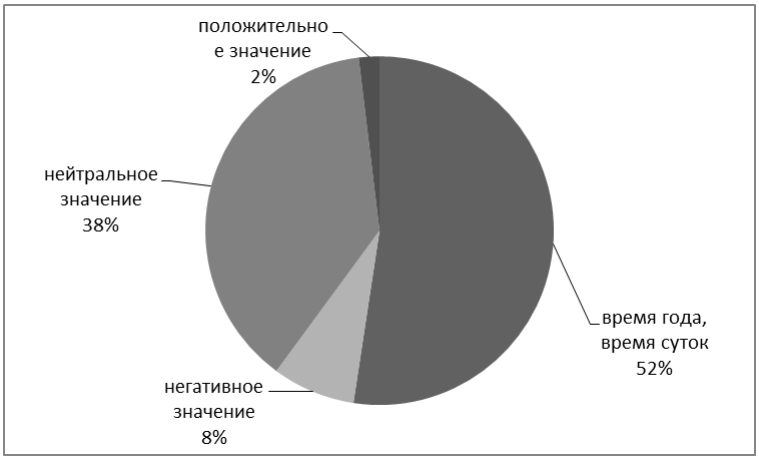

Рис. 20. Коннотации для серого цвета

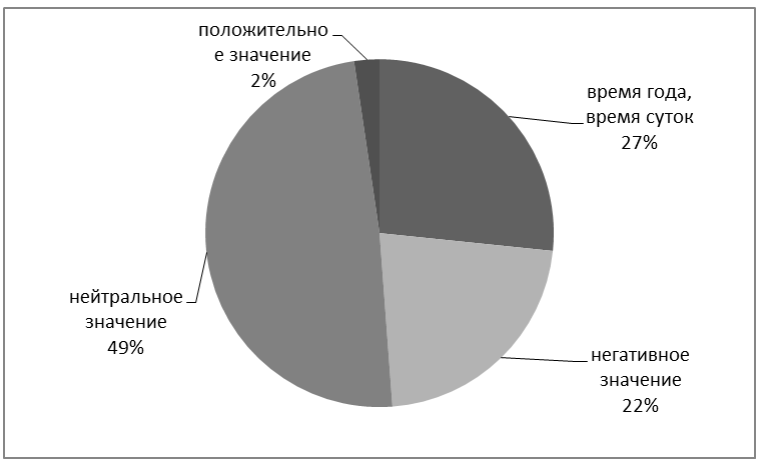

Рис. 21. Коннотации для черного цвета

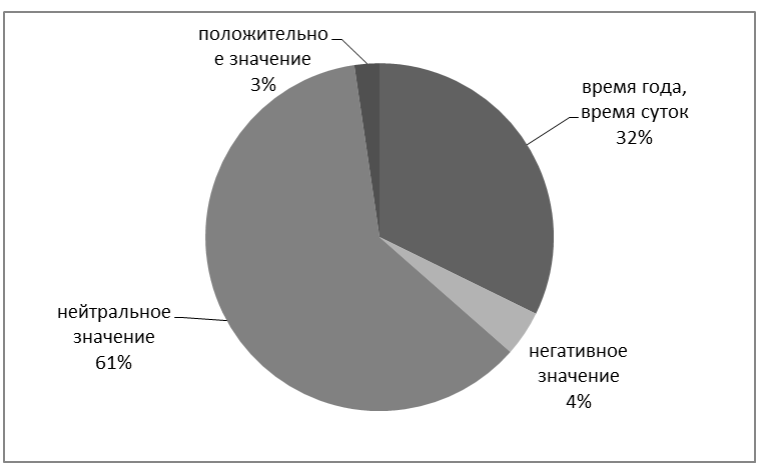

Рис. 22. Коннотации для белого цвета

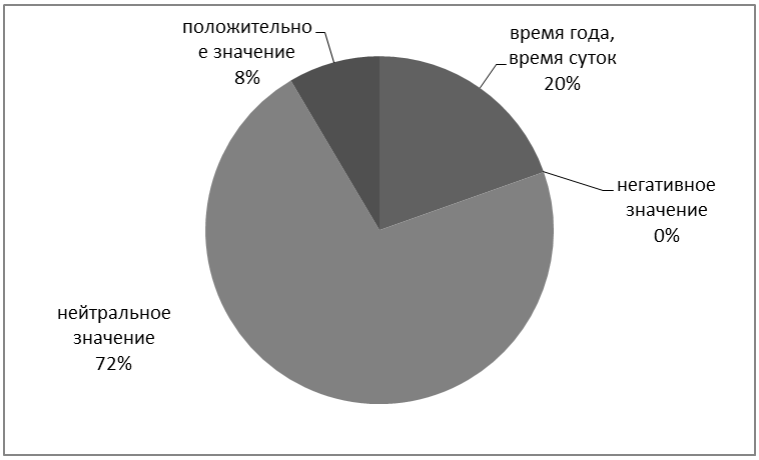

Рис. 23. Коннотации для голубого цвета

Теперь продемонстрируем на диаграмме (рис. 24), как распределяется частотность существительных, входящих в синестетические метафорические сочетания с цветообозначениями.

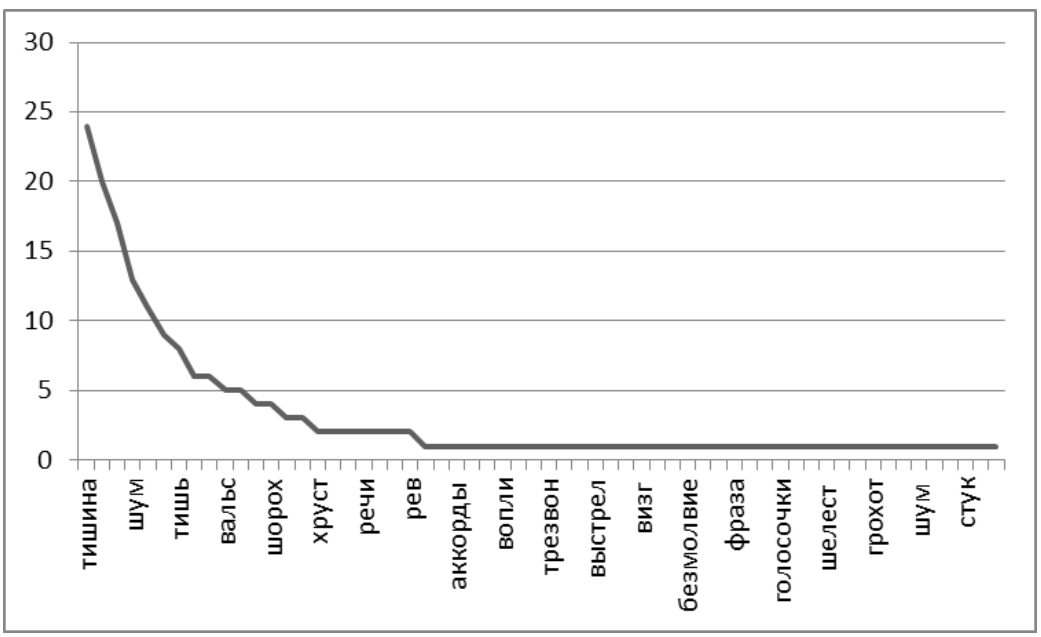

Рис. 24. Синестетические метафоры

Из всех видов синестетических метафор нам удалось найти лишь слухо-зрительные метафоры. «Тишина», «тишь», «безмолвие», по нашему мнению, могут в сочетании с цветообозначением составлять синестетическую метафору, несмотря на то, что они имеют значение «отсутствие звука», так как все равно речь идет о слуховом акте.

Всего было найдено 190 синестетических метафор из 1076 сочетаний. Как мы видим, синестетических метафор не так много, меньше 20% от общего числа найденных метафорических сочетаний.

4. Заключение

Все прилагательные, рассмотренные нами в данной работе, являются 12 базовыми цветообозначениями по теории Берлина и Кэя и имеют семантический признак «цвет» в поэтическом корпусе НКРЯ.

Утверждение Гете о коннотации 6 из 12 базовых цветов (красный, оранжевый, желтый, синий, фиолетовый и зеленый) подтвердилось не полностью. Согласно проведенному нами эксперименту, коннотация цветообозначений выглядит следующим образом:

- положительная коннотация свойственна цветам «голубой» и «розовый»;
- негативная коннотация свойственна цветам «черный», «серый» и «желтый»;
- нейтральная коннотация свойственна цветам «красный», «белый», «синий» и «зеленый».

Порядок снижения частотности исследованных прилагательных выглядит следующим образом: белый, черный, синий, красный, голубой, зеленый, серый, розовый, желтый, фиолетовый, оранжевый, коричневый.

Среди всех найденных метафорических сочетаний с цветообозначениями были найдены синестетические метафоры «слух-зрение» (примерно 18% от всего количества найденных метафорических сочетаний).

В приложении (рис. 25) приводятся метафорические сочетания, которые, с точки зрения автора, являются уникальными.

В дальнейших исследованиях планируется сравнить классификацию цветообозначений М. Люшера с полученным автором результатом. Также будут подробнее рассматриваться и изучаться синестетические метафоры различных видов.

5. Приложение

Смежают звезды	**голубые**	**веки**, И мачты корабельные скрипят
опасность все ближе, Двух **зрачков**	**голубые**	**выстрелы.**
срез, Запрокинутое ввысь лицо, В	**голубой**	**загар небес**?
в лицо мне Светлых **глаз**	**голубой**	**прилив.**
Здесь **травы теряют**	**зеленую**	**речь** И камни синеют от
прордело, и — **закат покрылся в**	**красный**	**кашель.**
самых нежных нежитей Засмеется **из**	**красной**	**трясины** ваших топких **губ.**
душу сжигаешь, как ночь, **На**	**белом**	**рассвете клавиш.**
А на закат наложен Был	**белый**	**траур черемух**, Что осыпался мелким
За	**белой**	**чащею бровей** Зажглись его глаза
солнечный песок, **мигнул и щелкнул**	**черным**	**веком фотографический глазок.**
бедным запахом домашней глажки, И	**черным**	**маникюром площадей.**
розах серая ограда, И **синий,**	**синий**	**плен очей...**
Город спятил. Людям надоели **Платья**	**серых**	**будней — пиджаки**. Люди тряпки пестрые
как зудом, Я дрожал **перед**	**серою**	**истиной глаз.**
переулков столице, В столице, /Покрытой	**серой**	**оберткой снегов**, Копошатся ночные лица
толкается боком, виснет росинка **на**	**розовой**	**мочке зари**,
Слушая	**розовый**	**сумрак смуглых ладоней**, Теплую музыку
отлетел твой гений; А **визги**	**желтой**	**клеветы Глупцов**, которые марали, Как
Довольно-таки весен, зим, лет. И	**желтый**	**осени билет**. На музыку спустя
Дымом половодье Зализало ил.	**Желтые**	**поводья Месяц уронил**. Еду на

Рис. 25. Уникальные метафоры

Литература

[1] Алымова Е.Н. Ассоциативная лингвоцветовая картина мира // Вестник Санкт-Петербургского университета. 2007. Вып.2 С. 74–76.
[2] Гете И.В. Избранные сочинения по естествознанию. М., 1957.
[3] Белов А.И. Цветовые этноэйдемы как объект этнопсихолингвистики // Этнопсихолингвистика. М: Наука, 1988. С. 49-58
[4] Косых Е.А. Система цветообозначений в русском языке: к созданию и публикации «Русской энциклопедии света» // Вестн. Барнаул. гос. пед. ун-та. Сер. «Психолого-педагогические науки». 2002. №2.
[5] Левчина И.Б. Развитие семантической структуры синестезических прилагательных. СПб, 2003. С.20.
[6] Захаров В.П., Богданова С.Ю. Корпусная лингвистика. СПб, 2013.

Distinctive Metaphorical Phrases Denoting Colors (Corpus Analysis)

A.V. Chekmeneva

Saint Petersburg State University

This article is concerned with the issue of the use of Russian metaphorical phrases that denote colors.
The author considered various theories, namely the color theory by Goethe, Basic color terms by Berlin and Kay, the theory by Belov. By combining different theories together there were selected 12 basic colors. The correlation between abstract nouns and adjectives that denote colors is being studied. There is an attempt to define the connotation of every basic color. All colors are divided into colors with positive, negative and neutral connotation.
The author also extracts synesthetic metaphors from all found metaphors.
The study is conducted in terms of corpus research. All the examples are provided due to National Russian Corpus.
In the further work the study will be concerned with synesthetic metaphors of sound and visual perception, with the classification by Max Luscher. The further work will be accompanied by setting up a unique corpora with the help of Sketch Engine tools.

Keywords: corpus linguistics, frequency of lexical units, Russian National Corpus, colors, Berlin and Kay theory, basic color terms, synesthetic metaphors

Корпусное исследование прилагательных лексико-семантической группы «цвет» в русском языке

Е.А. Шукшина

Санкт-Петербургский государственный университет

elena.shukshina@gmail.com

Аннотация

В работе представлена классификация прилагательных лексико-семантической группы «цвет» на основании количественных характеристик, полученных путем корпусного и лексикографического анализа: частота слова в основном корпусе Национального корпуса русского языка, процент употреблений в «цветном» значении, количество значений, выделяемых Большим толковым словарем русского языка.

В результате анализа более ста цветонаименований четко выделяются четыре группы прилагательных: основные цвета, оттенки (слова, использующиеся только в значении цвета), относительные цвета (относительные прилагательные, сравнительно редко употребляющиеся в «цветном» значении), относительные оттенки (относительные прилагательные, чаще употребляющиеся в значении цвета, чем в других значениях).

Представленный способ классификации позволяет объективно оценить состав и состояние «цветной» лексики в русском языке и предположить вектор ее дальнейшего развития.

Ключевые слова: корпусная лингвистика, цветовые термины, названия цветов, русский язык, лексико-семантическая группа

1. Введение

Цветонаименования в русском языке подвергались изучению с различных точек зрения на протяжении десятилетий. Существуют работы, посвященные описанию их состава, стилистических функций и семантической структуры, рассматривающие цветообозначения в психолингвистическом и социолингвистическом плане. Однако до сих пор лингвисты не пришли к единому мнению о классификации цветонаименований в русском языке. Количество выделяемых групп и их состав варьируются у разных авторов:

И.В. Макеенко [1] подразделяет цветовые обозначения на основные, к которым относятся ахроматические цвета (*белый, черный, серый*), цвета радуги, *коричневый* и *розовый*, и оттеночные, которые в свою очередь подразделяются на оттеночные слова без ясно прослеживаемой этимологии (*сизый, бурый, алый*), слова вторичной номинации (*абрикосовый, изумрудный*), контекстно ограниченные цветообозначения (*карий, русый, гнедой*).

Поддерживает выделение 12 основных цветов Р.М. Фрумкина [2, с.64–85], выделяя также в ходе психолингвистического эксперимента 32 «базовых» названия цветов, известных всем носителям русского языка (*алый, вишневый, бордовый, малиновый, морковный* и т.д.).

Б. Берлин и П. Кэй [3] также называют 12 основных цветовых терминов для славянских языков. Согласно их теории, основные цветовые термины должны быть непроизводными, несоставными, иметь широкую сочетаемость, не должны являться оттенками других

основных цветов (например, *алый* — оттенок красного), и должны восприниматься носителями языка как психологически выделяющиеся, значимые (salient).

Другой взгляд на проблему выделения основных цветов имеет В.И. Иваровская [4], которая на основе словарных определений выделяет 10 ядерных цветовых прилагательных, которые являются ядрами семантических полей, включающих все оттенки одного цвета. Таким образом, *розовый* и *голубой* не составляют отдельных полей, а включаются в поля для *красного* и *синего* цветов соответственно.

В части перечисленных работ в качестве цветообозначений рассматриваются единицы разных уровней: словосочетания, слова разных частей речи. В нашей работе мы ограничимся анализом прилагательных со значением цвета.

Целью работы является представление способа классификации прилагательных данной семантической группы, основанного на количественных характеристиках слов: частотности употребления в Национальном корпусе русского языка, проценте употребления слова в «цветном» значении, количестве значений, выделяемом авторитетным словарем (Большим толковым словарем русского языка [5]). Числовые показатели способны разрешить спор о принадлежности отдельных лексических единиц к той или иной группе, в том числе определить состав основных цветонаименований в русском языке. Представленная классификация может использоваться для оценки современного состояния лексико-семантического поля «цвет» в русском языке, а также наблюдения изменений в группировании слов и развития ЛСП в целом.

2. Материал исследования

Для исследования был взят объединенный список цветонаименований, собранных из двух источников. В него вошли леммы, найденные по семантическому признаку «цвет» в Национальном корпусе русского языка, и слова, вошедшие в каталог цветов [6, с.116–126].

Из списка были удалены повторы, существительные и обозначения цветов, образованные с помощью словосложения (например, *ярко-зеленый, красно-оранжевый*) и суффиксации от других слов, обозначающих цвета (например, *аленький*). Далее слова были разделены на 5 групп в соответствии с морфемной структурой и лексическим значением:

- основные цвета;
- оттенки;
- относительные цвета;
- «квази-цвета» (обозначающие не собственно цвет, а его характеристики — насыщенность, узор и т.д., например, *яркий, светлый, пятнистый*);
- прилагательные, обозначающие цвет определенного предмета, чаще всего части тела или шерсти животного (*голубоглазый, пегий*).

Эти группы примерно соответствуют классификации И.В. Макеенко [1]. Последние две группы в исследовании не рассматриваются, так как словарные статьи относящихся к ним слов вполне однозначно указывают на принадлежность к этим двум группам. Определения «квазицветов» не будут отсылать к какому-либо конкретному цвету спектра, определения слов пятой группы будут содержать пометы, указывающие на контекстные ограничения в употреблении слова (например, «о масти лошади», «о волосах»). В ходе исследования список дополнялся (были включены слова *салатовый, шафранный, шафрановый, сапфирный, охристый*).

Для первых трех групп процент употребления в «цветном» значении высчитывался как количество употреблений на 100 случайных контекстов в Национальном корпусе русского языка (использовались следующие настройки выдачи: упорядочить: случайно, документов на странице: 100, примеров в документе: 1). Для слов, число употреблений которых было меньше ста, приводится процентное соотношение «цветных» употреблений ко всем.

Процент «цветного» значения для слов, у которых в словаре указано только «цветное» значение, не подсчитывался и приравнивался к 100%.

Поиск по корпусу производился с применением фильтра по грамматическим признакам: прилагательное, полная форма (A,plen); Это было необходимо для отсева результатов с неснятой неоднозначностью, где существительное в косвенном падеже в одном из вариантов разбора трактовалось как краткая форма прилагательного (*кремов, абрикосов*). Имена собственные (фамилии и названия населенных пунктов), встречавшиеся в выдачи, также не были отсеяны.

Ноль в таблицах в столбце «Значений в БТС» означает, что для данного слова нет отдельной словарной статьи в словаре. Некоторые малочастотные прилагательные при этом указаны в словарной статье существительного, от которого они образованы (*антрацитовый, пурпурный, пурпуровый*), другие в словаре не встречаются вовсе (*шарлаховый, рубинный, кумачный*).

3. Закономерности внутри выделенных групп

После первичной обработки списка слов и проведения корпусного и лексикографического анализа списки слов были уточнены, полученные данные обобщены для каждой из групп.

3.1. Основные цвета

К этой группе было отнесено максимальное число цветонаименований, выделяемых различными исследователями в группу основных цветов. Как можно заметить из таблицы, лексемы *голубой* и *розовый*, статус которых оспаривается некоторыми исследователями, с точки зрения количественных характеристик являются типичными членами группы основных цветов и четко отделены от остальных цветообозначений, как мы увидим ниже.

Корпусный и лексикографический анализ показали, что цветам этой группы характерна относительно высокая частотность (c/y>3000 или ipm>11). Процент употребления в цветном значении и количество значений, указанных в словаре сильно варьируется, причем с числом словоупотреблений растет количество значений и падает процент употребления в цветном значении. Очевидно, переносные значения слова забирают на себя часть употреблений. Согласно А.А. Брагиной [7] развитие переносных значений у слов с устоявшимся цветным значением естественно и является одним из основных этапов развития цветовых прилагательных. Такой процесс возможен вследствие окказионального, образного употребления слов в новых контекстах и закрепления некоторых таких употреблений в языке. Поэтому не удивительно, что с частотностью слов увеличивается и количество их переносных значений.

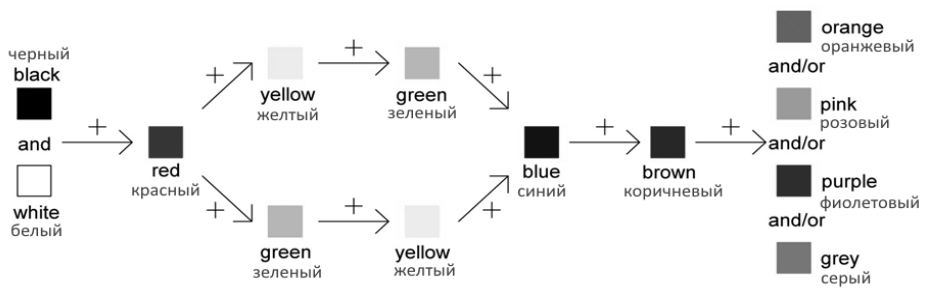

Рис.1. Порядок появления цветовых терминов по теории Берлина и Кэя

Порядок базовых терминов, отсортированных по частоте, количеству значений в словаре или проценту цветного значения соотносится с порядком появления цветовых терминов в языках мира (рис. 1.), описанным Берлином и Кэйем [3]. В таблице 1 показано,

что цвета *черный, красный* и *белый* находятся в первой тройке по числу значений и частоте, слова *коричневый*, *оранжевый, фиолетовый* — в последней.

Таблица 1. Базовые прилагательные со значением цвета в основном корпусе НКРЯ, отсортированные по количеству значений в БТС

Лемма	Число с/у в НКРЯ	ipm (instances per million)	% «цветного» значения	Число значений в БТС
черный	103613	365,57	65	12
красный	83620	295,03	63	8
белый	106945	377,32	72	6
зеленый	36573	129,04	80	6
серый	34125	120,40	82	5
розовый	15609	55,07	93	5
голубой	21916	77,32	92	3
желтый	15381	89,55	97	2
синий	27481	96,96	100	2 (оба цветные)
коричневый	6724	23,72	100	1
оранжевый	3731	13,16	100	1
фиолетовый	3166	11,17	100	1

3.2. Оттенки

Изначально к этой группе были отнесены прилагательные, имеющие непроизводную основу или образованные от названия цвета (*пурпурный, лазурный*). В процессе лексикографического анализа к этой группе присоединились ранее производные цвета, в современности, по-видимому, потерявшие связь с мотивирующим словом (*бежевый, бордовый, червленый*). Для них в словаре указано только одно, цветное значение. Некоторые другие прилагательные, имеющие только цветное значение, по-прежнему структурно отсылают нас к соответствующему существительному (*дымчатый, шафрановый*). Логично и эти слова отнести к группе оттенков. В таком случае немотивированность для слов этой группы станет частой сопутствующей, но не обязательной характеристикой.

Общими характеристиками для слов этой группы (табл. 2), таким образом, являются большой разброс по частоте, ограниченный сверху 14 ipm, наличие только цветного значения (одного или нескольких, как у слова *бурый*).

Таблица 2. Оттенки, отсортированные по частоте

Лемма	Число с/у в НКРЯ	ipm	Число значений в БТС
алый	3973	14,02	1
бурый	3450	12,17	3
лиловый	3212	11,33	1
багровый	2 903	10,24	1
сизый	2562	9,04	1
пунцовый	962	3,39	1
лазурный	785	2,77	1
пурпурный	713	2,52	0
багряный	553	1,95	1
бежевый	548	1,93	1
бордовый	496	1,75	1
палевый	445	1,57	1
лазоревый	389	1,37	1
пурпуровый	319	1,13	0
червленый	132	0,47	1
лазуревый	55	0,19	0
рдяный	36	0,13	1

3.3. Относительные цвета

Группа относительных отыменных прилагательных, имеющих «цветовое» значение, была названа нами «относительными цветами» (табл. 3). Такие прилагательные часто образованы от названий продуктов питания (*лимонный, вишневый, кремовый*) и природных материалов (*малахитовый, аквамариновый, изумрудный*), а также от названий других предметов (*кровавый, небесный, фиалковый*). Как правило, такие прилагательные могут быть переформулированы как «цвета этого предмета» (*небесный* — цвета неба, *угольный* — цвета угля и т.д.).

У слов этой группы в словаре зафиксировано в среднем два значения, из них цветное, как правило, указывается в словаре вторым. Наблюдается большой разброс в частотности (от десятков до десятков тысяч употреблений), связанный с частотой употребления слова в прямом значении. Как и у основных цветов, наблюдается зависимость количества значений от частоты употребления, однако, в отличие от них, никак не связанная с процентом употребления как цветообозначения. Большинство слов характеризуется достаточно малым процентом употребления в «цветном» значении. У части прилагательных этот показатель настолько мал, что употребление этих слов в качестве цвета можно считать окказиональным (*кирпичный, банановый, пламенный, ананасовый, яичный, небесный, порфировый* — % < 3).

Таблица 3. Фрагмент таблицы относительных цветов, отсортированных по частотности

Лемма	Число с/у в НКРЯ	ipm (instances per million)	% «цветного» значения	Число значений в БТС
золотой	45112	159,16	31	8
серебряный	15394	54,31	43	5
небесный	14399	50,80	2	6
кровавый	8797	31,04	13	6
огненный	6849	24,16	16	6
стальной	6498	22,93	8	3
молочный	4456	15,72	20	3
…				
салатный	162	0,57	33	2
опаловый	154	0,54	10	2
сапфировый	125	0,44	23	1 (нет цветного)
кумачный	122	0,43	11	0
ананасовый	52	0,18	2	1 (нет цветного)
порфировый	49	0,17	2	1 (нет цветного)
киноварный	32	0,11	1	0

Можно заметить, что часть прилагательных сильно выделяется тем, что употребляется чаще в значении цвета, чем в других значениях, и, следовательно, не вполне вписывается в описание этой группы. Такие цветонаименования были выделены нами в отдельную группу относительных оттенков.

3.4. Относительные оттенки

Относительные оттенки — переходная группа между оттенками и относительными цветами. Их морфемная структура по-прежнему указывает на связь с существительным, от которого оно образовано, но употребление тяготеет в сторону цвета, нежели просто относительного прилагательного. А.А. Брагина [7] называет первым этапом развития цветного прилагательного переход из относительного в качественное (цветовое) и утрату этимологических связей. На материале русского языка можно выделить ряд цветонаименований, уже завершивших этот процесс: *розовый*, *бордовый* (цвета вина бордо), *бежевый* и др. Неудивительно, что целый ряд относительных прилагательных

находится сейчас в переходном процессе, однако пока неясно, когда он завершится и завершится ли.

Таким образом, в плане количественных показателей с оттенками их объединяет малая частотность (<13 ipm) и высокий процент употребления в цветном значении, с относительными цветами — число значений, выделяемых словарем (табл. 4).

Таблица 4. Прилагательные с процентом употребления в цветном значении больше 50, отсортированные по частоте

Лемма	Число с/у в НКРЯ	ipm (instances per million)	% «цветного» значения	Число значений в БТС
золотистый	3491	12,32	97	2
серебристый	3016	10,64	97	3
малиновый	2895	10,21	71	2
сиреневый	1615	5,70	99	2
изумрудный	1343	4,74	88	2
янтарный	1134	4,00	65	2
каштановый	947	3,34	82	2
пепельный	793	2,80	85	2
бирюзовый	650	2,29	91	2
кремовый	536	1,89	78	2
рубиновый	419	1,48	65	2
васильковый	278	0,98	84	2
канареечный	192	0,68	76	2
фиалковый	141	0,50	50	1 (нет цветного)
шафранный	107	0,38	72	3
салатовый	89	0,31	98	0
аметистовый	88	0,31	72	2
ультрамариновый	70	0,25	94	2
аквамариновый	57	0,20	84	2
антрацитовый	41	0,14	76	0
сапфирный	25	0,09	88	2
рубинный	5	0,02	100	0

Сильно выделяются из общей картины относительных оттенков слова *золотистый, серебристый* и *малиновый*, имея в два раза большую частоту, чем все остальные слова этих групп. Их частоты сравнимы с самыми низкими среди основных цветов. Это говорит не только о более частом употреблении, но и об их более широкой сочетаемости. В будущем можно ожидать дальнейшего сближения этих прилагательных с группой основных цветов: рост употребления за счет расширения сочетаемости и приобретение обобщенного цветового значения.

Заключение

В работе представлен принципиально новый способ классификации цветных прилагательных, основанный на корпусном подходе и количественных измерениях. Для классификации предлагаются следующие меры: частотность слова (абсолютная частота слова в корпусе и число употреблений на миллион (ipm)), процент употребления в цветном значении и количество значений, выделяемых словарем. В качестве источников используются основной подкорпус Национального корпуса русского языка и Большой толковый словарь русского языка.

В процессе работы был проведен анализ более ста лексем, обозначающих цвет, с последующим обобщением полученных количественных значений. Данные по всем выделенным группам прилагательных отражены в таблице 5.

Таблица 5. Отличительные особенности групп прилагательных (пустые окошки указывают на отсутствие закономерности)

Группа	**с/у в НКРЯ**	**ipm**	**%**	**Значений**
Основные цвета	>3000	>11	>60 (в среднем 87)	
Оттенки	<4000	<14	100	1
Относительные цвета			в среднем 18	в среднем 2
Относительные оттенки	<2000	<13	>50	в среднем 2

Можно заметить, что в большинстве случаев данных о частоте слова и проценте его употребления в цветном значении может быть достаточно для отнесения слова к той или иной группе, количество выделяемых словарем значений, таким образом, является дополнительным критерием.

Предложенный метод классификации подтверждает взгляды Берлина и Кэя на состав группы основных цветов в русском языке. Исключаемые некоторыми исследователями из этой группы цветонаименования *голубой* и *розовый* по количественным характеристикам совершенно однозначно являются типичными членами группы основных цветов.

Безусловно, такой метод классификации сильно зависит от используемых источников: корпуса и словаря, выбранных для подсчета значений мер. Однако можно считать, что наиболее авторитетные, нормативные словари русского языка и наиболее крупные, сбалансированные и репрезентативные корпуса все же могут предоставить достаточно точное представление об употреблении слов в языке.

Частичная проверка системы классификации на корпусе Araneum Russicum Maius, превосходящем основной корпус НКРЯ по размерам более чем в три раза, показала, что для большинства цветонаименований относительные частоты и процент употребления в цветном значении сильно отличаются в меньшую сторону. Такие серьезные отличия вызваны в первую очередь различным составом корпусов. Корпуса семейства Aranea созданы по технологии wacky, основанной на сборе текстов для корпуса из интернета. Особенность таких корпусов заключается в том, что мы ничего не можем сказать об их сбалансированности — во-первых, отбор текстов из интернета — это во многом случайный процесс, во-вторых, в веб-документах вообще отсутствует жанровая метаразметка. Проверка адекватности выделенных критериев классификации на корпусах большого размера и сравнение результатов классификации цветонаименований является предметом дальнейшего исследования.

Выражаю благодарность Захарову Виктору Павловичу за идею исследования, а также за ценные советы и предложения, связанные с проведением классификации и написанием статьи.

Литература

[1] Макеенко И.В. Семантика цвета в разноструктурных языках: универсальное и национальное. Дис. … канд. филол. наук. Саратов, 1999.
[2] Фрумкина Р.М. Психолингвистика. М., 2001.
[3] Berlin B., Kay P. Basic Color Terms: Their University and Evolution. California: University of Californian Press, 1969.
[4] Иваровская В.И. Лексическое значение цветовых прилагательных в синагматико-парадигматическом и словообразовательном аспектах // Вестник СПбГУ. Сер. 2. 1998. № 2. С. 104–109.
[5] Большой толковый словарь русского языка. Гл. ред. С.А. Кузнецов. СПб: Норинт, 1998.
[6] Василевич А.П., Кузнецова С.Н., Мищенко С.С. Цвет и названия цвета в русском языке / Под общ. ред. А. П. Василевича. М.: КомКнига, 2005. 216 с.

[7] Брагина А.А. «Цветовые» определения и формирование новых значений слов и словосочетаний // Лексикология и лексикография. М., 1972. С. 73–104.

Corpus-Based Research of the Adjectives of the Lexico-Semantic Group “Color” in the Russian Language

E.A. Shukshina

Saint Petersburg State University

The paper presents a classification of the adjectives of the lexico-semantic group “color” based on quantitative measures obtained by corpus-based and lexicographic analysis: frequency of the word in the Russian National Corpus, the percentage of its use in the sense of color, the number of its senses in the Large Explanatory Dictionary of the Russian Language.

The analysis of more than 100 words enables us to distinguish four distinct classes of color terms: basic color terms, hues, relative colors and relative hues. Basic color terms have high frequency of more than 11 ipm (instances per million) and high percentage of use in the sense of color (more than 60%). Hues are the words that have exclusively color meaning and frequency less than 14 ipm. Relative colors and relative hues are relative adjectives that are used to denote colors; both groups have on average 2 senses in the dictionary. Relative colors include those that are rarely used in the sense of color (less than 50%) and their frequency in the corpus almost entirely depends on its use in its non-color meanings. As to relative hues, their color meaning is dominant (percentage of usage as color is more than 50) and the upper limit of their frequency is close to that of regular hues (13 ipm).

This method allows us to describe the contemporary state of the lexico-semantic group in question and predict its further development.

Keywords: corpus linguistics, color terms, the Russian language, lexico-semantic group

Сведения об авторах

Белов Сергей Александрович, кандидат юридических наук, доцент, Санкт-Петербургский государственный университет, декан юридического факультета, доцент кафедры конституционного права, директор НИИ Проблем государственного языка, ORCID 0000-0003-4935-9658

Блинова Ольга Владимировна, кандидат филологических наук, Санкт-Петербургский государственный университет, доцент кафедры общего языкознания, ORCID 0000-0002-5665-3495

Бойков Владимир Николаевич, Ярославский государственный университет им. П. Г. Демидова, математик, ORCID 0000-0001-6640-4129

Гулида Виктория Борисовна, кандидат филологических наук, доцент, Санкт-Петербургский государственный университет, доцент кафедры общего языкознания

Еникеева Екатерина Владимировна, Санкт-Петербургский государственный университет, аспирант

Захаров Виктор Павлович, кандидат филологических наук, доцент, Санкт-Петербургский государственный университет, доцент, ORCID 0000-0003-0522-7469

Зубов Владислав Иванович, Санкт-Петербургский государственный университет, магистрант кафедры общего языкознания

Каряева Мария Сергеевна, магистр, Ярославский государственный университет им. П. Г. Демидова, аспирант, ORCID 0000-0003-4466-1735

Кириллова Алёна Олеговна, Санкт-Петербургский государственный университет, магистрант

Коган Марина Самуиловна, кандидат технических наук, Санкт-петербургский политехнический университет Петра Великого

Колотаева Анна Мизайловна, магистр, Санкт-петербургский политехнический университет Петра Великого

Ларионова Елена Юрьевна, Европейский Университет в Санкт-Петербурге, старший лаборант

Мартыненко Григорий Яковлевич, доктор филологических наук, профессор, Санкт-Петербургский государственный университет, профессор, ORCID 0000-0002-5962-2395

Масевич Андрей Цезаревич, Санкт-Петербургский государственный институт культуры, старший преподаватель, ORCID 0000-0002-1752-8915

Мельник Алексей Геннадиевич, бакалавр, Санкт-Петербургский государственный университет, магистрант

Микони Станислав Витальевич, доктор технических наук, профессор, Санкт-Петербургский институт информатики и автоматизации Российской Академии Наук, ведущий научный сотрудник

Митрофанова Ольга Александровна, Санкт-Петербургский государственный университет, доцент

Москвина Анна Денисовна, Санкт-Петербургский государственный университет, аспирант, ORCID 0000-0001-7400-8097

Мочалов Владимир Анатольевич, кандидат технических наук, Институт космофизических исследований и распространения радиоволн ДВО РАН, старший научный сотрудник, ORCID 0000-0003-0588-0361

Мочалова Анастасия Викторовна, кандидат технических наук, Институт космофизических исследований и распространения радиоволн ДВО РАН, научный сотрудник

Пильщиков Игорь Алексеевич, доктор филологических наук, Московский государственный университет им. М. В. Ломоносова, ведущий научный сотрудник; Таллиннский университет, старший научный сотрудник, ORCID 0000-0003-0153-6598

Плетнева Анастасия Дмитриевна, Санкт-Петербургский государственный университет, магистрант

Попов Андрей Михайлович, ООО Инфо-Кьюбс, лингвист-разработчик

Попова Татьяна Ивановна, бакалавр, Санкт-Петербургский государственный университет, магистрант.

Райков Александр Николаевич, доктор технических наук, профессор, Институт проблем управления РАН, ведущий научный сотрудник; Институт философии РАН, главный научный сотрудник, ORCID 0000-002-6726-9619

Тильманс Анна, магистр, Ганноверский университет им. Лейбница

Толстикова Полина Сергеевна, Институт лингвистических исследований РАН, аспирант

Хохлова Мария Владимировна, кандидат филологических наук, Санкт-Петербургский государственный университет, доцент, ORCID 0000-0001-9085-0284

Чекменева Анна Владимировна, Санкт-Петербургский государственный университет, студент, ORCID 0000-0001-5982-8919

Шерстинова Татьяна Юрьевна, кандидат филологических наук, Санкт-Петербургский государственный университет, доцент, ORCID 0000-0002-9085-3378

Шрот-Вихерт Зигрун, кандидат филологических наук, Ганноверский университет им. Лейбница

Шукшина Елена Александровна, Санкт-Петербургский государственный университет, студент, ORCID 0000-0002-6014-9136

Ярошевич Анна Михайловна, магистр, Санкт-петербургский политехнический университет Петра Великого

Авторский указатель

Оглавление

Компьютерная лингвистика и вычислительные онтологии. Выпуск 2 (Труды XXI Международной объединенной конференции «Интернет и современное общество, IMS-2018, Санкт-Петербург, 30 мая - 2 июня 2018 г. Сборник научных статей). — СПб: Университет ИТМО, 2018. — 160 с.

Сборник научных статей

Компьютерная лингвистика и вычислительные онтологии

Выпуск 2

Под редакцией В.П. Захарова
Дизайн обложки С. Н. Ушаков
Оригинал-макет Е. Е. Нестерова
Редакционно-издательский отдел Университета ИТМО
Зав. РИО Н.Ф. Гусарова
Подписано к печати 28.06.2018
Заказ № 4136
Тираж 50 экз.

Университет ИТМО. 197101, Санкт-Петербург,
Кронверкский пр., 49